UL: PRINCIPAL COMPONENT ANALYSIS

In this Chapter, we focus on methods for dimension reduction. In particular, we are interested in projective methods = find low dimensional projections that extract / summarize useful information about the data.

- Why dimension reduction?
  - ↘ Data visualization [biplots]
  - ↘ Extracting key low dimensional features (can be used as a pre-processing step in a SL problem)
  - ↘ Curse of dimensionality / Rates of convergence.
- In this chapter we discuss Principal Component Analysis (PCA), kernel PCA (kPCA) and probabilistic PCA (pPCA).

I - PRINCIPAL COMPONENT ANALYSIS.

I.1. Matrix Decompositions.

We first review key concepts in linear algebra, which will be useful later on.

- The SPECTRAL DECOMPOSITION of a $(p \times p)$ symmetric matrix $\underline{\underline{S}}$ is

$$\underline{\underline{S}} = \lambda_1 \underline{e}_1 \underline{e}_1^t + \cdots + \lambda_p \underline{e}_p \underline{e}_p^t,$$

($\lambda_1, \ldots, \lambda_p \in \mathbb{R}$, $\underline{e}_1, \ldots, \underline{e}_p \in \mathbb{R}$)

where
- $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\underline{\underline{S}}$
- $\underline{e}_1, \ldots, \underline{e}_p$ are the associated normalized eigenvectors, i.e. $\underline{e}_i^t \underline{e}_i = 1$ and $\underline{e}_i^t \underline{e}_j = 0 \quad \forall i \neq j$

Put $\underset{(k\times k)}{\underline{\underline{P}}} = \begin{pmatrix} | & & | \\ \underline{e}_1 & \cdots & \underline{e}_p \\ | & & | \end{pmatrix}$ so that $\underline{\underline{P}}\,\underline{\underline{P}}^t = \underline{\underline{P}}^t\underline{\underline{P}} = \underline{\underline{I}}$

$\underline{\underline{P}}$ is ORTHOGONAL

The spectral decomposition of $\underline{\underline{S}}$ can be written

$$\boxed{\underline{\underline{S}} = \underline{\underline{P}}\,\underline{\underline{\Lambda}}\,\underline{\underline{P}}^t},$$

where $\underline{\underline{\Lambda}}$ is a diagonal matrix, $\underline{\underline{\Lambda}} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$

→ A $(p\times p)$ symmetric matrix $\underline{\underline{S}}$ is (strictly) positive definite iff every eigenvalue is positive: $\lambda_i > 0$.

→ If $\underline{\underline{S}}$ is a (strictly) positive definite symmetric matrix, we have that

$$\underline{\underline{S}}^{-1} = \underline{\underline{P}}\,\underline{\underline{\Lambda}}^{-1}\underline{\underline{P}}^t = \sum_{i=1}^{p} \frac{1}{\lambda_i}\,\underline{e}_i\,\underline{e}_i^t,$$

since

$$(\underbrace{\underline{\underline{P}}\,\underline{\underline{\Lambda}}^{-1}\,\underline{\underline{P}}^t}_{\underline{\underline{S}}^{-1}})(\underbrace{\underline{\underline{P}}\,\underline{\underline{\Lambda}}\,\underline{\underline{P}}^t}_{\underline{\underline{S}}}) = \underline{\underline{P}}\,\underline{\underline{\Lambda}}^{-1}\underline{\underline{\Lambda}}\,\underline{\underline{P}}^t = \underline{\underline{P}}\,\underline{\underline{P}}^t = \underline{\underline{I}}$$

→ Put $\underline{\underline{\Lambda}}^{1/2} = \begin{pmatrix} \Lambda_1^{1/2} & & 0 \\ & \ddots & \\ 0 & & \Lambda_p^{1/2} \end{pmatrix}$.

The SQUARE ROOT of a Matrix $\underline{\underline{S}}$ is (symm positive definite)

$$\boxed{\underline{\underline{S}}^{1/2} = \underline{\underline{P}}\,\underline{\underline{\Lambda}}^{1/2}\,\underline{\underline{P}}^t}.$$

Note that
- $S^{1/2}S^{1/2} = S$ (→ hence the name)
- $S^{-1/2} \equiv (S^{1/2})^{-1} = P\,\Lambda^{-1/2}\,P^t$
- $S^{1/2}S^{-1/2} = S^{-1/2}S^{1/2} = I$
- $S^{-1/2}S^{-1/2} = S^{-1}$

- The SINGULAR VALUE DECOMPOSITION (SVD) of a $(n \times p)$ matrix X of rank $r \leq p$ $(p < n)$ is:

$$\boxed{\underset{=}{X} = \underset{=}{U}_r \underset{=}{\Lambda}_r \underset{=}{V}_r^t}$$

where . $\underset{=}{U}_r = (n \times r)$ orthonormal matrix $\underset{=}{U}_r^t \underset{=}{U}_r = I$

. $\underset{=}{V}_r = (p \times r)$ orthonormal matrix $\underset{=}{V}_r^t \underset{=}{V}_r = I$

. $\underset{=}{\Lambda}_r = (r \times r)$ diagonal matrix with entries > 0

The singular values.

→ Put $\underset{=}{U}_r = \begin{pmatrix} | & & | \\ \underset{-}{u}_1 & \cdots & \underset{-}{u}_r \\ | & & | \end{pmatrix}$, $\underset{-}{u}_i$ is $(n \times 1)$

$$\underset{=}{V}_r = \begin{pmatrix} | & & | \\ \underset{-}{v}_1 & \cdots & \underset{-}{v}_r \\ | & & | \end{pmatrix}, \quad \underset{-}{v}_i \text{ is } (p \times 1)$$

Then

$$\underset{=}{X}^t \underset{=}{X} = V \Lambda \underline{U^t U} \Lambda V^t = V \Lambda^2 V^t$$

$$\Rightarrow (\underset{=}{X}^t \underset{=}{X}) \underset{-}{v}_i = \lambda_i^2 \underset{-}{v}_i$$

and

$$\underset{=}{X} \underset{=}{X}^t = U \Lambda V^t V \Lambda U^t = U \Lambda^2 U^t$$

$$\Rightarrow (\underset{=}{X} \underset{=}{X}^t) \underset{-}{u}_i = \lambda_i^2 \underset{-}{u}_i$$

ie $\underset{=}{X} \underset{=}{X}^t$ has eigenvalue-eigenvector pair $(\lambda_i^2, \underset{-}{u}_i)$

$\underset{=}{X}^t \underset{=}{X}$ has " " $(\lambda_i^2, \underset{-}{v}_i)$

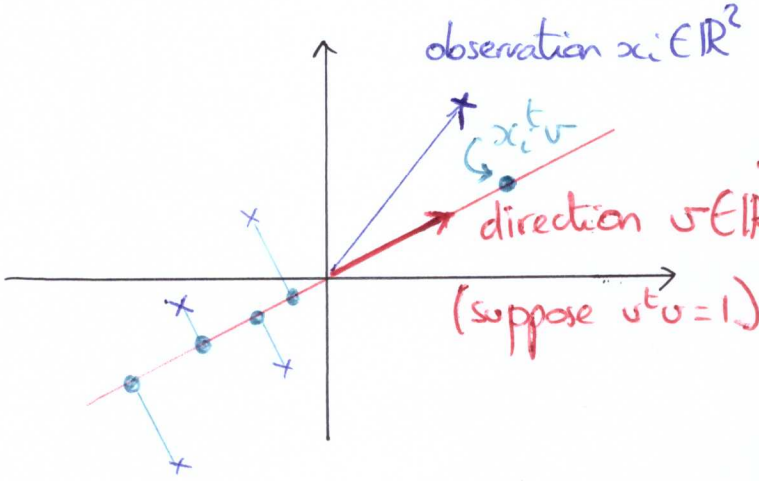
XX^t and $X^t X$ have the same eigenvalues.

I.2 Derivation of Principal Components.

Given a set of observations $\mathcal{L}_n = \{x_1, \dots, x_n\}$, $x_i \in \mathbb{R}^d$, you want to find a direction $v \in \mathbb{R}^d$ along which the projections $\langle x_i, v \rangle = x_i^t v$ give a 'good' one-dimensional approximation of the original data.

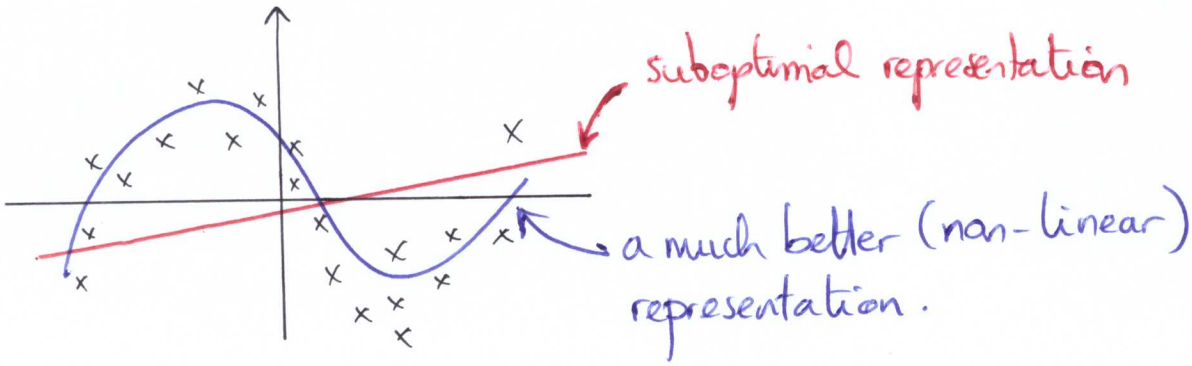
we consider linear combinations of the covariates.

loose as little information as possible.



Idea: The variability observed can be accounted for by a small number a such linear combinations.
(→ the so-called principal components)

Drawback: We will only be able to detect/find linear trends, which in many cases will be suboptimal:



Notation: let $X = (X_1, \dots, X_d)^t \in \mathbb{R}^d$ be a random vector with covariance matrix $\underline{\underline{\Sigma}}$.
Eigenvalues of $\underline{\underline{\Sigma}}$ are given by $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d > 0$.

We consider linear combinations $\begin{cases} z_1 = a_1^t \underline{X} \\ \vdots \\ z_d = a_d^t \underline{X} \end{cases}$ $\underline{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_d \end{pmatrix}$

(number of linear comb. considered = dimension of $\underline{X}$)

a_i is a $(d \times 1)$ vector

Then → $\text{Var } z_i = \underline{a}_i^t \underline{\underline{\Sigma}} \underline{a}_i$

→ $\text{Cov}(z_i, z_j) = \underline{a}_i^t \underline{\underline{\Sigma}} \underline{a}_j$

The principal components are uncorrelated linear combinations $z_1, \dots, z_d$ whose variances are as large as possible.
(variance = indicator of how much variability is in the data)

- First Principal Component = Linear combination $\underline{a}_1^t \underline{X}$ that maximizes $\text{Var}(\underline{a}_1^t \underline{X})$, subject to $\underline{a}_1^t \underline{a}_1 = 1$

(otherwise $\text{Var}(a_1^t X)$ could be made arbitrarily large)

- Second Principal Component = Linear combination $\underline{a}_2^t \underline{X}$ that maximizes $\text{Var}(\underline{a}_2^t \underline{X})$, subject to

$$\begin{cases} \underline{a}_2^t \underline{a}_2 = 1 \\ \text{Cov}(\underbrace{\underline{a}_1^t \underline{X}}_{z_1}, \underbrace{\underline{a}_2^t \underline{X}}_{z_2}) = 0 \end{cases}$$

Second PC is uncorrelated with the first one.

- And So on.../...

Q: How to find directions $\underline{a}_1, \dots, \underline{a}_d$? We need the following lemma:

Lemma: Let $\underline{\underline{S}}_{(p \times p)}$ be a symmetric positive definite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$ and associated eigenvectors $e_1, \dots, e_p$.

