Ex 1

(1) $$X = \left(\vec{x}_1 \; \dots \; \vec{x}_n \right) \Big\}_p = \begin{pmatrix} X_{11} & \dots & X_{1n} \\ X_{21} & & \\ \vdots & & \\ X_{p1} & \dots & X_{pn} \end{pmatrix},$$

where X_{ij} i is the index for the i-th component of the random vector $\vec{x}$, and j is the j-the observation.

We see that the first row of matrix X contains all n observations of the 1st random variable.

Let $V_i^T = (X_{i1} \; X_{i2} \dots X_{in})$, hence

$$X = \begin{pmatrix} V_1^T \\ \vdots \\ V_p^T \end{pmatrix}$$

The rows of X span thus a p-dimensional subspace of $\mathbb{R}^n$.

(2) $\operatorname{cov}(X_1, X_2) = E(x_1 x_2)$ by assumption of zero mean of $\vec{x}$.

Sample version is $\frac{1}{n} \sum_{i=1}^{n} X_{1i} X_{2i} = \frac{1}{n} V_1^T V_2$.

The covariance matrix C is $E(xx^T)$, (for $E(x) = 0$)

which equals

$$E(xx^T) = \begin{pmatrix} E(x_1x_1) & E(x_1x_2) & \dots & E(x_1x_p) \\ \vdots & & & \\ E(x_px_1) & E(x_px_2) & \dots & E(x_px_p) \end{pmatrix}.$$

A sample version $\hat{C}$ is thus

$$\hat{C} = \frac{1}{n} \begin{pmatrix} V_1^T V_1 & \dots & V_1^T V_p \\ \vdots & & \\ V_p^T V_1 & \dots & V_p^T V_p \end{pmatrix} = \frac{1}{n} X X^T$$

(3) $\vec{z} = U^T \vec{x}$

$$Z = U^T X = U^T (\vec{x}_1 \ldots \vec{x}_n) = (U^T \vec{x}_1 \ldots U^T \vec{x}_n)$$

$$U^T = \begin{pmatrix} \vec{u}_1^T \\ \vdots \\ \vec{u}_m^T \end{pmatrix} \longrightarrow = \begin{pmatrix} \vec{u}_1^T \vec{x}_1 & \ldots & \vec{u}_1^T \vec{x}_n \\ \vec{u}_2^T \vec{x}_1 & & \vdots \\ \vdots & & \vdots \\ \vec{u}_m^T \vec{x}_1 & & \vec{u}_m^T \vec{x}_n \end{pmatrix}$$

$z_i = u_i^T \vec{x}$, which is the ith principal componect.

The i-th row of Z contains thus all the realizations of the i-th principal component, which also often called the i-th principal component. Language is often not so precise here.

(4) The i-th row of Z is $u_i^T X$. Taking the scalar product with $j \neq i$-th row gives

$$u_i^T X X^T u_j = n\, u_i^T \left(\tfrac{1}{n} X X^T\right) u_j \overset{\text{by (2)}}{=} n\, u_i^T \hat{C} u_j .$$

By assumption $\hat{C} = U D U^T$, thus

$$n\, \underbrace{u_i^T U}\, D\, \underbrace{U^T u_j} = n\, (0 \ldots \underset{\substack{\uparrow \\ i\text{-th slot}}}{1} \ldots 0) \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_p \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \text{ j-th slot}$$

$$= n\, (0 \ldots d_i \ldots 0) \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \text{ j-th slot}$$

$$= 0 .$$

(5) The principal components are an orthogonal basis for the data space.

Ex. 2

(1) $Cx \underset{\substack{\uparrow \\ \lambda \text{ Eigenvalue}}}{=} \lambda x \Rightarrow (C - I\lambda)x = 0 \Rightarrow \det(C - I\lambda) = 0$

$$\det(C - I\lambda) = \det \begin{pmatrix} 1-\lambda & \rho \\ \rho & 1-\lambda \end{pmatrix} = (1-\lambda)^2 - \rho^2 \overset{!}{=} 0$$

$\Rightarrow 1 - \lambda = \pm \rho$

$\Rightarrow \underline{\lambda = 1 \pm \rho}$

If $|\rho| \to 1$, then one $\lambda \to 0$, i.e. if they are highly correlated, we get one small eigenvalue.

