4-4-2025 (Shrinkage methods)

Ridge Regression

$$\min \Big\{ \underbrace{(y - X\beta)^T (y - X\beta)}_{\sum_{i=1}^{n} (y_i - \beta_0 - \dots - \beta_p x_{ip})^2} + \lambda \underbrace{\beta^T \beta}_{\sum_{j=1}^{p} \beta_j^2} \Big\} \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\nabla_\beta = 0$$

$$\Rightarrow 2(X^T X)\beta - 2X^T y + 2\lambda \beta = 0$$

$$\Rightarrow (X^T X + \lambda I)\beta = X^T y \Rightarrow \boxed{\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y}$$

(για $\lambda = 0$: $\hat{\beta} = (X^T X)^{-1} \cdot X^T y = \hat{\beta}_{LSE}$

Παράδειγμα

x_1 = εισόδημα (1000 ευρώ)
x_2 = απόσταση (m

$\Rightarrow \hat{b}_1, \hat{b}_2$

$\tilde{x}_1$ (σε ευρώ) : $\tilde{x}_1 = 1000 \cdot x_1$
$\tilde{x}_2$ (σε km) : $\tilde{x}_2 = 10^{-3} \cdot x_2$

$\Rightarrow \tilde{b}_1 = 10^{-3} \cdot \hat{b}_1$, $\tilde{b}_2 = 10^{3} \cdot \hat{b}_2$

$$\hat{y}_j = \tilde{y}_j$$

Οι γραμμικοί μετ/σμοί των μεταβλητών επηρεάζουν τη σχετική μεταβολή των συντελεστών (δηλ. τη σχετική σπουδαιότητα των μεταβλητών στο μοντέλο) σε αντίθεση με την ελαχιστοποίηση τετραγώνων.

Κανονικοποίηση των X_i πριν την εφαρμογή της μεθόδου

$$\tilde{X}_i = \frac{X_i - \overline{X}_i}{s(X_i)}$$

① Έστω $X_2 = X_1^2$, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

$$\tilde{X}_1 = \frac{X_1 - \overline{X}_1}{s(X_1)} , \quad \tilde{X}_2 = \frac{X_2 - \overline{X}_2}{s(X_2)}$$