Then

$$\max_{x\neq 0} \frac{x^t S x}{x^t x} = \lambda_1, \text{ attained at } x = e_1$$

$$\min_{x\neq 0} \frac{x^t S x}{x^t x} = \lambda_p, \text{ attained at } x = e_p$$

$$\max_{x \perp e_1, \dots, e_l} \frac{x^t S x}{x^t x} = \lambda_{l+1}, \text{ attained at } x = e_{l+1}$$

$$(l = 1, \dots, k-1)$$

proof: Let $\cdot P = \begin{pmatrix} | & & | \\ e_1 & \cdots & e_p \\ | & & | \end{pmatrix}$ $\cdot \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$

Put $\cdot S^{1/2} = P \Lambda^{1/2} P^t$

$\cdot z = P^t x$

(note that $x \neq 0 \Rightarrow P^t x = z \neq 0$)

since orthogonal $\Rightarrow 0 \neq x = Pz$

* We have

$$\frac{x^t S x}{x^t x} = \frac{x^t S^{1/2} S^{1/2} x}{x^t x} = \frac{x^t (P \Lambda^{1/2} \underbrace{P^t}_{I})(P \Lambda^{1/2} P^t) x}{x^t (P P^t) x} = \frac{z^t \Lambda z}{z^t z}$$

$$\Rightarrow \frac{x^t S x}{x^t x} = \frac{\sum \lambda_i z_i^2}{\sum z_i^2} \leq \frac{\sum (\max \lambda_j) z_i^2}{\sum z_i^2} = \lambda_1$$

Moreover, setting $x = e_1$ gives $z = P^t e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ since

$$e_j^t e_i = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{o/w} \end{cases}$$

For this choice of x we have $\frac{z^t \Lambda z}{z^t z} = \lambda_1 =$ the max value.

$$\Rightarrow \max_{x\neq 0} \frac{x^t S x}{x^t x} = \frac{e_1^t \Lambda e_1}{e_1^t e_1} = \lambda_1$$

* A similar argument replacing max by min gives

$$\min_{x \neq 0} \frac{x^t \Sigma x}{x^t x} = \frac{e_p^t \Sigma e_p}{e_p^t e_p} = \lambda_p .$$

* Now, $x = Pz$

$$= z_1 e_1 + \cdots + z_p e_p$$

So $x \perp e_1, \ldots, e_\ell \Rightarrow 0 = e_i^t x = z_1 e_i^t e_1 + \cdots + z_p e_i^t e_p$

$$(i \leq \ell) \qquad = z_i$$

Thus,

$$\frac{x^t \Sigma x}{x^t x} = \frac{\sum_{i=\ell+1}^{k} \lambda_i z_i^2}{\sum_{\ell+1}^{k} z_i^2} \leq \lambda_{\ell+1}$$

Taking $z_{\ell+1} = 1$ and $z_{\ell+2} = \cdots = z_p$ gives the max ∎

We have the following result:

> The i-th principal component of $\underline{X}$ is given by
>
> $$z_i = \underline{e}_i^t \underline{X} \quad (= \text{coordinate of } \underline{X} \text{ in the } \underline{e}_i \text{ direction}),$$
>
> where $\underline{e}_1, \ldots, \underline{e}_d$ are the eigenvectors of $\underline{\underline{\Sigma}}$ (cov matrix of $\underline{X}$) associated with the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d > 0$.
>
> Moreover, $\begin{cases} \text{var } z_i = \underline{e}_i^t \underline{\underline{\Sigma}} \underline{e}_i = \lambda_i \\ \text{cov}(z_i, z_j) = \underline{e}_i^t \underline{\underline{\Sigma}} \underline{e}_j = 0 \end{cases}$

Remark: In practice, we use the eigenvalues and eigenvectors of the sample covariance matrix $\underline{\underline{S}}$, calculated from a random sample $\mathcal{L}_n = \{\underline{X}_1, \ldots, \underline{X}_n\}$, $\underline{X}_i \in \mathbb{R}^d$:

$$\underset{(d\times d)}{\underline{\underline{S}}} = \frac{1}{n} \underset{(d\times n)}{\underline{\underline{X}}^t} \underset{(n\times d)}{\underline{\underline{X}}}, \text{ where } \underline{\underline{X}} = \begin{pmatrix} X_{11} & \cdots & X_{1d} \\ \vdots & & \\ X_{n1} & \cdots & X_{nd} \end{pmatrix} = \begin{pmatrix} \underline{X}_1^t \\ \vdots \\ \underline{X}_n^t \end{pmatrix}$$

Using the SVD decomposition of $\underline{\underline{X}} = U \Lambda V^t$,
$\underline{\underline{X}}^t \underline{\underline{X}} = V \Lambda^2 V^t$ so that $\underline{\underline{X}}^t \underline{\underline{X}}$ has eigenvalue-eigenvector pair (λ_i^2, v_i) (see page 3)

Thus $\underline{\underline{S}}$ has e-e pair $\left(\frac{\lambda_i^2}{n}, v_i\right)$ [Assume matrix $\underline{\underline{X}}$ centered : $\underline{\underline{S}} = n^{-1} \underline{\underline{X}}^t \underline{\underline{X}}$]

- The i-th principal component is thus $\underline{z}_i = \underline{\underline{X}}\, \underline{v}_i$,
  $(n \times 1)$ $(n \times d)(d \times 1)$

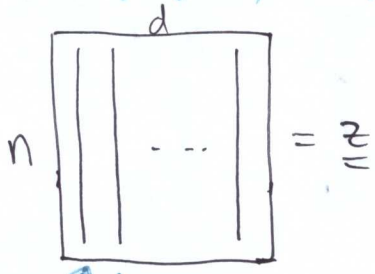
  and its variance is $\frac{\lambda_i^2}{n}$.

  [n realizations of the i-th PC ↑ of X]

  ↳ Putting all d PC into a matrix =

$$\underline{\underline{Z}} = \underline{\underline{X}}\,\underline{\underline{V}} = \underline{\underline{U}}\,\underline{\underline{\Lambda}}\,\underline{\underline{V}}^t\,\underline{\underline{V}} = \underline{\underline{U}}\,\underline{\underline{\Lambda}} \quad \Rightarrow \underline{z}_i = \lambda_i \underline{u}_i$$

$(n \times d)$ $(n \times d)(d \times d)$ $(n \times d)(d \times d)$ $(n \times 1)$ $(n \times 1)$



n realizations of the first PC

n realizations of the second PC, and so on .../...

Proof = To find the first principal component, we need to solve

$$\text{maximize}_{\underline{a} \neq 0} \ \underline{a}^t \underline{\underline{\Sigma}}\, \underline{a} \quad \text{subject to} \ \underline{a}^t \underline{a} = 1$$

$$\Leftrightarrow \quad \text{maximize}_{\underline{a} \neq 0} \ \frac{\underline{a}^t \underline{\underline{\Sigma}}\, \underline{a}}{\underline{a}^t \underline{a}} \quad \text{subject to} \ \underline{a}^t \underline{a} = 1$$

Recall that $\max_{\underline{a} \neq 0} \frac{a^t \Sigma a}{a^t a} = \frac{e_1^t \Lambda e_1}{e_1^t e_1} = \lambda_1$, with $e_1^t e_1 = 1$

(Lemma p.5) $e_1^t \Sigma e_1 = \text{Var}\, Z_1$

And similarly for the subsequent principal components ■

I.3. Eckart-Young Theorem.

Let $\underline{\underline{X}}$ be an $(n \times d)$ matrix with $n \geq d$. A common problem (with diverse applications) is to find approximations of $\underline{\underline{X}}$ that are of rank r lower than rank $\underline{\underline{X}}$.

↑ "low-rank approximation problem"

Finding a 'best' low-rank approximation requires to specify an objective function to minimize, in terms of the matrix norm of the difference between the original matrix $\underline{\underline{X}}$ and its approximate. When the norm is the FROBENIUS NORM, the Eckart-Young (EY) theorem provides an answer.

EY Theorem.

Let
- X be an $(n \times d)$ matrix of observations of rank $d \leq n$. (centered)
- $X = U \Lambda V^t$ its SVD decomposition. $(n \times d)(n \times d)(d \times d)(d \times d)$

Then the solution to the optimization problem

$$\underline{\underline{A}}^* = \underset{\substack{\underline{\underline{A}} \in \mathbb{R}^{n \times d} \\ \text{rank } \underline{\underline{A}} = r \leq d}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{d} (x_{ij} - a_{ij})^2 \right\}$$

↑ $\underline{\underline{A}}$ has entries a_{ij}

↑ Frobenius norm of $\underline{\underline{X}} - \underline{\underline{A}}$.

is given by

$$A^* = U \Lambda^* V^t,$$

where $\Lambda^* = \begin{pmatrix} \lambda_1 & & & 0 \\ & \ddots & & \\ & & \lambda_r & \\ 0 & & & 0 \end{pmatrix}$

Moreover, $\sum_{i=1}^{n} \sum_{j=1}^{d} (x_{ij} - a_{ij})^2 = \sum_{k=r+1}^{d} \lambda_k^2$.

↑ Sum of the variance of the remaining PC.

- **Interpretation.**
- Recall that the i-th PC is $\underset{(n\times 1)}{\underline{z}_i} = \underset{(n\times d)}{\underline{\underline{X}}}\,\underset{(d\times 1)}{\underline{v}_i} = \lambda_i \underset{(n\times 1)}{\underline{u}_i}$.

Put $\underline{z}_i = \begin{pmatrix} z_{1i} \\ \vdots \\ z_{ni} \end{pmatrix} \in \mathbb{R}^n$, $i = 1, \dots, d$, $\underset{(n\times d)}{\underline{\underline{Z}}} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1d} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & & z_{nd} \end{pmatrix}$

- The j-th row of $\underline{\underline{X}}$ (corresponding to the j-th observation) is thus approximated by the j-th row of $\underline{\underline{A}}^* = \underline{\underline{U}}\,\underline{\underline{\Lambda}}^*\,\underline{\underline{V}}^t$, i.e.

our rank r approx

$$\underset{(1\times d)}{\hat{\underline{x}}_j^t} \equiv \underset{(1\times d)}{\underline{a}_j^{*t}} = \underbrace{z_{j1}\,\underline{v}_1^t}_{(1\times d)} + \underbrace{z_{j2}\,\underline{v}_2^t}_{(1\times d)} + \cdots + \underbrace{z_{jr}\,\underline{v}_r^t}_{(1\times d)}, \quad r \le d, \; j = 1, \dots, n$$

Principal Components

Principal directions = eigenvectors of the sample cov matrix $S = n^{-1} X^t X$, X assumed centered

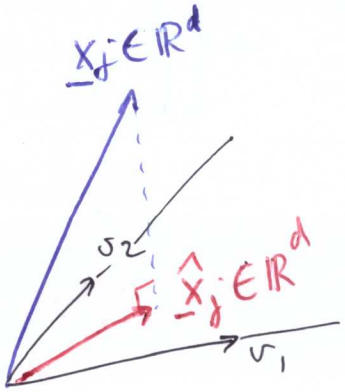
with $\underline{\underline{A}}^* = \begin{pmatrix} — & \underline{a}_1^{*t} & — \\ & \vdots & \\ — & \underline{a}_n^{*t} & — \end{pmatrix}$

- Note that $z_{ji} = \underline{x}_j^t\,\underline{v}_i$ since $\begin{bmatrix} \\ \underline{z}_i \\ \\ \end{bmatrix} = \begin{bmatrix} — & \underline{x}_1^t & — \\ & \vdots & \\ — & \underline{x}_n^t & — \end{bmatrix} \begin{bmatrix} \\ \underline{v}_i \\ \\ \end{bmatrix}$

j-th obs, $1 \le j \le n$; i-th PC, $1 \le i \le d$

$$\Rightarrow \boxed{\hat{\underline{x}}_j^t = \langle \underline{x}_j, \underline{v}_1 \rangle \underline{v}_1^t + \cdots + \langle \underline{x}_j, \underline{v}_r \rangle \underline{v}_r^t}$$

$\hat{\underline{x}}_j \in \mathbb{R}^d$ = $\perp$ projection of $\underline{x}_j$ onto the column space of the first r columns of $\underline{\underline{V}}$ $(d\times d)$.