(2) $V(x_2) = V(ax_1 + n) \underset{x_1, n \text{ indep.}}{=} a^2 V(x_1) + V(n) = a^2 + V(n) \overset{!}{=} 1 \quad (*)$

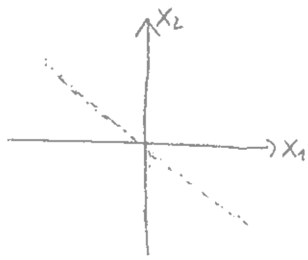
$$\operatorname{cov}(x_1 x_2) \underset{\substack{\uparrow \\ E(x_1)=0}}{=} E(x_1 x_2) = E(x_1(ax_1 + n)) = a \underbrace{E(x_1^2)}_{1} + \underbrace{E(x_1)}_{0} E(n) = \rho$$

$\Rightarrow \underline{a = \rho}$

$\Rightarrow (*) \quad \rho^2 + V(n) = 1 \iff \underline{V(n) = 1 - \rho^2}$

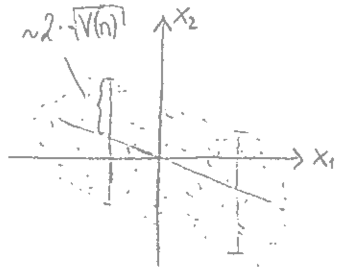
(3) $\rho = -1$

$V(n) = 0$



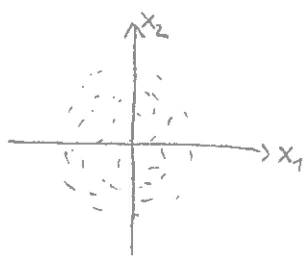
$\rho = -0.25$

$V(n) = 0.9375$



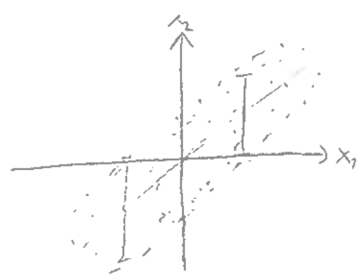
$\rho = 0$

$V(n) = 1$



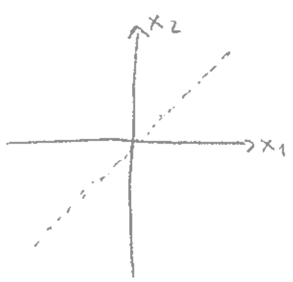
$\rho = 0.5$

$V(n) = 0.75$



$\rho = 1$

$V(n) = 0$



(4) $X = \begin{pmatrix} x_{11} & \dots & x_{1n} \\ x_{21} & \dots & x_{2n} \end{pmatrix} = \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix}$ $\qquad X^T = (v_1, v_2)$

If $|\varrho| = 1$ then the variance of the noise variable is 0; x_2 is deterministically related to x_1 ($x_2 := x_1$), hence $v_1 = v_2$ linearly dependent.

Generally, if $|\varrho|$ is close to 1, v_1 and v_2 are „close to linearly dependent".

The conditioning number of C is given by $\frac{\lambda_{max}}{\lambda_{min}} = \frac{1+|\varrho|}{1-|\varrho|}$. This becomes arbitrary large as $|\varrho| \to 1$.

The conditioning number of X^T is a measure of the linear dependencies of v_1 and v_2.
For any matrix M (not necessarily square) the conditioning number is defined as

$$\mathrm{cond}(M) = \sqrt{\frac{\text{biggest Eigenvalue of } M^TM}{\text{smallest Eigenvalue of } M^TM}}$$

In our case $M = X^T$ so that $M^TM = XX^T = n \cdot C$ (C covariance matrix)
The conditioning number of X^T is thus

$$\mathrm{cond}(X^T) = \sqrt{\frac{n \cdot \lambda_{max}}{n \cdot \lambda_{min}}} = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = \sqrt{\frac{1+|\varrho|}{1-|\varrho|}} = \sqrt{\mathrm{cond}(C)}$$

If $|\varrho| \to 1$, this clearly shows that the conditioning number of X^T goes up, that is v_1 and v_2 become „more linearly dependent".