$$\overline{X}_2 = \overline{X_1^2} \neq \left(\overline{X}_1\right)^2$$

② Έστω training set $\overline{X}_1 = 5$, $s(X_1) = 2$

test set $\overline{X}_1 = 5{,}5$, $s(X_1) = 1.5$

Training : Training set $\tilde{X}_1 = \frac{X_1 - 5}{2}$

Test set $\tilde{X}_1 = \frac{X_1 - 5.5}{1.5}$

Training : $\hat{\beta}_0, \hat{\beta}_1$: $\forall\, i \in$ Training

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \tilde{x}_{i1}$$

$$= \hat{\beta}_0 + \hat{\beta}_1 \cdot \frac{x_{i1} - 5}{2}$$

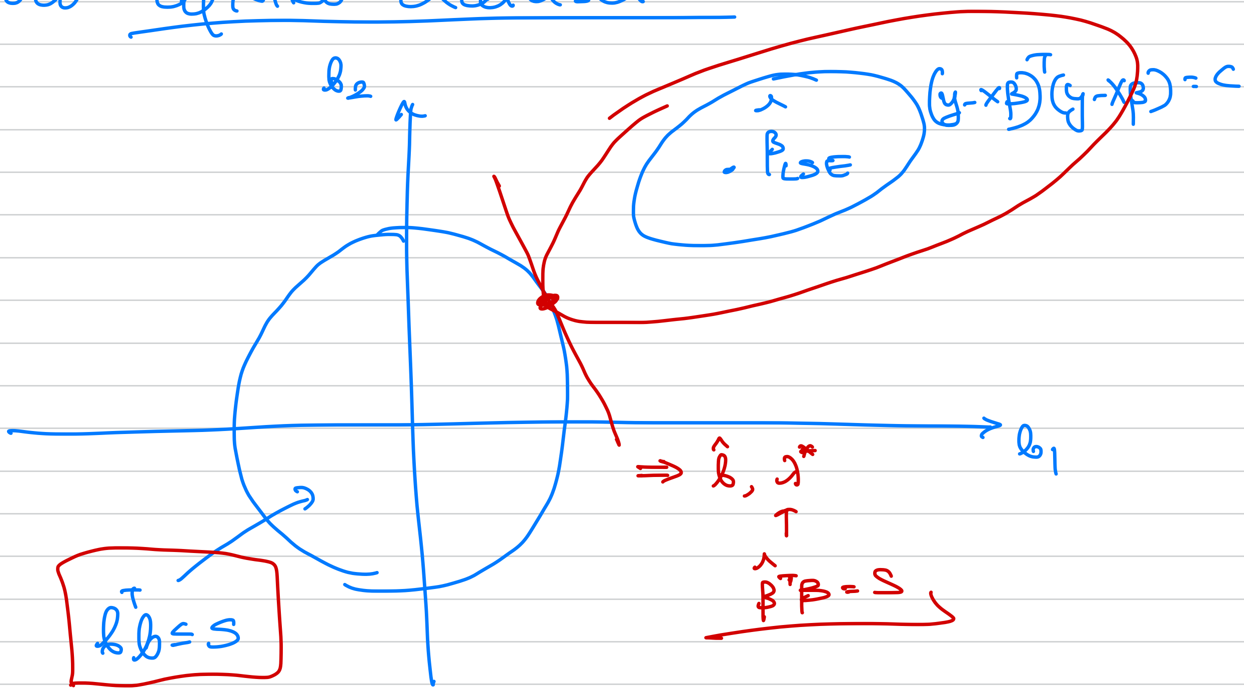
Αν στο testing set χρησιμοποιήσουμε

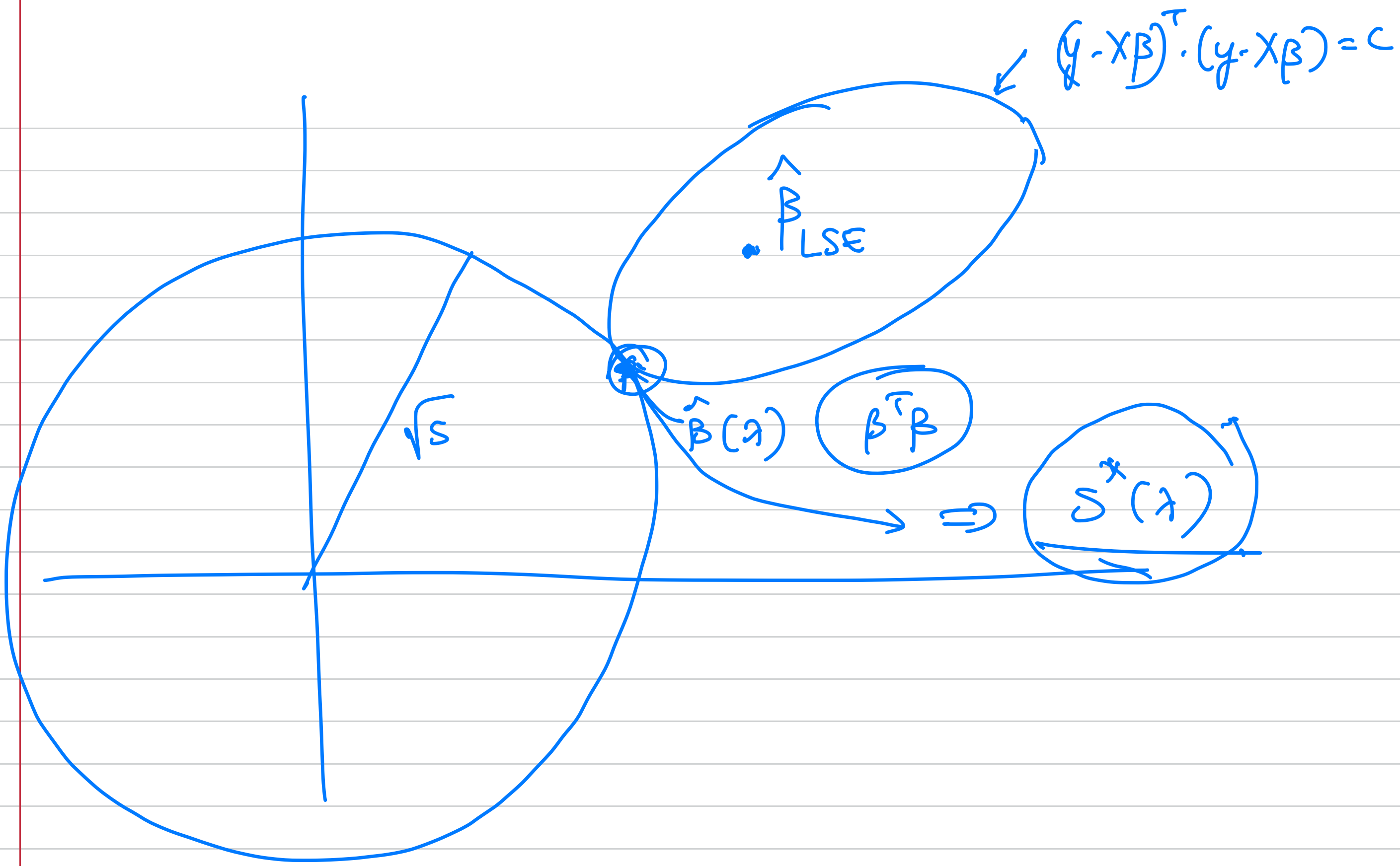
$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 \underbrace{\tilde{x}_{j1}}_{test} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \frac{x_{j1} - 5.5}{1.5} \neq$$

!!