#TakeAwayMessage

The approximation $\hat{\underline{x}}_j$ is an orthogonal projection of $\underline{x}_j$, minimizing the squared error between $\underline{x}_j$ and any approximating vector

Proof of the EY theorem.

- **Step I**

Let $\underset{(d\times 1)}{e_1}, \dots, \underset{(d\times 1)}{e_r}$ be r orthonormal vectors, and

$$\underset{(d\times r)}{E} = \begin{pmatrix} | & & | \\ e_1 & \cdots & e_r \\ | & & | \end{pmatrix} ; \qquad E^t E = \underset{(r\times r)}{I}$$

↑ our goal will be to find the entries of E.

Our goal is to minimize the quantity

$$\sum_{i=1}^{n}\sum_{j=1}^{d} (x_{ij} - a_{ij})^2 = \sum_{i=1}^{n} (x_i - a_i)^t (x_i - a_i)$$

over the set of matrices $A \in \mathbb{R}^{n\times d}$, with rank $A = r \le d$.

$$A = \begin{pmatrix} — & a_1^t & — \\ & \vdots & \\ — & a_n^t & — \end{pmatrix}$$

$$X = \begin{pmatrix} — & x_1^t & — \\ & \vdots & \\ — & x_n^t & — \end{pmatrix}$$

This guy is our a_i. We will first derive the optimal b_i minimizing the sq error, and then the columns of E

First, let's compute the square norm of $\underset{(d\times 1)}{x_i} - \underset{(d\times r)}{E}\underset{(r\times 1)}{b_i}$, for an arbitrary vector $b_i \in \mathbb{R}^r$.

$$\begin{aligned} x_i - E b_i &= x_i - EE^t x_i + EE^t x_i - E b_i \\ &= (I - EE^t) x_i + E(E^t x_i - b_i) \end{aligned}$$

- $(x_i - E b_i)^t (x_i - E b_i) = x_i^t (I - EE^t) x_i$ ← idempotent
$+ (E^t x_i - b_i)^t E^t E (E^t x_i - b_i)$

since the cross products vanish : $(I - EE^t)E = E - E = 0$
$E^t (I - EE^t) = E^t - E^t = 0$

$\Rightarrow (x_i - E b_i)^t (x_i - E b_i) = x_i^t (I - EE^t) x_i$
$+ \underbrace{(E^t x_i - b_i)^t (E^t x_i - b_i)}$

This term is positive unless b_i is chosen such that $b_i = E^t x_i$
$\Rightarrow E b_i = EE^t x_i$

Now, we are looking for the columns of $\underline{\underline{E}}$

$\Rightarrow$ With $\underset{(d\times 1)}{a_i} = \underset{(d\times r)}{\underline{\underline{E}}} \underset{(r\times d)}{\underline{\underline{E}}^t} \underset{(d\times 1)}{x_i}$, we get

$$\sum_{i=1}^{n} (x_i - a_i)^t (x_i - a_i) = \sum_{i=1}^{n} x_i^t (I - \underline{\underline{E}}\,\underline{\underline{E}}^t) x_i$$
$$= \underbrace{\sum_{i=1}^{n} x_i^t x_i}_{\text{independent of } \underline{\underline{E}}} - \sum_{i=1}^{n} x_i^t EE^t x_i$$

minimizing this $\equiv$ maximizing this one.

OK, so it remains to find a matrix $\underline{\underline{E}}$ which will make $\sum_{i=1}^{n} x_i^t EE^t x_i$ as large as possible.

(⚠ Recall that E has $\perp$ vectors of unit length.)

Well,

$$\sum_{i=1}^{n} x_i^t EE^t x_i = \sum_{i=1}^{n} Tr\left(x_i^t \underbrace{EE^t x_i}\right) = \sum_{i=1}^{n} Tr\left(EE^t x_i x_i^t\right)$$

and introduce $S = n^{-1} \sum_{i=1}^{n} x_i x_i^t$ = sample cov matrix (obs are centered)
(dxd)
$= n^{-1} X^t X$

$$\Rightarrow \sum_{i=1}^{n} x_i^t E E^t x_i = n \, Tr(E E^t S)$$
$$= n \, Tr(E^t S E).$$

In other words, the best choice for E maximizes the sum of the diagonal elements of $E^t S E$.

↳ Selecting e_1 to maximize $e_1^t S e_1$, the first diagonal element of $E^t S E$, subject to $e_1^t e_1 = 1$, gives $e_1 = v_1$ (= eigenvector of the sample covariance matrix see page 8)

↳ For $e_2 \perp v_1$, $e_2^t S e_2$ is maximized by $e_2 = v_2$.
$e_2^t e_2 = 1$

↳ And so on.

⇒ We find that the matrix E maximizing $\sum_{i=1}^{n} x_i^t E E^t x_i$, denoted E^*, is thus $E^* = \begin{pmatrix} | & & | \\ v_1 & \dots & v_r \\ | & & | \end{pmatrix}$.
(dxr) (dx1) (dx1)

→ The best approximation matrix $A^* = \begin{pmatrix} — \, a_1^{*t} \, — \\ \vdots \\ — \, a_n^{*t} \, — \end{pmatrix}$ is such that

- $(A^*)^t = \begin{pmatrix} | & & | \\ a_1^* & \dots & a_n^* \\ | & & | \end{pmatrix}$
$= \begin{pmatrix} | & & | \\ E^* E^{*t} x_1 & \dots & E^* E^{*t} x_n \\ | & & | \end{pmatrix}$

- $A^* = \begin{pmatrix} — \, x_1^t E^* E^{*t} \, — \\ \\ — \, x_n^t E^* E^{*t} \, — \end{pmatrix}$
(nxd)

$$\Rightarrow A^* = X E^* E^{*t}$$
$$= U \Lambda \underbrace{V^t E^*}_{(d\times d)\,(d\times r)} E^{*t}$$

$$V^t E^* = \begin{pmatrix} - v_1^t - \\ \vdots \\ - v_d^t - \end{pmatrix} \begin{pmatrix} | & & | \\ v_1 & \cdots & v_r \\ | & & | \end{pmatrix}$$

$$= \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \\ \hline & 0 & \end{pmatrix} \begin{matrix} \updownarrow r \\ \updownarrow d \end{matrix}$$
$$\overset{\longleftrightarrow}{r}$$

so that $V^t E^* E^{*t} = \underset{(d\times d)}{\begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ \hline & 0 & \end{pmatrix}} \underset{(d\times r)}{\begin{pmatrix} - v_1^t - \\ \vdots \\ - v_r^t - \end{pmatrix}} \underset{(r\times d)}{}$

$$= \begin{pmatrix} — & v_1^t & — \\ & \vdots & \\ — & v_r^t & — \\ & 0 & \\ & \vdots & \\ & 0 & \end{pmatrix} \updownarrow d$$
$$\overset{\longleftrightarrow}{d}$$

Complete the missing rows with $v_{r+1}^t, \dots, v_d^t$, and set the associated singular values $\lambda_{r+1}, \dots, \lambda_d$ to zero. It's the same!

$$\Rightarrow A^* = U \Lambda^* V^t, \text{ where } \Lambda^* = \begin{pmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & \lambda_r & \\ & & & 0 \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix},$$

as asserted.

Step II. Residual error.

With this choice of $\underline{\underline{A}}$, it remains to compute the residual term

$$\sum_{i=1}^n (x_i - a_i^*)^t (x_i - a_i^*) = \underbrace{\sum_{i=1}^n x_i^t x_i}_{= n \operatorname{Tr} S} - \underbrace{\sum_{i=1}^n x_i^t E^* E^{*t} x_i}_{= n \operatorname{Tr}(E^{*t} S E^*)}$$

$$= n\left(\frac{\lambda_1^2}{n} + \cdots + \frac{\lambda_d^2}{n}\right)$$

$$= \lambda_1^2 + \cdots + \lambda_d^2$$

Term $Tr(E^{*t} S E^{*})$ $\rightarrow$ the i-th diagonal element of $E^{*t} S E^{*}$ is
$(r \times r)$

$$v_i^t S v_i = v_i^t \frac{\lambda_i^2}{n} v_i = \frac{\lambda_i^2}{n}$$

page 8

$$\Rightarrow Tr(E^{*t} S E^{*}) = \frac{\lambda_1^2}{n} + \dots + \frac{\lambda_r^2}{n},$$

and

$$\sum_{i=1}^{n} (x_i - a_i^*)^t (x_i - a_i^*) = (\lambda_1^2 + \dots + \lambda_d^2) - (\lambda_1^2 + \dots + \lambda_r^2)$$
$$= \sum_{k=r+1}^{d} \lambda_k^2 \qquad \blacksquare$$

Application - Visualisation of data with more than two predictors (aka BIPLOTS)

→ With only two variables, a scatterplot provides a visual information on both the sampling unit (observation) and the predictor

→ With more than two predictors, one may consider a matrix array of scatterplots. Alternatively, we may plot the first two principal components:

Recall that the j-th observation (j-th row of $\underline{\underline{X}}$ $(n \times d)$) is approximated by

best rank r approxim. in the Frobenius norm →

$$\hat{\underline{x}}_j^t \approx z_{j1} \underline{v}_1^t + z_{j2} \underline{v}_2^t + \dots + z_{jr} \underline{v}_r^t, \qquad j = 1, \dots, n$$

corresponding to the j-th row of $\underline{\underline{A}}^*$, where rank $\underline{\underline{A}}^* = r \leq d$ (top of page 10)

The rank-2 approximation is thus

$$\underset{(1 \times d)}{\hat{\underline{x}}_j^t} \approx z_{j1} \underset{(1 \times d)}{\underline{v}_1^t} + z_{j2} \underset{(1 \times d)}{\underline{v}_2^t} \qquad j = 1, \dots, n$$

In other words:

$$\begin{pmatrix} \hat{x}_{j1} \\ \vdots \\ \hat{x}_{jd} \end{pmatrix} = z_{j1} \begin{pmatrix} v_{11} \\ \vdots \\ v_{d1} \end{pmatrix} + z_{j2} \begin{pmatrix} v_{12} \\ \vdots \\ v_{d2} \end{pmatrix}$$

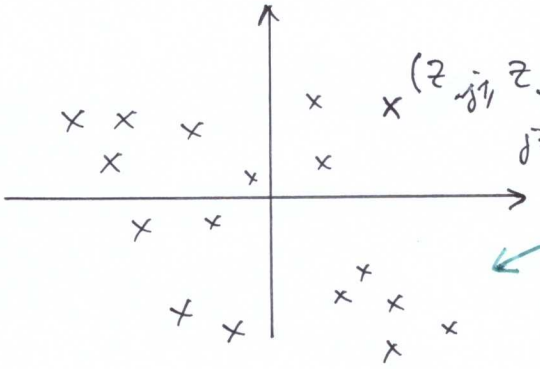
← 1st predicte

← dth predict

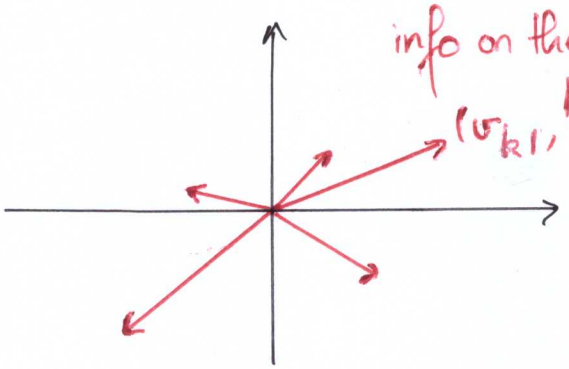
weights

(z_{j1}, z_{j2}) = coordinate of $\underline{x}_j$ in the plane defined by the first two eigenvectors.

= information about the sampling unit j.