Ex. 3

(1) We show that the columns are linearly dependent:

$$\underbrace{\cos\alpha \begin{pmatrix} 1 \\ 0 \\ \cos\alpha \end{pmatrix}}_{1^{st}\text{ column}} + \underbrace{\sin\alpha \begin{pmatrix} 0 \\ 1 \\ \sin\alpha \end{pmatrix}}_{2^{nd}\text{ column}} = \begin{pmatrix} \cos\alpha \\ \sin\alpha \\ \cos^2\alpha + \sin^2\alpha \end{pmatrix} = \underbrace{\begin{pmatrix} \cos\alpha \\ \sin\alpha \\ 1 \end{pmatrix}}_{3^{rd}\text{ column}}.$$

(2) $$Cu_1 = \frac{1}{\sqrt{2}} \cdot \begin{pmatrix} 1 & 0 & \cos\alpha \\ 0 & 1 & \sin\alpha \\ \cos\alpha & \sin\alpha & 1 \end{pmatrix} \begin{pmatrix} \cos\alpha \\ \sin\alpha \\ 1 \end{pmatrix} = \frac{1}{\sqrt{2}} \cdot \begin{pmatrix} 2\cos\alpha \\ 2\sin\alpha \\ \underbrace{\cos^2\alpha + \sin^2\alpha}_{1} + 1 \end{pmatrix} = 2 \cdot \underbrace{\frac{1}{\sqrt{2}} \begin{pmatrix} \cos\alpha \\ \sin\alpha \\ 1 \end{pmatrix}}_{u_1}$$

since $Cu_1 = \lambda_1 u_1 \Rightarrow \underline{\lambda_1 = 2}$

$$Cu_2 = \begin{pmatrix} 1 & 0 & \cos\alpha \\ 0 & 1 & \sin\alpha \\ \cos\alpha & \sin\alpha & 1 \end{pmatrix} \begin{pmatrix} -\sin\alpha \\ \cos\alpha \\ 0 \end{pmatrix} = \begin{pmatrix} -\sin\alpha \\ \cos\alpha \\ -\cos\alpha\sin\alpha + \cos\alpha\sin\alpha \end{pmatrix} = 1 \cdot \underbrace{\begin{pmatrix} -\sin\alpha \\ \cos\alpha \\ 0 \end{pmatrix}}_{u_2}$$

$\longrightarrow \underline{\lambda_2 = 1}$

$Cu_3 \overset{!}{=} 0 \cdot u_3 = 0$: $\quad Cu_3 = \frac{1}{\sqrt{2}} \begin{pmatrix} -\cos\alpha + \cos\alpha \\ -\sin\alpha + \sin\alpha \\ -\cos^2\alpha - \sin^2\alpha + 1 \end{pmatrix} = 0$, $u_3 \neq 0$

(3) Use the formula from math. ex. 1: $C = \sum_{i=1}^{3} \lambda_i u_i u_i^T$

$$\lambda_1 u_1 u_1^T = \cancel{\lambda} \cdot \frac{1}{2} \begin{pmatrix} \cos^2\alpha & \cos\alpha\sin\alpha & \cos\alpha \\ \cos\alpha\sin\alpha & \sin^2\alpha & \sin\alpha \\ \cos\alpha & \sin\alpha & 1 \end{pmatrix}$$

$$\lambda_2 u_2 u_2^T = 1 \cdot \begin{pmatrix} \sin^2\alpha & -\sin\alpha\cos\alpha & 0 \\ -\sin\alpha\cos\alpha & \cos^2\alpha & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\lambda_3 u_3 u_3^T = \lambda_3 \frac{1}{2} \begin{pmatrix} \cos^2\alpha & \cos\alpha\sin\alpha & -\cos\alpha \\ \cos\alpha\sin\alpha & \sin^2\alpha & -\sin\alpha \\ -\cos\alpha & -\sin\alpha & 1 \end{pmatrix} \longrightarrow 2^{nd} \text{ part in } C$$

$$\lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T = \begin{pmatrix} 1 & 0 & \cos\alpha \\ 0 & 1 & \sin\alpha \\ \cos\alpha & \sin\alpha & 1 \end{pmatrix} \longrightarrow 1^{st} \text{ part in } C$$

(4) The principal component weights correspond to the eigenvectors of the covariance matrix C. Since we want to explain as much variance as possible, we would use the two PC with the biggest eigenvalues, that means $s_1 = \underset{\substack{\uparrow \\ \text{weights}}}{u_1^T} \underset{\substack{\nwarrow \\ \text{random vector}}}{x}$ and $s_2 = u_2^T x$

(5) The proportion of variance explained is defined as $\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$, where k is the number of selected components, and n the total dimension, hence we get: $\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{3}{3.1} \approx 0.97$, meaning that 97% of the variance is explained by the first two PCs.

(6) Here we wanted you to make plots similar to the illustration in the lecture in Chapter 3.4.
The projection of a point x is defined as $(u_1^T x, u_2^T x)$.

For $y_1 = \begin{pmatrix} x_1 \\ 0 \\ 0 \end{pmatrix}$ the projection is $(u_1^T y_1, u_2^T y_1) = \left(\frac{x_1}{\sqrt{2}} \cos\alpha, -x_1 \sin\alpha \right) = x_1 \cdot \underbrace{\left(\frac{1}{\sqrt{2}} \cos\alpha, -\sin\right.}_{pe_1}$

For $y_2 = \begin{pmatrix} 0 \\ x_2 \\ 0 \end{pmatrix}$ — " — $(u_1^T y_2, u_2^T y_2) = \left(\frac{x_2}{\sqrt{2}} \sin\alpha, x_2 \cos\alpha \right) = x_2 \cdot \underbrace{\left(\frac{1}{\sqrt{2}} \sin\alpha, \cos\alpha\right.}_{pe_2}$

For $y_3 = \begin{pmatrix} 0 \\ 0 \\ x_3 \end{pmatrix}$ — " — $(u_1^T y_3, u_2^T y_3) = \left(\frac{x_3}{\sqrt{2}}, 0 \right) = x_3 \cdot \underbrace{\left(\frac{1}{\sqrt{2}}, 0\right)}_{pe_3}$

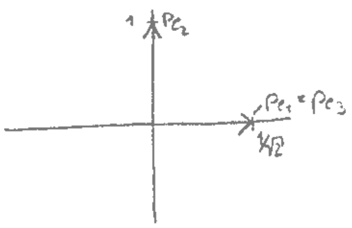
From this formulas we see, that projecting y_i is the same as projecting the i-th unit vector scaled by the value x_i. In the plots we thus only show the projection of the unit vectors e_i; the projection of any other vector of the form y_i lies along the same axes.