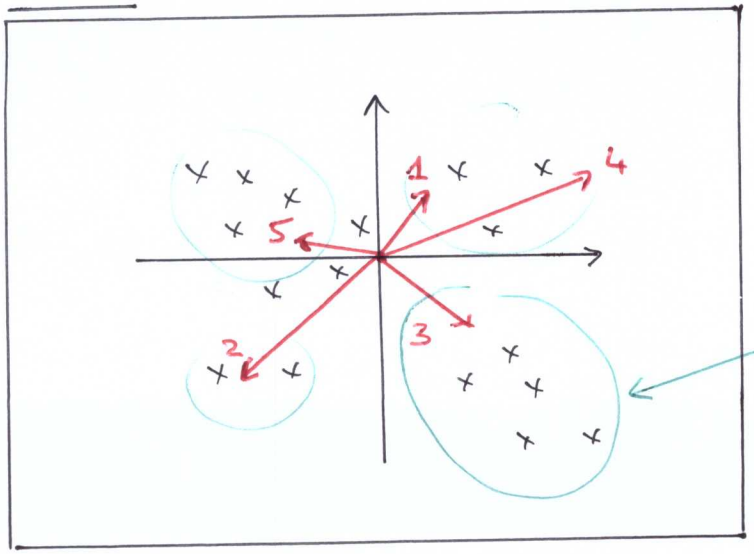


good, but we would like to add information about the predictors



Putting these two in a single plot yields:

BIPLOT



The third variable is responsible for grouping these points here, which all have a high value associated with the 3rd predictor

Before introducing probabilistic PCA (pPCA), let's summarize the conventional PCA results we obtained:

(i) Conventional PCA is a non-probabilistic approach: no probability distribution is attached to $\underset{(n\times d)}{\underline{\underline{X}}} = \begin{pmatrix} — & x_1^t & — \\ & \vdots & \\ — & x_n^t & — \end{pmatrix}$

(ii) The principal directions of $\underline{\underline{X}}$ are given by the eigenvectors of the sample covariance matrix

→ page 8

$$\underset{(d\times d)}{\underline{\underline{S}}} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^t = \frac{1}{n}X^tX$$

↑ if $\underline{\underline{X}}$ is centered

corresponding to the columns of the matrix $\underline{\underline{V}}$ in the SVD decomposition of $\underline{\underline{X}} = \underline{\underline{U}}\,\underline{\underline{\Lambda}}\,\underline{\underline{V}}^t$

→ page 3

(iii) The best rank r approximation (in the square error sense) of $\underline{x}_j$ is

$$\hat{\underline{x}}_j = z_{j1}\underline{v}_1 + \cdots + z_{jr}\underline{v}_r + \bar{x}$$

$$= \langle \underline{x}_j, \underline{v}_1\rangle \underline{v}_1 + \cdots + \langle \underline{x}_j, \underline{v}_r\rangle \underline{v}_r + \bar{x}$$

$$= \begin{bmatrix} | & & | \\ \underline{v}_1 & \cdots & \underline{v}_r \\ | & & | \end{bmatrix}\begin{bmatrix} z_{j1} \\ \vdots \\ z_{jr} \end{bmatrix} + \bar{x}$$

↖ $(d\times 1)$

(1) $$\hat{\underline{x}}_j = \underline{\underline{V}}_r\,\underline{z}_r^j + \bar{x}$$

→ page 10

"reconstruction from the PC"

↖ vector $\in \mathbb{R}^r$ whose entries correspond to the first r PC of the j-th observation

The reduced-dimensionality transformation of the j-th observation is

(2) $$\underline{z}_r^j = \underline{\underline{V}}_r^t(\underline{x}_j - \bar{x})$$

"extraction of the PC".

→ pPCA challenges (i) by deriving the principal components using a probabilistic model for the $\underline{x}$s. The model is motivated by relation (1), which links the principal components (soon to be the hidden / latent variables) to the original observations.

→ In view of (1), consider the model:

For $i=1,\dots,n$,

$$(*)\quad \underset{(d\times 1)}{\underline{x}_i} = \underset{(d\times r)}{\underline{\underline{W}}}\,\underset{(r\times 1)}{\underline{z}_i} + \underset{(d\times 1)}{\underline{\mu}} + \underset{(d\times 1)}{\underline{\varepsilon}_i} \qquad (r \le d)$$

($\underline{x}_i$: i-th observation; $\underline{z}_i$: $\equiv$ Principal Component)

where
- $\underline{z}_i \in \mathbb{R}^r$ is an r-dimensional vector of latent variables. Generally, $r < d$ so that the variable $\underline{z}$ offers a more parcimonious description of the data.
- $\underline{z}_i$ are assumed to be independent, with distribution
$$\underline{z}_i \sim \mathcal{N}(0, \underline{\underline{I}}_r)$$
- $\underline{\underline{W}} \in \mathbb{R}^{d\times r}$ = matrix of parameters to be estimated.
- $\underline{\varepsilon}_i \sim \mathcal{N}(0, \underline{\underline{\Psi}})$, with Ψ diagonal.
The choice of a diagonal covariance matrix for $\underline{\varepsilon}_i$ ensures that conditionally on $\{\underline{z}_i\}$ observations $\{\underline{x}_i\}$ are independent $\Rightarrow$ the dependency structure is explained by a small number of latent variable. What is left unexplained appears in the noise term $\underline{\varepsilon}_i$, which differs for each observation.

Given model (*), observations $\underline{x}_i$ are also normally distributed:

$$\underline{x}_i \sim \mathcal{N}(\underline{\mu}, \underbrace{\underline{\underline{W}}\underline{\underline{W}}^t + \underline{\underline{\Psi}}}_{\text{invariant under post-multiplication of } W \text{ by an orthogonal } (\equiv \text{rotation}) \text{ matrix } Q:\ (WQ)(Q^tW^t) = W(QQ^t)W^t = WW^t.})$$

FACTOR ANALYSIS (FA) (not pPCA yet)

The probabilistic formulation allows ML estimation of the parameters $\underline{\underline{W}}$ and $\underline{\underline{\Psi}}$, using an iterative procedure (EM algorithm), since no closed analytical expression of the MLE of $\underline{\underline{W}}$ and $\underline{\underline{\Psi}}$ are available in general.

Q: Is there a clear correspondance between $\underline{\underline{W}}$ and $\underline{\underline{V}}_r$?
For example, do the columns of $\underline{\underline{W}}$ and $\underline{\underline{V}}_r$ span the same space ?
In general, the answer is no.
However, links between the two exist in the special case of an ISOTROPIC noise model, ie when $\underline{\underline{\Psi}} = \sigma^2 \underline{\underline{I}}_d$.
And this is exactly the model considered by Tipping & Bishop
They named this model pPCA.

Informally, PCA $\leftrightarrow$ pPCA $\leftrightarrow$ Factor Analysis

pPCA is like a bridge, linking concepts from PCA and FA

pPCA model

$$\underset{(dx1)}{\underline{x}_i} = \underset{(dxr)}{\underline{\underline{W}}} \underset{(rx1)}{\underline{z}_i} + \underset{(dx1)}{\underline{\mu}} + \underset{(dx1)}{\underline{\varepsilon}_i} \quad , \quad i = 1, \dots, n$$

$$\underline{z}_i \sim \mathcal{N}(0, I_r)$$

$$\underline{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 I_d) \quad , \quad r \leq d$$

$$\Rightarrow \underline{x}_i \sim \mathcal{N}(\underline{\mu}, \underline{\underline{W}}\underline{\underline{W}}^t + \sigma^2 I_d)$$

Put $\underset{(dxd)}{\underline{\underline{C}}} = \underset{(dxr)}{\underline{\underline{W}}}\underset{(rxd)}{\underline{\underline{W}}}^t + \sigma^2 I_d$

Goal: Estimation of $\underline{\underline{W}}$ and σ^2 + Extraction of the hidden variable and/or reconstruction from the hidden variables.

How: Maximum likelihood, of course.

$$\mathcal{L} = \ln\left[\prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2}|C|^{1/2}} \exp\left\{-\frac{1}{2}(x_i-\mu)^t C^{-1}(x_i-\mu)\right\}\right]$$

$$= -\frac{n}{2}\left\{ d\ln(2\pi) + \ln|C| \right\} - \frac{1}{2}\underbrace{\sum_{i=1}^{n}(x_i-\mu)^t C^{-1}(x_i-\mu)}\Bigg\}$$

$$\sum_{i=1}^{n} \mathrm{Tr}\left\{(x_i-\mu)^t C^{-1}(x_i-\mu)\right\}$$

$$= \sum_{i=1}^{n} \mathrm{Tr}\left\{ C^{-1}(x_i-\mu)(x_i-\mu)^t \right\}$$

$$= n\,\mathrm{Tr}(C^{-1}S),$$

where we defined the (sample) covariance matrix as

$$\underline{\underline{S}} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^t.$$

$$\boxed{\mathcal{L} = -\frac{n}{2}\left\{ d\ln(2\pi) + \ln|C| + \mathrm{Tr}(C^{-1}S) \right\}}$$

→ The MLE of $\underline{\mu}$ is the mean of the observations:

$$\hat{\underline{\mu}}_{ML} = \frac{1}{n}\sum_{i=1}^{n}\underline{x}_i.$$

You may again pre-process the data by centering $\underline{\underline{X}}$.

→ The MLE of $\underline{\underline{W}}$, σ^2 → Explicit expression by direct maximization of $\mathcal{L}$. Gives no insight about the link between $\hat{\underline{\underline{W}}}_{MLE}$ and $\underline{\underline{V}}_r$.

→ Using EM algorithm.
(why bother? It allows us to deal with missing components in the $\underline{x}_i$'s.)

Remark: $\underline{\underline{S}} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_{ML})(x_i-\hat{\mu}_{ML})^t$, so centering the data yields $\underline{\underline{S}} = \frac{1}{n}\underline{\underline{X}}^t\underline{\underline{X}}$.

↖ Substitute this expression back to $\mathcal{L}$

(A) Direct Maximization of $\mathcal{L}$.

The derivative of $\mathcal{L}$ with respect to W is found to be:

$$\frac{\partial \mathcal{L}}{\partial W} = n\left(C^{-1}SC^{-1}W - C^{-1}W\right), \text{ where } \underset{(d\times d)}{C} = WW^t + \sigma^2 I$$

$\Rightarrow$ Stationary points satisfy the relation $\boxed{SC^{-1}W = W}$

Assume $W \neq 0$ [Solution $W = 0$ can be shown to be a minimum of $\mathcal{L}$]

&

$C \neq S$ [Unrealistic Scenario of exact covariance model]

- Consider the SVD decomposition of W:

$W = ULV^t$, where (W: $d\times r$, U: $d\times r$, L: $r\times r$, V: $r\times r$)

Our goal: identify U, L and V.

$U = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{pmatrix}$, columns are orthonormal

$L = \begin{pmatrix} l_1 & & 0 \\ & \ddots & \\ 0 & & l_r \end{pmatrix}$ ← some of these might be zero.

V = orthogonal matrix.

- First, we compute the product $C^{-1}W$:

$$C^{-1}W = (WW^t + \sigma^2 I)^{-1} W$$
$$= W(\sigma^2 I + W^t W)^{-1}$$
$$= ULV^t(\sigma^2 I + VLU^t ULV^t)^{-1}$$
$$= ULV^t(\sigma^2 I + VL^2V^t)^{-1}$$
$$= UL(\sigma^2 V + VL^2)^{-1}$$
$$= UL(V(\sigma^2 I + L^2))^{-1}$$
$$= UL(L^2 + \sigma^2 I)V^t$$

($V^tV = I$ and thus $(V^t) = V^{-1}$)

Useful: $\forall P, Q$,

$\sigma^2 P + PQP = P(\sigma^2 I + QP)$

$(\sigma^2 I + PQ)P$

$\Downarrow$

$P(\sigma^2 I + QP)^{-1} = (\sigma^2 I + PQ)^{-1}P$

Now, $S C^{-1} W = S \underline{U L (L^2 + \sigma^2 I)^{-1} V^t}$

$= W$

$= U L V^t$

Multiplying by V on the right-hand side and making use of $V^t V = I$, we get

$$S U L (L^2 + \sigma^2 I)^{-1} = U L$$

$$S U L = U \underbrace{L}_{\text{diagonal matrices. They commute}} \underbrace{(L^2 + \sigma^2 I)}_{\text{diagonal matrices. They commute}}$$

$$\boxed{(\Delta) \quad S U L = U (L^2 + \sigma^2 I) L}$$

We now need to consider two cases; depending on whether $\ell_j = 0$ or not.