$\alpha = 0$

$pe_1 = \left(\frac{1}{\sqrt{2}}, 0\right)$

$pe_2 = (0, 1)$

$pe_3 = \left(\frac{1}{\sqrt{2}}, 0\right)$

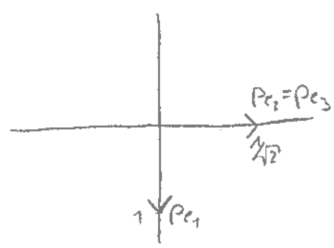


$\alpha = \frac{\pi}{2}$

$pe_1 = (0, 1)$

$pe_2 = \left(\frac{1}{\sqrt{2}}, 0\right)$

$pe_3 = \left(\frac{1}{\sqrt{2}}, 0\right)$

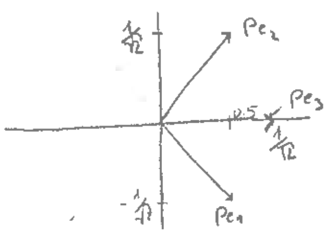


$\alpha = \frac{\pi}{4}$

$pe_1 = \left(0.5, -\frac{1}{\sqrt{2}}\right)$

$pe_2 = \left(0.5, \frac{1}{\sqrt{2}}\right)$

$pe_3 = \left(\frac{1}{\sqrt{2}}, 0\right)$

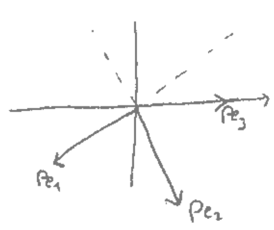


$\alpha = \frac{5\pi}{6}$

$p_1 \approx (-0.61, -0.5)$

$p_2 \approx (0.35, -0.86)$

$p_3 = \left(\frac{1}{\sqrt{2}}, 0\right)$



(7) The correlation between the 1st and 3rd variable is given by $\cos\alpha$. Between the 2nd and 3rd variable by $\sin\alpha$: For the α's above we obtain:

$\alpha = 0$: $\rho_{13} = 1$, $\rho_{23} = 0$ $\qquad \alpha = \frac{\pi}{2}$: $\rho_{13} = 0$, $\rho_{23} = 1$

$\alpha = \frac{\pi}{4}$: $\rho_{13} = \frac{1}{\sqrt{2}}$, $\rho_{23} = \frac{1}{\sqrt{2}}$ $\qquad \alpha = \frac{5\pi}{6}$: $\rho_{13} = -0.86$, $\rho_{23} = 0.5$

If axes pe_3 is closer to axis pe_1 then to pe_2, the 3rd variable is more correlated to the first var. then to the 2nd. If the arrows point in the same direction they are positively correlated. otherwise negatively.

Ex 4

(1) $y_k = x_k^T \beta + \varepsilon_k \qquad k=1\dots n \qquad$ with $\varepsilon_k \sim \mathcal{N}(0,\sigma^2)$ iid.

In matrix notation $\quad \underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

$$X = (x_1 \dots x_n), \qquad x_k \in \mathbb{R}^p$$
$$X^T = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$$

$$\underline{y} = X^T \beta + \underline{\varepsilon}$$

Observed are only $\underline{y}$ and X, minimizing $J(\beta)$

$$J(\beta) = \frac{1}{N} (\underline{y} - X^T\beta)^T (\underline{y} - X^T\beta)$$

gives an estimate $\hat\beta = \underset{\beta}{\arg\min}\, J(\beta)$ of the true value of $\beta \in \mathbb{R}^p$.

$$J(\beta) = \frac{1}{n}\left[\underline{y}^T\underline{y} - \underline{y}^T X^T \beta - \beta^T X \underline{y} + \beta^T X X^T \beta \right]$$

$$\nabla_\beta J = \frac{1}{n}\left[-2X\underline{y} + 2XX^T\beta \right] \qquad \text{(cf. math. assignment 2)}$$

$$\nabla_\beta J \doteq 0$$

$$\Rightarrow \boxed{\begin{aligned} \hat\beta &= (XX^T)^{-1} X\underline{y} \\ &= \left(\tfrac{1}{n}XX^T\right)^{-1} \tfrac{1}{n} X \underline{y} \end{aligned}}$$

$\frac{1}{n}XX^T = \hat C_x$, the sample covariance matrix

$\hat\beta = \hat C_x^{-1}\left(\frac{1}{n}X\underline{y}\right)$.