- $\underline{\ell_j \neq 0}$, then relation (Δ) becomes

$$\underline{\underline{S}} \begin{pmatrix} | & & | \\ \ell_1 \underline{u}_1 & \cdots & \ell_r \underline{u}_r \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ \ell_1 (\ell_1^2 + \sigma^2) \underline{u}_1 & \cdots & \ell_r (\ell_r^2 + \sigma^2) \underline{u}_r \\ | & & | \end{pmatrix}$$

So that the j-th column satisfies

$$\underline{\underline{S}}\, \underline{u}_j = \ell_j (\ell_j^2 + \sigma^2) \underline{u}_j \qquad (\ell_j \neq 0)$$

$$\Rightarrow \underline{\underline{S}}\, \underline{u}_j = (\ell_j^2 + \sigma^2) \underline{u}_j$$

$\Rightarrow \underline{u}_j$ is an eigenvector of $\underline{\underline{S}}$, with eigenvalue $(\ell_j^2 + \sigma^2)$.

Let λ_j^S the j-th eigenvalue of $\underline{\underline{S}}$ (can be computed), so that

$$\boxed{\ell_j = (\lambda_j^S - \sigma^2)^{1/2}}$$

We identified one column of $\underline{\underline{U}}$ ... and also one term in the diagonal of $\underline{\underline{L}}$.

Related to the singular value λ_j of $\underline{\underline{X}}$ by $\lambda_j^S = n^{-1} \lambda_j^2$ (if $\underline{\underline{X}}$ is centered)

- $l_j = 0$, then u_j is arbitrary.

All stationary points can thus be written:

$$(0) \quad \boxed{W = U_r (K_r - \sigma^2 I_r)^{1/2} R}$$

where
- U_r is a $(d \times r)$ matrix comprising r column eigenvectors of S (which ones ?)
- K_r is a $(r \times r)$ diagonal matrix with elements:

$$k_j = \begin{cases} \lambda_j & = \text{the corresponding eigenvalue to } u_j. \\ \sigma^2 & \end{cases}$$

↖ gives $l_j = 0$ so that the associated column in U_r can indeed be chosen arbitrarily.

- R = an arbitrary orthogonal matrix ≡ rotation
  (indeed, the matrix V in the SVD decomposition of W is left unspecified.)

Now the question is = which column eigenvectors of S shall we select to get to the global maximum ?

→ Substitute expression (0) for W back into the likelihood $\mathcal{L}$, whose expression is given on page 19.
We need to compute $\ln|C|$ and $\mathrm{Tr}(C^{-1}S)$, $C = \sigma^2 I + WW^t$.

Useful facts: for an $(n \times n)$ matrix A,

$$|A| = \prod_{i=1}^{n} \lambda_i \quad , \quad \lambda_i = \text{eigenvalues of } A.$$

$$\mathrm{Tr}\, A = \sum_{i=1}^{n} \lambda_i$$

→ First, $\ln |\underline{\underline{C}}|$.

The determinant of $\underline{\underline{C}} = \sigma^2 I + \underline{\underline{W}}\underline{\underline{W}}^t$ is equal to the product of its eigenvalues.

$C = \sigma^2 I + U_r (K_r - \sigma^2 I) U_r^t$

Clearly, u_j are eigenvectors of C:

$$C u_j = \sigma^2 u_j + U_r \begin{pmatrix} \ddots & & \\ & \lambda_k^s - \sigma^2 & \\ & & \ddots \\ & 0 & \ddots \end{pmatrix} \begin{pmatrix} \\ \\ 1 \\ \\ \end{pmatrix} \leftarrow j\text{-th position}$$

Terms on the diagonal are either $\lambda_k^s - \sigma^2$ or 0.
Let q' = number of non-zero terms.

$$= \begin{cases} \lambda_j^s \, u_j \\ \text{or} \\ \sigma^2 \, u_j \end{cases}$$, depending on whether the j-th diagonal term is zero or not.

⇒ Denote by $\lambda_1^s, \dots, \lambda_{q'}^s$ the eigenvalues corresponding to the eigenvectors 'retained' in W. It follows that

$$\ln |C| = \sum_{j=1}^{q'} \ln \lambda_j^s + (d - q') \ln \sigma^2. \qquad 0 \le q' \le r$$

→ Next, $\mathrm{Tr}(C^{-1} S)$.

The trace of $C^{-1}S$ is equal to the sum of its eigenvalues:

$$C^{-1} S \underline{e}_j = \left(\sigma^2 I + U_r (K_r - \sigma^2) U_r^t\right)^{-1} S \underline{e}_j = \rho \underline{e}_j$$

pair (eigenvalue, eigenvector) we are looking for.

$$\Rightarrow \quad S \underline{e}_j = \rho C \underline{e}_j .$$

Take $\underline{e}_j = \underline{u}_j$ = eigenvector of S.

Then $C\,\underline{u}_j = \left(\sigma^2 I + U_r (K_r - \sigma^2 I) U_r^t\right) \underline{u}_j$

$$= \sigma^2 \underline{u}_j + \begin{cases} (\lambda_j^S - \sigma^2)\,\underline{u}_j \\ 0 \end{cases}$$

$$= \begin{cases} \lambda_j^S\ \underline{u}_j \\ \sigma^2\ \underline{u}_j \end{cases}$$, depending on whether the j-th diagonal term of $K_r - \sigma^2$ is zero or not.

↳ Eigenvalues of $C^{-1}S$ satisfy: $\lambda_j^S = \mu \times \begin{cases} \lambda_j^S \\ \sigma^2 \end{cases}$

since $S\underline{u}_j = \lambda_j^S \underline{u}_j$

$\Rightarrow$ $C^{-1}S$ has q' eigenvalues equal to 1 and $(d-q')$ eigenvalues equal to λ_j^S/σ^2, where $\lambda_{q'+1}^S, \ldots, \lambda_d^S$ ($q'+1 \leq j \leq d$) are the eigenvalues corresponding to the eigenvectors not 'retained' in W.

$$\boxed{\mathrm{Tr}(C^{-1}S) = \frac{1}{\sigma^2} \sum_{j=q'+1}^{d} \lambda_j^S + q'}$$

The likelihood can now be written:

$$\boxed{\mathcal{L} = -\frac{n}{2}\left\{ d \ln 2\pi + \sum_{j=1}^{q'} \ln \lambda_j^S + (d-q')\ln\sigma^2 + \frac{1}{\sigma^2} \sum_{j=q'+1}^{d} \lambda_j^S + q' \right\}}$$

Maximizing $\mathcal{L}$ with respect to σ^2 gives:

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = -\frac{n}{2}\left\{ \frac{d-q'}{\sigma^2} - \frac{1}{\sigma^4} \sum_{j=q'+1}^{d} \lambda_j^S \right\} = 0$$

$$\Rightarrow \boxed{\sigma^2 = \frac{1}{d-q'} \sum_{j=q'+1}^{d} \lambda_j^S}$$

Note that we do not know yet at this stage which λ_j^S are selected.

Substituting this expression back into $\mathcal{L}$ gives:

$$\mathcal{L} = -\frac{n}{2}\left\{ d \ln 2\pi + \sum_{j=1}^{q'} \ln \lambda_j^S + (d-q') \ln\left(\frac{1}{d-q'} \sum_{j=q'+1}^{d} \lambda_j^S\right) + (d-q') + q' \right\}$$

$$\mathcal{L} = -\frac{n}{2}\left\{ d \ln 2\pi + d + \sum_{j=1}^{q'} \ln \lambda_j^S + (d-q') \ln\left(\sum_{j=q'+1}^{d} \lambda_j^S / d-q'\right)\right\}.$$

Since $\sum_{j=1}^{d} \lambda_j^S = \operatorname{Tr} S$, (constant), $\sum_{j=1}^{q'} \ln \lambda_j^S = \text{Constant} - \sum_{j=q'+1}^{d} \ln \lambda_j^S$.

$\Rightarrow$ Maximizing $\mathcal{L}$ is equivalent to minimizing the quantity

$$\ln\left(\frac{1}{d-q'} \sum_{j=q'+1}^{d} \lambda_j^S\right) + \frac{1}{d-q'} \sum_{j=1}^{q'} \ln \lambda_j^S$$

↑ for a fixed q'.

$\Updownarrow$

Minimizing

$$\boxed{\ln\left(\frac{1}{d-q'} \sum_{j=q'+1}^{d} \lambda_j^S\right) - \frac{1}{d-q'} \sum_{j=q'+1}^{d} \ln \lambda_j^S =: E}$$

Advantage: this quantity to minimize depends only on the discarded values.

Remark: Jensen's inequality $\ln\left(\frac{1}{n}\sum x_i\right) \geq \frac{1}{n}\sum \log x_i$ ensures $E \geq 0$.

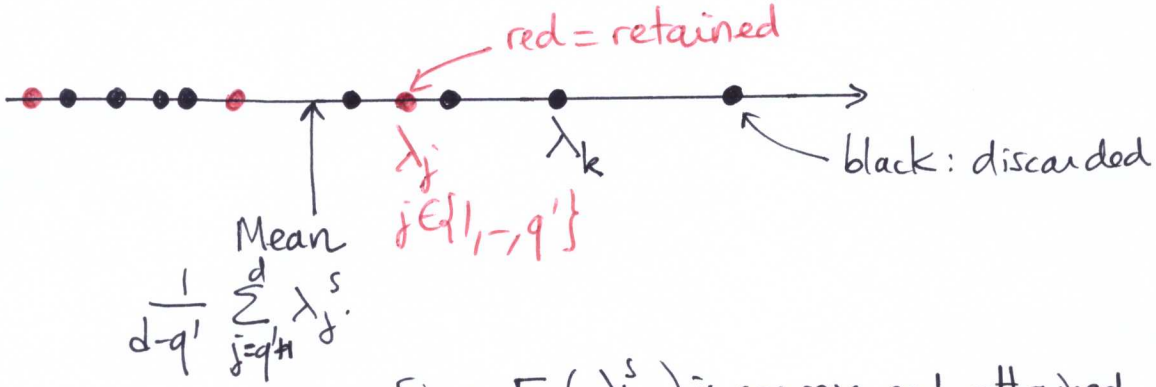
We consider the minimization of E by selecting appropriately the eigenvalues of S.

Suppose that the $(d-q')$ discarded eigenvalues have been chosen arbitrarily, and consider how a single value $\lambda_k^S \in \{\lambda_{q'+1}^S, \dots, \lambda_d^S\}$ affects the value of E:

$$\frac{\partial E}{\partial \lambda_k^S} = \frac{1}{\sum_{j=q'+1}^{d} \lambda_j^S} - \frac{1}{(d-q')\lambda_k^S}$$

$\Rightarrow E(\lambda_k^S)$ as a function of λ_k^S has a unique minimum at $\frac{1}{d-q'} \sum_{j=q'+1}^{d} \lambda_j^S$ = mean of the discarded eigenvalues.

Consider the following scenario:



Since $E(\lambda_k^S)$ is convex and attained its minimum at the mean of the discarded eigenvalues, if there is a λ_j^S, $j \in \{1, \dots, q'\}$ between λ_k^S and the mean, the swapping λ_k^S with λ_j^S will necessarily decrease E.

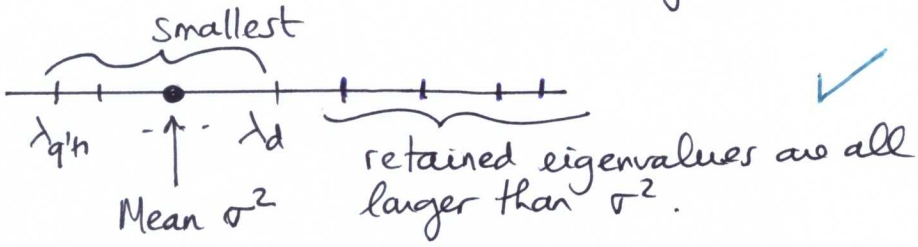
$\Rightarrow$ Unless the discarded eigenvalues $\lambda_{q'+1}^S, \dots, \lambda_d^S$ are chosen to be adjacent amongst the ordered eigenvalues of S, we can always swap two eigenvalues to decrease the value of E.

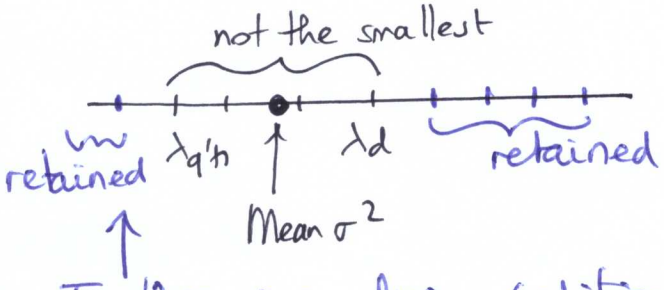
⇒ The discarded eigenvalues must be adjacent.
We now conclude that they need to be the smallest:

↘ Recall the expression of the MLE of σ^2 (top of page 25)

$$(*)\quad \sigma^2 = \frac{1}{d-q'} \sum_{j=q'+1}^{d} \lambda_j^S .$$

↘ The retained eigenvectors must have eigenvalues $\lambda_j^S > \sigma^2$ (corresponding to $\ell_j \neq 0$). In view of (*), this condition is violated unless the discarded eigenvalues are the smallest ones:





For this eigenvalue, condition $\lambda_j^S > \sigma^2$ is violated.

Conclusion: E must be minimized when $\lambda_{q'+1}^S, \ldots, \lambda_d^S$ are the smallest $(d-q')$ eigenvalues.

⇒ $\mathcal{L}$ is maximized when $\lambda_1^S, \ldots, \lambda_{q'}^S$ are the principal eigenvalues of $\underline{\underline{S}}$.

Also, E is further minimized as a function of q' when there are the fewest terms in the sum, which corresponds to $q' = r$; i.e. when no ℓ_j is zero.

Rk: $\mathcal{L}$ is minimized when $q' = 0$, ie when $\underline{\underline{W}} = 0$.

Summary: Consider the pPCA model

$$\begin{cases} \underline{x} = \underline{\underline{W}}\,\underline{z} + \underline{\mu} + \underline{\varepsilon} \\ \underline{z} \sim \mathcal{N}(0, I_r) \\ \underline{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_d) \end{cases} \qquad \begin{cases} x \in \mathbb{R}^d \\ z \in \mathbb{R}^r \\ W \in \mathbb{R}^{d \times r} \end{cases}$$

Let $\mathcal{L} = \{\underline{x}_1, \dots, \underline{x}_n\}$ be a sample of size n, and $\underline{\underline{S}} = \frac{1}{n} \sum_{i=1}^{n} (\underline{x}_i - \bar{x})(x_i - \bar{x})^t$ be the sample covariance matrix, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d > 0$.

The MLE of $\underline{\underline{W}}$ and σ^2 are given by

$$\begin{cases} \hat{W}_{ML} = V_r \left(\Lambda_r - \sigma^2 I \right)^{1/2} R \\ \hat{\sigma}^2_{ML} = \dfrac{1}{d-r} \displaystyle\sum_{j=r+1}^{d} \lambda_j , \end{cases}$$

where

- $\underset{(r \times r)}{\Lambda_r} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_r \end{pmatrix}$ = diagonal matrix of the r largest eigenvalues of S.
- $\underset{(d \times r)}{V_r}$ = matrix of the associated eigenvectors of S.
- $\underset{(r \times r)}{R}$ = an arbitrary orthogonal rotation matrix.

Remarks.

(i) $\hat{\sigma}^2_{ML} \equiv$ average variance 'lost' per discarded dimension.

(ii) Reduced-dimensionality transformation & reconstruction.

PCA

$z_r^j = V_r^t (x_j - \bar{x})$

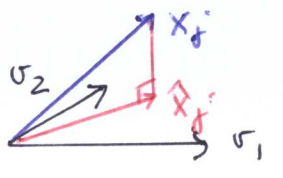
= reduced dim. trans.

z_r^j in $\mathbb{R}^r$, x_j in $\mathbb{R}^d$

$\hat{x}_j = V_r z_r^j + \bar{x}$

= reconstruction

= orthogonal projection of x_j onto the space spanned by $\{v_1, \dots, v_r\}$



pPCA

Posterior distribution is

$$z|x \sim \mathcal{N}(M^{-1}W^T(x-\mu), \sigma^2 M^{-1})$$

where

$$M = W^T W + \sigma^2 I$$

$(r \times r)$

$\Rightarrow$ An observation $x_j \in \mathbb{R}^d$ in the original space is summarized by its posterior mean in the latent space:

$$E(z_j|x_j) = M^{-1}\hat{W}_{ML}^T(x_j - \bar{x})$$

$(r \times 1)$ = reduced dim. trans.

$$\hat{x}_j = \hat{W}_{ML} E(z_j|x_j) + \bar{x}$$

= reconstruction from the latent variable.

Q: Is $\hat{x}_j$ = orthogonal projection of x_j onto the subspace spanned by $\{v_1, \dots, v_r\}$?

A: In general, NO. However:

$\rightarrow$ As $\sigma^2 \to 0$, the posterior mean

$$E(z_j|x_j) \to (\hat{W}_{ML}^T \hat{W}_{ML})^{-1} \hat{W}_{ML}^T (x_j - \bar{x})$$

so that in the limit,

$$\hat{x}_j = \hat{W}_{ML}(\hat{W}_{ML}^T \hat{W}_{ML})^{-1} \hat{W}_{ML}^T (x_j - \bar{x}) + \bar{x}$$

$= $ orthogonal projection of x_j onto the column space of $\hat{W}_{ML}$ = space spanned by $\{v_1, \dots, v_r\}$. We recover conventional PCA.

Note that as $\sigma^2 \to 0$ the posterior covariance is zero and the density becomes singular.

(B) EM algorithm for pPCA.

Alternatively, we can use the EM algorithm to maximize numerically the likelihood (potential advantage in the EM approach will be discussed later).

- The complete data log-likelihood is:

$$\mathcal{L}_c = \sum_{i=1}^{n} \ln f(\underline{x}_i, \underline{z}_i),$$

where

$$f(\underline{x}_i, \underline{z}_i) = f(\underline{x}_i | \underline{z}_i) f(\underline{z}_i)$$

$$= \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{-\frac{1}{2\sigma^2}\|x_i - Wz_i - \mu\|^2\right\} \frac{1}{(2\pi)^{r/2}} \exp\left\{-\frac{1}{2} z_i^t z_i\right\}$$

$$\Rightarrow \mathcal{L}_c = \sum_{i=1}^{n} \left\{-\frac{d}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|x_i - Wz_i - \mu\|^2 - \frac{r}{2}\ln 2\pi - \frac{1}{2} z_i^t z_i\right\}$$

Omitting terms independent of the model parameters:

$$\mathcal{L}_c = -\sum_{i=1}^{n} \left\{\frac{d}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}\|x_i - Wz_i - \mu\|^2 + \frac{1}{2} z_i^t z_i\right\}$$

$$= -\sum_{i=1}^{n} \left\{\frac{d}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}\|x_i - \mu\|^2 + \frac{1}{2\sigma^2}\|Wz_i\|^2 - \frac{1}{\sigma^2}\langle Wz_i, x_i - \mu\rangle + \frac{1}{2} z_i^t z_i\right\},$$

where
- $\|Wz_i\|^2 = \langle Wz_i, Wz_i\rangle = z_i^t W^t W z_i$
  $= \mathrm{Tr}(z_i^t W^t W z_i)$
  $= \mathrm{Tr}(W^t W z_i z_i^t)$
- $z_i^t z_i = \mathrm{Tr}(z_i^t z_i) = \mathrm{Tr}(z_i z_i^t)$

$\Rightarrow$

$$\boxed{\mathcal{L}_c = -\sum_{i=1}^{n} \left\{\frac{d}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}\|x_i - \mu\|^2 + \frac{1}{2\sigma^2}\mathrm{Tr}(W^t W z_i z_i^t) - \frac{1}{\sigma^2} z_i^t W^t (x_i - \mu) + \frac{1}{2}\mathrm{Tr}(z_i z_i^t)\right\}}$$

- **E-step.**

We need to compute the expected value of the complete log-likelihood, keeping the observed variables fixed, with respect to the conditional distribution $p(\underline{z}|\underline{x})$.

Here $p(\underline{z}|\underline{x}) \sim \mathcal{N}(M^{-1}W^t(x-\mu), \sigma^2 M^{-1})$

where
$M = W^tW + \sigma^2 I$.
$(r \times r)$

Thus

Computed with the 'old' values of the parameters

$\rightarrow \langle z_i \rangle \equiv E_{p(\underline{z}_i|\underline{x}_i)} \underline{z}_i = M^{-1}W^t(x_i - \mu)$

$\rightarrow \langle z_i z_i^t \rangle \equiv E_{p(\underline{z}_i|\underline{x}_i)}(\underline{z}_i \underline{z}_i^t) = \underbrace{\sigma^2 M^{-1}}_{\text{Covariance matrix}} + \langle z_i \rangle \langle z_i \rangle^t$

$\Rightarrow$

$$E_{p(z|x)} \mathcal{L}_c = \langle \mathcal{L}_c \rangle = -\sum_{i=1}^{n} \left\{ \frac{d}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}\|x_i - \mu\|^2 + \frac{1}{2\sigma^2} Tr\left(W^tW\langle z_i z_i^t\rangle\right) - \frac{1}{\sigma^2}\langle z_i^t\rangle W^t(x_i - \mu) + \frac{1}{2} Tr\left(\langle z_i z_i^t\rangle\right) \right\}$$

- **M-step.** $\langle \mathcal{L}_c \rangle$ is maximized with respect to W and σ^2.

$$\frac{\partial \langle \mathcal{L}_c \rangle}{\partial W} = -\sum_{i=1}^{n} \left\{ \frac{1}{2\sigma^2} \frac{\partial}{\partial W} Tr(W^tW\langle z_i z_i^t\rangle) - \frac{1}{\sigma^2}\frac{\partial}{\partial W}\langle z_i\rangle^t W^t(x_i - \mu) \right\}$$

Useful: $\frac{\partial}{\partial W} Tr(W^tWA) = WA^t + WA$
$(r \times d)(d \times r)(r \times r)$

$\frac{\partial}{\partial W}(u^tW^tv) = vu^t$
$(1 \times r)(r \times d)(d \times 1) \quad (d \times 1)(1 \times r)$

$W \in \mathbb{R}^{d \times r}$

$$\Rightarrow \frac{\partial \langle \mathcal{L}_c \rangle}{\partial W} = -\sum_{i=1}^{n} \left\{ \frac{2}{2\sigma^2} W \langle z_i z_i^t \rangle - \frac{1}{\sigma^2}(x_i - \mu)\langle z_i \rangle^t \right\} = 0 \quad (32)$$

$$\Rightarrow \sum_{i=1}^{n} \tilde{W} \langle z_i z_i^t \rangle = \sum_{i=1}^{n} (x_i - \mu)\langle z_i \rangle^t$$

$$\boxed{\tilde{W} = \left(\sum_{i=1}^{n} (x_i - \mu)\langle z_i \rangle^t \right) \left(\sum_{i=1}^{n} \langle z_i z_i^t \rangle \right)^{-1}}$$

Next,

$$\frac{\partial \langle \mathcal{L}_c \rangle}{\partial \sigma^2} = -\sum_{i=1}^{n} \left\{ \frac{d}{2\sigma^2} - \frac{1}{2\sigma^4} \|x_i - \mu\|^2 - \frac{1}{2\sigma^4} Tr(W^t W \langle z_i z_i^t \rangle) + \frac{1}{\sigma^4} \langle z_i \rangle^t W^t (x_i - \mu) \right\} = 0$$

$$\Rightarrow \tilde{\sigma}^2 nd = \sum_{i=1}^{n} \left\{ \|x_i - \mu\|^2 + Tr(\tilde{W}^t \tilde{W} \langle z_i z_i^t \rangle) - 2\langle z_i \rangle^t \tilde{W}^t (x_i - \mu) \right\}.$$

$$\boxed{\tilde{\sigma}^2 = \frac{1}{nd} \sum_{i=1}^{n} \left\{ \|x_i - \mu\|^2 + Tr(\langle z_i z_i^t \rangle \tilde{W}^t \tilde{W}) - 2\langle z_i \rangle^t \tilde{W}^t (x_i - \mu) \right\}.}$$

↳ EM algorith Version #1.