(2) For n large $\quad \beta \to \underset{p\times p}{C_x^{-1}} \underset{p\times 1}{C_{xy}}$ ← cross correlation between x & y. (cf. ex. 1)

(3) $\hat\beta = (XX^T)^{-1} X \underline{y}, \quad \underline{y} = X^T\beta + \underline{\varepsilon}$

$$\Rightarrow \hat\beta = (XX^T)^{-1}(XX^T)\beta + (XX^T)^{-1}X\underline{\varepsilon} = \beta + (XX^T)^{-1}X\underline{\varepsilon}$$

$$E(\hat\beta \mid X) = \beta + (XX^T)^{-1}X\underbrace{E(\varepsilon \mid X)}_{0} = \beta.$$

$$V(\hat\beta \mid X) = \underset{p\times p}{(XX^T)^{-1}}\underset{p\times n}{X}\underbrace{E(\varepsilon\varepsilon^T \mid X)}_{\sigma^2 I_n} X^T (XX^T)^{-1} = \sigma^2 (XX^T)^{-1}(XX^T)(XX^T)^{-1}$$

$$= \sigma^2 (XX^T)^{-1} = \frac{\sigma^2}{n}\left(\frac{1}{n}XX^T\right)^{-1} = \frac{\sigma^2}{n}\hat C_x^{-1}.$$

For $n \to \infty$ $\left(n \cdot V(\hat\beta \mid X)\right) = \sigma^2 C_x^{-1}$

(4) $MSE = E\left(\|\beta - \hat\beta\|^2 \mid X \right) = E\left(tr\left[(\beta-\hat\beta)(\beta-\hat\beta)^T \right] \mid X \right)$

$= E\left(tr\left[(\beta - m + m - \hat\beta)(\beta - m + m - \hat\beta)^T \right] \mid X \right)$ where $m = E(\hat\beta \mid X)$

$$\left((\beta-m) + m-\hat\beta\right)\left(\beta - m + m - \hat\beta\right)^T = (\beta-m)(\beta-m)^T + (\beta-m)(m-\hat\beta)^T + (m-\hat\beta)(\beta-m)^T + (m-\hat\beta)(m-\hat\beta)^T.$$

The trace is a linear operator, we can thus take the expectation inside to get

$$MSE = tr\Big[E\left((\beta-m)(\beta-m)^T \mid X\right) + E\left((\beta-m)(m-\hat\beta)^T + (m-\hat\beta)(\beta-m)^T \mid X\right) + E\left((m-\hat\beta)(m-\hat\beta)^T \mid X\right) \Big]$$

- $E\left((\beta-m)(m-\hat\beta)^T\right) = \left(E(\beta \mid X) - m\right)(m-\hat\beta)^T = (m-m)(m-\beta)^T = 0$

  same for $E\left((m-\hat\beta)(\beta-m)^T \mid X\right)$.

- $E\left((\beta-m)(\beta-m)^T \mid X\right) = (\beta-m)(\beta-m)^T$ (everything is deterministic here)
- $E\left((m-\hat\beta)(m-\hat\beta)^T \mid X\right) = V(\hat\beta \mid X)$ by definition of the variance and the fact that $m = E(\hat\beta \mid X)$.

$$\Rightarrow MSE = \underbrace{tr\left[(\beta-m)(\beta-m)^T\right]}_{\|\beta-m\|^2} + tr V(\hat\beta \mid X)$$

$m = E(\hat\beta|X)$

$$\Rightarrow MSE = \underbrace{\|\beta - E(\hat\beta \mid X)\|^2}_{bias^2} + \underbrace{tr\, V(\hat\beta \mid X)}_{\text{variance-term}}$$

Since $E(\hat\beta \mid X) = \beta$, there is no bias.

$$\Rightarrow MSE = tr\, V(\hat\beta \mid X) \underset{(3)}{=} \frac{\sigma^2}{n} tr\, \hat C_x^{-1} = \frac{\sigma^2}{n} \sum_{i=1}^{p} \frac{1}{d_i}$$

$tr(A^{-1}) = \sum \frac{1}{\text{eigenvalues of } A}$ (cf math assignment 1)

The small eigenvalues of $\hat C_x$ cause the MSE to be large.
The eigenvalues of $\hat C_x$ are small if some random variables x_i (elements of the vector $\vec{x}$) are highly correlated $\equiv$ some rows of X are (linearly) dependent (cf Ex. 2).