- Initialise W, σ^2.
- Repeat until convergence

E-step:
$$\langle z_i \rangle = M^{-1} W^t (x_i - \mu)$$
$$\langle z_i z_i^t \rangle = \sigma^2 M^{-1} + \langle z_i \rangle \langle z_i^t \rangle$$

M-step:
$$\tilde{W} = \left(\sum_{i=1}^{n} (x_i - \mu)\langle z_i \rangle^t \right) \left(\sum_{i=1}^{n} \langle z_i z_i^t \rangle \right)^{-1}$$
$$\tilde{\sigma}^2 = \frac{1}{nd} \sum_{i=1}^{n} \left\{ \|x_i - \mu\|^2 + Tr(\langle z_i z_i^t \rangle \tilde{W}^t \tilde{W}) - 2\langle z_i \rangle^t \tilde{W}^t (x_i - \mu) \right\}$$

$$W \leftarrow \tilde{W}$$
$$\sigma \leftarrow \tilde{\sigma}$$

Simplifications:

- $$\tilde{W} = \left(\sum_{i=1}^{n} (x_i - \mu)\langle z_i \rangle^t \right) \left(\sum_{i=1}^{n} \langle z_i z_i^t \rangle \right)^{-1}$$

$$= \left(\sum_{i=1}^{n} (x_i - \mu) \underbrace{(x_i - \mu)^t W M^{-1}}_{= \langle z_i \rangle^t} \right) \left(\sum_{i=1}^{n} \left\{ \sigma^2 M^{-1} + M^{-1} W^t (x_i - \mu)(x_i - \mu)^t W M^{-1} \right\} \right)^{-1}$$

$$= \frac{1}{n} \left\{ \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^t \right\} W M^{-1} \left(\sigma^2 M^{-1} + M^{-1} W^t \left\{ \frac{1}{n} \Sigma (x_i - \mu)(x_i - \mu)^t \right\} W M^{-1} \right)^{-1}$$

$$= S W M^{-1} \left(\sigma^2 M^{-1} + M^{-1} W^t S W M^{-1} \right)^{-1}.$$

$$\boxed{\tilde{W} = S W \left(\sigma^2 I + M^{-1} W^t S W \right)^{-1}}$$

$\equiv$ fusion of E and M steps for $\tilde{W}$.

- $$\tilde{\sigma}^2 = \frac{1}{nd} \sum_{i=1}^{n} \left\{ \mathrm{Tr}(x_i - \mu)(x_i - \mu)^t + \mathrm{Tr}\left(\langle z_i z_i^t \rangle \tilde{W}^t \tilde{W} \right) - 2 \langle z_i \rangle^t \tilde{W}^t (x_i - \mu) \right\}$$

These are computed with the 'old' versions of W and σ^2.

$$= \frac{1}{d} \left\{ \mathrm{Tr}\, S + \frac{1}{n} \sum_{i=1}^{n} \mathrm{Tr}\left(\left[\sigma^2 M^{-1} + M^{-1} W^t (x_i - \mu)(x_i - \mu)^t W M^{-1} \right] \tilde{W}^t \tilde{W} \right) \right.$$

$$\left. - \frac{2}{n} \sum_{i=1}^{n} \underbrace{(x_i - \mu)^t W M^{-1} \tilde{W}^t (x_i - \mu)}_{\| \; \mathrm{Tr}\left((x_i - \mu)^t W M^{-1} \tilde{W}^t (x_i - \mu) \right)} \right\}$$

$$= \mathrm{Tr}\left(W M^{-1} \tilde{W}^t (x_i - \mu)(x_i - \mu)^t \right)$$

$$\Rightarrow \frac{2}{n} \sum_{i=1}^{n} (\ldots) = 2\, \mathrm{Tr}\left(W M^{-1} \tilde{W}^t S \right) = 2\, \mathrm{Tr}\left(S W M^{-1} \tilde{W}^t \right)$$

$(d \times d)(d \times r)(r \times r)(r \times d)$

$$= \mathrm{Tr}\, \sigma^2 M^{-1} \tilde{W}^t \tilde{W}$$

$$+ \mathrm{Tr}\, M^{-1} W^t S W M^{-1} \tilde{W}^t \tilde{W}$$

$$= \mathrm{Tr}\left\{ \tilde{W} \left(\sigma^2 M^{-1} + M^{-1} W^t S W M^{-1} \right) \tilde{W}^t \right\}$$

$$= \mathrm{Tr}\left\{ S W \left(\sigma^2 I + M^{-1} W^t S W \right)^{-1} \left(\sigma^2 I + M^{-1} W^t S W \right) M^{-1} \tilde{W}^t \right\}$$

$$= \mathrm{Tr}\left\{ S W M^{-1} \tilde{W}^t \right\}$$

Putting everything together, the M-step for $\tilde{\sigma}^2$ simplifies to:

$$\tilde{\sigma}^2 = \frac{1}{d}\left[\mathrm{Tr}\,S + \mathrm{Tr}\{ S W M^{-1} \tilde{W}^t \} - 2\,\mathrm{Tr}\{ S W M^{-1} \tilde{W}^t \}\right]$$

$$\boxed{\tilde{\sigma}^2 = \frac{1}{d}\mathrm{Tr}\left(S - S W M^{-1} \tilde{W}^t \right)}$$

$\equiv$ fusion of E and M steps for $\tilde{\sigma}^2$.

↳ EM algorithm Version #2.

- Initialise W, σ^2
- Repeat until convergence:

$$\tilde{W} = S W \left(\sigma^2 I + M^{-1} W^t S W\right)^{-1}$$

$$\tilde{\sigma}^2 = \frac{1}{d}\mathrm{Tr}\left(S - S W M^{-1} \tilde{W}^t \right)$$

$$W \leftarrow \tilde{W}$$

$$\sigma \leftarrow \tilde{\sigma}$$

pPCA

Remarks.

(i) Data enters into the EM algorithm only through the covariance matrix $\underline{\underline{S}}$ (compare with the analytical solution).

(ii) $\underline{\underline{S}} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^t$ → computed in $O(nd^2)$ operations.

$(x_i - \bar{x})$: $(d \times 1)$, $(x_i - \bar{x})^t$: $(1 \times d)$

$O(d^2)$ operations

Computing $\underline{\underline{S}}$ may be expensive & so time may be saved by not computing $\underline{\underline{S}}$ directly, as suggested by the EM algorithm.

↳ S appears through $\mathrm{Tr}\,S$ and SW only.

- $\mathrm{Tr}\, S = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Tr}\,(x_i - \mu)(x_i - \mu)^t$

$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{\underbrace{(x_i - \mu)^t}_{(1\times d)} \underbrace{(x_i - \mu)}_{(d\times 1)}}_{d \text{ operations, repeated } n \text{ times}}$ → $O(nd)$ operations. (S not computed)

- $SW = \frac{1}{n} \left\{ \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^t \right\} W$

$\underbrace{\qquad\qquad\qquad\qquad}_{(d\times d)} \quad \underbrace{\quad}_{(d\times r)}$

$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{O(d^2 r)}$

$\Rightarrow O(d^2 r) + \underbrace{O(nd^2)}_{\text{computation of } S}$, to compute SW this way.

However,

$$SW = \frac{1}{n} \sum_{i=1}^{n} \underbrace{(x_i - \mu)}_{(d\times 1)} \Big[\underbrace{(x_i - \mu)^t W}_{(1\times d)(d\times r)} \Big] \quad \text{(S not computed)}$$

$(1\times d)(d\times r) \to O(dr)$

$(d\times 1)\,O(dr) \to O(dr) + O(dr)$

$\to O(ndr)$

Total of $O(ndr)$ operations ⇒ If $r \ll d$, then computing SW without explicitely computing S is more efficient.

- Analytical solution requires the computation of $\underline{\underline{S}}$.
- We might get away in the EM algorithm without direct computation of $\underline{\underline{S}}$,
  ⇒ Iterative approach MAY be more efficient.

Why 'may': because $\underline{\underline{S}}$ is computed only once, while the efficient way of calculating SW requires its evaluation at each iteration of the algorithm ⇒ potential efficiency gain depends on the number of iterations needed to converge.

III_ KERNEL PCA

• Conventional PCA is looking for direction in the data with maximum variance. The directions are given by the eigenvectors of the covariance matrix Σ. Assuming the data is centered, the eigenvectors can be computed from the sample covariance matrix

$$\underset{(d\times d)}{S} = \frac{1}{n}\underset{(d\times n)}{X^t}\underset{(n\times d)}{X} = \frac{1}{n}\sum_{j=1}^{n} x_j x_j^t .$$

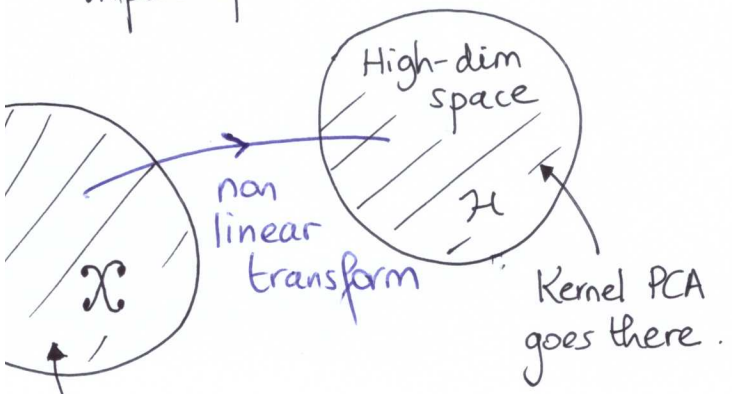
Eigenvalue-eigenvector pair (λ, v) of S satisfy $Sv = \lambda v$;

$$\frac{1}{n}\sum_{i=1}^{n} x_i \underbrace{x_i^t v}_{=\langle x_i, v\rangle} = \lambda v \quad \Rightarrow \quad v = \frac{1}{n\lambda}\sum_{i=1}^{n}\langle x_i, v\rangle x_i = \sum_{i=1}^{n}\alpha_i x_i$$

$\langle x_i, v\rangle$ = a dot product !

ie all solutions lie in the span of $x_1, \dots, x_n$.

• Kernel PCA is a non-linear generalization of conventional PCA: PCA is performed in a high-dimensional space $\mathcal{H}$ (possibly of infinite dimension), which is non-linearly related to the original input space $\mathcal{X}$.



The technique used is that of the 'kernel trick', that we encountered when we discussed kernel SVM: the kernel substitution allows us to take an algorithm expressed in terms of dot products $x^t x$ and generalize it by replacing the dot products with a kernel.

⇒ The new space we live in is an RKHS, with reproducing kernel $K(\cdot,\cdot)$. ($\mathcal{H}$ = our $\mathcal{H}$)

⇒ Original features x_i are now transformed using the feature map Φ: $\Phi(x_i)$ (Φ not unique)

$1 \le i \le n$

However, the feature map is not unique, and we want to express the principal components in $\mathcal{H}$ in terms of $K \Rightarrow$ we need to formulate the problem in terms of dot products in $\mathcal{H}$. (recall that $K(x,y) = \Phi^t(x)\,\Phi(y)$; the feature map can be extracted using Mercer theorem for example)

$\Phi(x_i)$ — D-dimensional

For the moment, assume that the new features $\Phi(x_i)$ are centered:

$$\sum_{i=1}^{n} \Phi(x_i) = 0$$ ← We return to this later

The sample covariance matrix S in $\mathcal{H}$ is $S = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i)\,\Phi(x_i)^t$.

$(D \times D)$ $(D \times 1)$ $(1 \times D)$

Its eigenvector expansion is $\boxed{S v_k = \lambda_k v_k}$ ← $k = 1, \ldots, M$.

As mentioned before, we are looking for solution to this eigenvalue-eigenvector problem without working explicitely in the feature space; i.e. working with kernels

From the definition of S, we get

$$\frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \boxed{\Phi(x_i)^t v_k} = \lambda_k v_k \quad \text{—— (*)}$$

$\boxed{\Phi(x_i)^t v_k}$ — a scalar

$\Rightarrow v_k$ can be expressed as a linear combination of the features:

$$v_k = \sum_{j=1}^{n} \alpha_{kj} \Phi(x_j)$$

Substituting this expression back into (*) yields

$$\frac{1}{n} \sum_{i,j} \alpha_{ki} \Phi(x_i) \underbrace{\Phi(x_i)^t \Phi(x_j)}_{K(x_i, x_j)} = \lambda_k \sum_{j=1}^{n} \alpha_{kj} \Phi(x_j)$$

← We are on the right track!

$\Phi^t(x_m)$ for some m

$$\sum_{i,j} \alpha_{ki} \underbrace{\Phi^t(x_m)\Phi(x_i)}\, K(x_i, x_j) = n \lambda_k \sum_{j=1}^{n} \alpha_{kj} \underbrace{\Phi^t(x_k)\,\Phi(x_j)}$$

$$\Rightarrow \quad \sum_{i,j} \alpha_{ki}\, K(x_m, x_i) K(x_i, x_j) = n\lambda_k \sum_j \alpha_{kj} K(x_k, x_j)$$

We can rewrite this equation in matrix notation, noticing that the m-th line of K^2 is

$$K^2_{mj} = \sum_{i=1}^{n} K(x_m, x_i) K(x_j, x_i)$$

j-th column

We obtain $\boxed{K^2 \alpha_k = n\lambda_k K \alpha_k}$ (*), where $\alpha_k = \begin{pmatrix} \alpha_{k1} \\ \vdots \\ \alpha_{kn} \end{pmatrix} \in \mathbb{R}^n$

Solution to this problem can be found by solving the following eigenvalue problem

$$\boxed{K \alpha_k = n \lambda_k \alpha_k} \quad (**)$$

← We have removed a factor K on both sides.

Clearly, solutions to this problem are also solutions to $K^2 \alpha_k = n\lambda_k K \alpha_k$. Moreover, we will show shortly that for any solution to (**) having eigenvalue λ_k, we can add any multiple of an eigenvector of K having 0 eigenvalue, and obtain a solution to (*) that also has eigenvalue λ_k. We also show that these modifications do not affect the principal components, that we shortly derive. Good.

As for conventional PCA, we normalize the solution so that $v_k^t v_k = 1$

ie $\left(\sum_{i=1}^{n} \alpha_{ki} \Phi(x_i) \right)^t \left(\sum_{i=1}^{n} \alpha_{ki} \Phi(x_i) \right) = 1$

$$\sum_{i,j} \alpha_{ki} \alpha_{kj} \Phi(x_i)^t \Phi(x_j)$$

$$\sum_{i,j} \alpha_{ki} \alpha_{kj} K(x_i, x_j) = \alpha_k^t K \alpha_k = 1$$

From (**) we get that $K\alpha_k = n\lambda_k \alpha_k$, so that

$$n\lambda_k \alpha_k^t \alpha_k = 1$$

Summary: (i) Solve the eigenvalue problem for $K \rightarrow (\mu_k, \alpha_k)$ $k = 1, \dots, D$

(ii) Eigenvalues of S satisfy $\mu_k = n\lambda_k$

(iii) Normalize eigenvectors s.t. $n\lambda_k \alpha_k^t \alpha_k = 1$.

(iv) Compute v_k using $v_k = \sum_{j=1}^{n} \alpha_{kj} \Phi(x_j)$

Once the eigenvalue problem is solved, the resulting PC can also be expressed in term of K:

⇒ we never do step (iv) in practice since

→ projection of a point x onto the eigenvector v_k is:

$$\Phi(x)^t v_k = \sum_{j=1}^{n} \alpha_{kj} \Phi(x)^t \Phi(x_j)$$

$$z_k(x) := \sum_{j=1}^{n} \alpha_{kj} K(x, x_j)$$

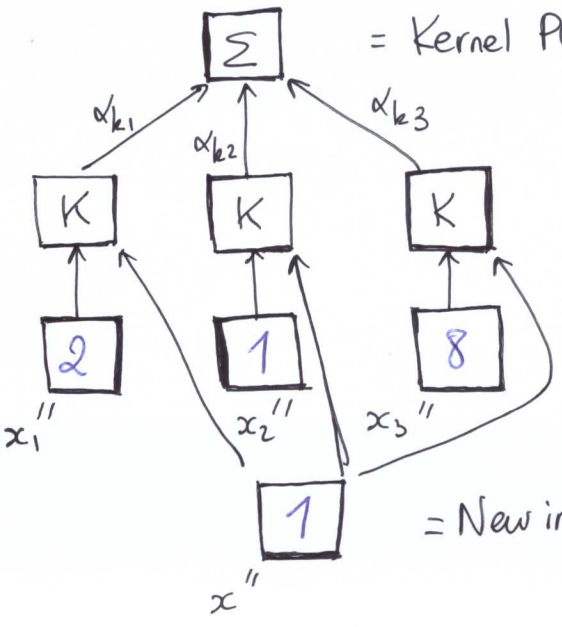
k-th KERNEL PC.

$z_k(x)$ = k-th principal component of a new input vector x.

⇒ Replace step (iv) in the summary above by:

(iv) Compute the k-th PC by $z_k(x) = \sum_{j=1}^{n} \alpha_{kj} K(x, x_j)$

A weighted linear combination of comparisons of the new input vector x with training features $x_1, \dots, x_n$.



Once the features are extracted, you can use your favorite classifier on the kernel PC

Remarks: (i) As mentioned before, the solutions of (*) and (**) differ only by eigenvectors of K having 0 eigenvalues that do not affect the kernel PC:

(*) $K^2 \alpha_k = n \lambda_k K \alpha_k$ → let $\tilde{a}_k$ be a solution with eigenvalue λ_k

(**) $K \alpha_k = n \lambda_k \alpha_k$ → let a_k be a solution with the same λ_k.

Write $\tilde{a}_k = a_k + b_k$. Then, with $Kb_k = 0$,

$$K a_k = n \lambda_k a_k \Rightarrow K \tilde{a}_k = n \lambda_k \tilde{a}_k$$
$$\Rightarrow K^2 \tilde{a}_k = n \lambda_k K \tilde{a}_k$$

$$K b_k = 0 \Rightarrow \sum_{j=1}^{n} b_{kj} K(x, x_j) = 0 \quad \forall x \;\&$$

$$\Phi^t(x) \tilde{v}_k = \sum_j \tilde{a}_{kj} \Phi^t(x) \Phi(x_j) = \sum_j a_{kj} K(x, x_j) = z_k(x$$

(ii) The data is usually not centered in $\mathcal{H}$. We do this manually:

$$\tilde{\Phi}(x_i) = \Phi(x_i) - \frac{1}{n} \sum_{\ell=1}^{n} \Phi(x_\ell).$$

Put $\tilde{K}_{ij} = \tilde{\Phi}^t(x_i) \tilde{\Phi}^t(x_j)$

The eigenvalue problem becomes $\tilde{K} \tilde{\alpha}_k = n \lambda_k \tilde{\alpha}_k$.

$$\tilde{K}_{ij} = \left(\Phi(x_i) - \frac{1}{n} \sum_{\ell=1}^{n} \Phi(x_\ell) \right)^t \left(\Phi(x_j) - \frac{1}{n} \sum_{\ell=1}^{n} \Phi(x_\ell) \right)$$

$$= K_{ij} - \frac{1}{n} \sum_\ell \Phi(x_i)^t \Phi(x_\ell) - \frac{1}{n} \sum_\ell \Phi(x_\ell)^t \Phi(x_j) + \frac{1}{n^2} \sum_{\ell,m} \Phi(x_\ell)^t \Phi(x_m)$$

$$\Rightarrow \tilde{K}_{ij} = K_{ij} - \frac{1}{n} \sum_\ell K_{i\ell} - \frac{1}{n} \sum_\ell K_{j\ell} + \frac{1}{n^2} \sum_{\ell,m} K_{\ell,m}$$

↑ Symmetric!

In matrix form,

$$\tilde{K} = K - 1_{\frac{1}{n}} K - K 1_{\frac{1}{n}} + 1_{\frac{1}{n}} K 1_{\frac{1}{n}},$$

where $1_{\frac{1}{n}} = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots \\ \frac{1}{n} & & \\ \vdots & & \end{pmatrix}$.

$\Rightarrow$ In practice, pick a kernel K, compute $\tilde{K}$, solve the eigenvalue problem $\tilde{K}\tilde{\alpha}_k = n\lambda_k \tilde{\alpha}_k$, and compute the k-th kernel PC.

(iii) The eigenvalue problem $K\alpha_k = n\lambda_k \alpha_k$ reduces to the conventional PCA eigenvalue problem $S v_k = \lambda_k v_k$, for the linear kernel $K(x, y) = x^t y$. Indeed,

Feature map $\Phi(x) = x$

$$(K\alpha_k)_i = \sum_{j=1}^{n} \alpha_{kj} K(x_i, x_j) = \sum_{j=1}^{n} \alpha_{kj} x_i^t x_j.$$

i-th row of $K\alpha_k$

Plugging this expression back into $K\alpha_k = n\lambda_k \alpha_k$ yields:

$$\sum_{j=1}^{n} \alpha_{kj} x_i^t x_j = n\lambda_k \alpha_{ki}$$

$(\times x_i)$

$$\sum_{j=1}^{n} \alpha_{kj} x_i x_i^t x_j = n\lambda_k \alpha_{ki} x_i$$

sum over i

$$\sum_{j=1}^{n} \left(\sum_{i=1}^{n} x_i x_i \right)^t \alpha_{kj} x_j = n\lambda_k \left(\sum_{i=1}^{n} \alpha_{ki} x_i \right)$$

$$n S \underbrace{\left(\sum_{j=1}^{n} \alpha_{kj} x_j \right)}_{:= v_k} = n\lambda_k \underbrace{\left(\sum_{i=1}^{n} \alpha_{ki} x_i \right)}_{:= v_k}$$

$z_k(x) = \Phi(x)^t v_k = x^t v_k = k\text{-th PC}.$

$\Rightarrow \quad S v_k = \lambda_k v_k$, as required.

(iv) A disadvantage of kernel PCA is that it involves finding the eigenvectors of a $(D \times D)$ matrix, rather than an $(d \times d)$ matrix for conventional PCA.

(v) In conventional PCA, the number of PC cannot exceed d.
In kernel PCA, the dimension of the feature space is usually much larger than d.
$\Rightarrow$ we can find a number of non-linear PC which exceeds the dimension of $\underline{x}$.

(vi) Schölkopf, Smola & Müller used kernel PCA to train an SVM classifier on hand-written digit data, and showed that they can improve the classification performance compared with linear PCA.