(5) Let $\beta = U_m \gamma$, $U_m = (u_1 \dots u_m)$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$.

$$J(U_m\gamma) = \frac{1}{n}\sum_{k=1}^{n} (y_k - x_k^T U_m \gamma)^2 = \frac{1}{n}\sum_{k=1}^{n} (y_k - (U_m^T x_k)^T \gamma)^2$$

$$= \frac{1}{n}\sum_{k=1}^{n} (y_k - z_k^T \gamma)^2 , \quad z_k = U_m^T x_k$$ (vector of the k-th observation of the principal components)

$$\Rightarrow J_{pc}(\gamma) := \frac{1}{n}\sum_{k=1}^{n} (y_k - \underbrace{z_k^T}_{\text{vector with principal components}}\gamma)^2$$

$$J(\beta) = \frac{1}{n}\sum_{k=1}^{n} (y_k - \underbrace{x_k^T}_{\text{original inputs}}\beta)^2$$

J_{pc} has the same form as J but the principal components are used instead of the original inputs.

(6) As in (1) but with $Z = (z_1 \dots z_n)$, $z_k \in \mathbb{R}^m$ instead of X:

$$\hat{\gamma} = \left(\frac{1}{n} Z Z^T\right)^{-1} \frac{1}{n} Z y$$

$$= \left(\frac{1}{n} U_m^T X X^T U_m\right)^{-1} \frac{1}{n} Z y \underset{\uparrow}{=} (U_m^T U D U^T U_m)^{-1} \frac{1}{n} Z y$$

$$\underbrace{\frac{1}{n} X X^T}_{= \hat{C}_x} = U D U^T$$

- $U_m^T U = \begin{pmatrix} u_1^T \\ \vdots \\ u_m^T \end{pmatrix} (u_1 \dots u_m \; u_{m+1} \dots u_p)$

$$= m\Big\{ \underbrace{\left(\begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix} \,\middle|\, 0 \right)}_{m \,\to\, p}$$

- $m\Big\{\left(\underbrace{\begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix}}_{m} \,\middle|\, 0 \right) D \begin{pmatrix} \begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix} \\ 0 \end{pmatrix}_m \Big\} m \Big| p = D_m$, where $D_m = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_m \end{pmatrix}$ has the first m diagonal elements of D on its diagonal.

$$\Rightarrow \underline{\underline{\hat{\gamma} = D_m^{-1} \frac{1}{n} Z y = D_m^{-1} U_m^T \frac{1}{n} X y.}}$$

$$\underline{\underline{\hat{\beta}_{pc} = U_m \hat{\gamma} = U_m D_m^{-1} U_m^T \frac{1}{n} X y.}}$$

Note: For $m = p$ $U_m D_m^{-1} U_m^T$ becomes the inverse of the sample covariance matrix $\hat{C}_x$.

(7) $$\hat{\beta}_{pc} = (U_m D_m^{-1} U_m^T)\, \tfrac{1}{n} X \,(X^T\beta + \underline{\varepsilon})$$

$$= (U_m D_m^{-1} U_m^T)\, \tfrac{1}{n} X X^T \beta + (U_m D_m^{-1} U_m^T)\, \tfrac{1}{n} X \underline{\varepsilon}$$

$$E(\hat{\beta}_{pc} | X) = (U_m D_m^{-1} U_m^T)\, \tfrac{1}{n} X X^T \beta + (U_m D_m^{-1} U_m^T)\, \tfrac{1}{n} X \underbrace{E(\underline{\varepsilon}|X)}_{0}$$

$\frac{1}{n} X X^T = \hat{C}_x = U D U^T$, hence

$$E(\hat{\beta}_{pc} | X) = U_m D_m^{-1} \underbrace{U_m^T U}_{\left(\begin{smallmatrix}1 & & \\ & \ddots & \\ & & 1\end{smallmatrix} \middle| 0\right)} D U^T \beta = U_m D_m^{-1} \left(\begin{smallmatrix}d_1 & & \\ & \ddots & \\ & & d_m\end{smallmatrix} \middle| 0\right) U^T \beta$$

$$= U_m \left(\begin{smallmatrix}1 & & \\ & \ddots & \\ & & 1\end{smallmatrix} \middle| 0\right) U^T \beta = U_m \Bigg(\underbrace{U \begin{pmatrix}\begin{smallmatrix}1 & & \\ & \ddots & \\ & & 1\end{smallmatrix} \\ 0\end{pmatrix}}_{U_m}\Bigg)^T = U_m U_m^T \beta$$

$$\Rightarrow \quad E(\hat{\beta}_{pc} | X) = U_m U_m^T \beta .$$

$$V(\hat{\beta}_{pc} | X) = E\left[(U_m D_m^{-1} U_m^T) \left(\tfrac{1}{n}\right)^2 X \underline{\varepsilon}\,\underline{\varepsilon}^T X^T U_m D_m^{-1} U_m^T\right]$$

$$= \tfrac{1}{n}(U_m D_m^{-1} U_m^T)\, \tfrac{1}{n} X\, E(\underline{\varepsilon}\,\underline{\varepsilon}^T) X^T U_m D_m^{-1} U_m^T$$

$E(\underline{\varepsilon}\,\underline{\varepsilon}^T) = \sigma^2 I_n$

$$= \frac{\sigma^2}{n} (U_m D_m^{-1} U_m^T) \underbrace{\left(\tfrac{1}{n} X X^T\right)}_{\hat{C}_x = U D U^T} U_m D_m^{-1} U_m^T$$

$$= \frac{\sigma^2}{n} (U_m D_m^{-1} \underbrace{U_m^T)\, U}_{\left(\begin{smallmatrix}1 & & \\ & \ddots & \\ & & 1\end{smallmatrix} \middle| 0\right)} D \underbrace{U^T (U_m}_{\begin{pmatrix}\begin{smallmatrix}1 & & \\ & \ddots & \\ & & 1\end{smallmatrix} \\ 0\end{pmatrix}} D_m^{-1} U_m^T)$$

$$= \frac{\sigma^2}{n} U_m D_m^{-1} D_m D_m^{-1} U_m^T = \frac{\sigma^2}{n} U_m D_m^{-1} U_m^T$$

$$\Rightarrow \quad V(\hat{\beta}_{pc} | X) = \frac{\sigma^2}{n} U_m D_m^{-1} U_m^T$$

For $m = p$ $\; U_m D_m^{-1} U_m^T = \hat{C}_x^{-1}$, that is for $m = p$,

$$V(\hat{\beta}_{pc} | X) = V(\hat{\beta} | X) .$$

(8) $$MSE_{pc} = \| \beta - U_m U_m^T \beta \|^2 + \frac{\sigma^2}{n} \underbrace{tr(U_m D_m^{-1} U_m^T)}_{tr(U_m^T U_m D_m^{-1}) = tr(D_m^{-1}) = \sum_{i=1}^{m} \frac{1}{d_i}}$$

$$= \| \beta (I_p - U_m U_m^T) \|^2 + \frac{\sigma^2}{n} \sum_{i=1}^{m} \frac{1}{d_i}$$

If $m = p$, $U_m = U$ and $UU^T = I_p$, so that

MSE_{pc} becomes $\frac{\sigma^2}{n} \sum_{i=1}^{p} \frac{1}{d_i}$, i.e. equal to the MSE in (4). (PCA regression boils then down to ordinary regression)

If $m < p$, the variance is reduced by $\frac{\sigma^2}{n} \sum_{i=m+1}^{p} \frac{1}{d_i}$, but we incur a bias since $U_m U_m^T \neq I_p$.

This is called the bias-variance trade-off: By choosing m, one can choose a certain reduction in variance, at the cost of more bias. The best m is the one which leads to the smallest

MSE $(= \underset{\uparrow \text{ as } m\downarrow}{\text{bias}^2} + \underset{\downarrow \text{ as } m\downarrow}{\text{variance}})$.

The formula for the MSE_{pc} show that the best m is essentially a function of d_i and u_i, i.e. the covariance matrix of X.