Είναι λάθος να χρησιμοποιήσουμε χωριστή τυποποίηση στο test set !!

Εναλλακτικά: Αν τα training + test sets προέρχονται από ένα αρχικό dataset η τυποποίηση μπορεί να γίνει εξ αρχής στο αρχικό dataset

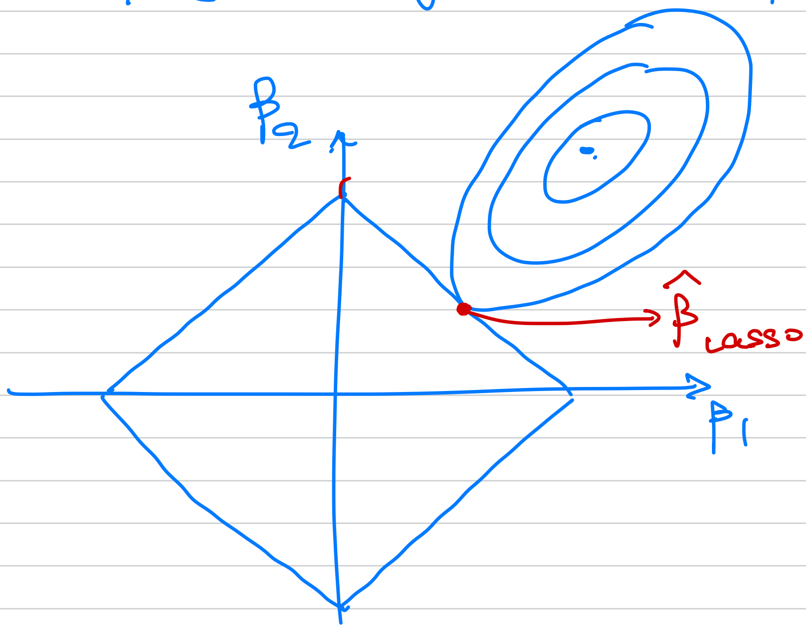


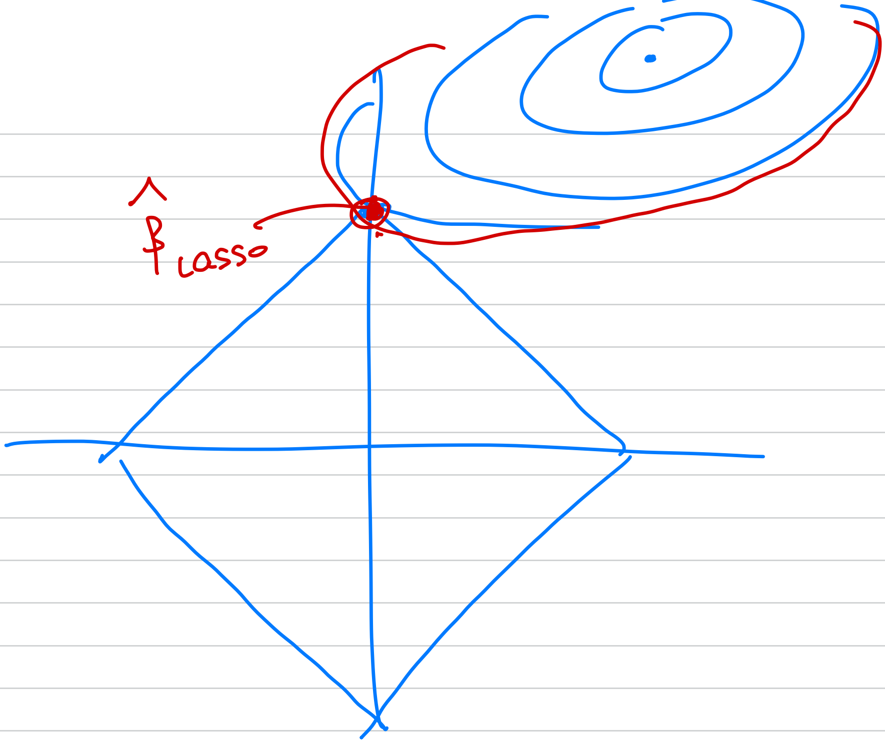


Lasso Regression

(quadratic programming)

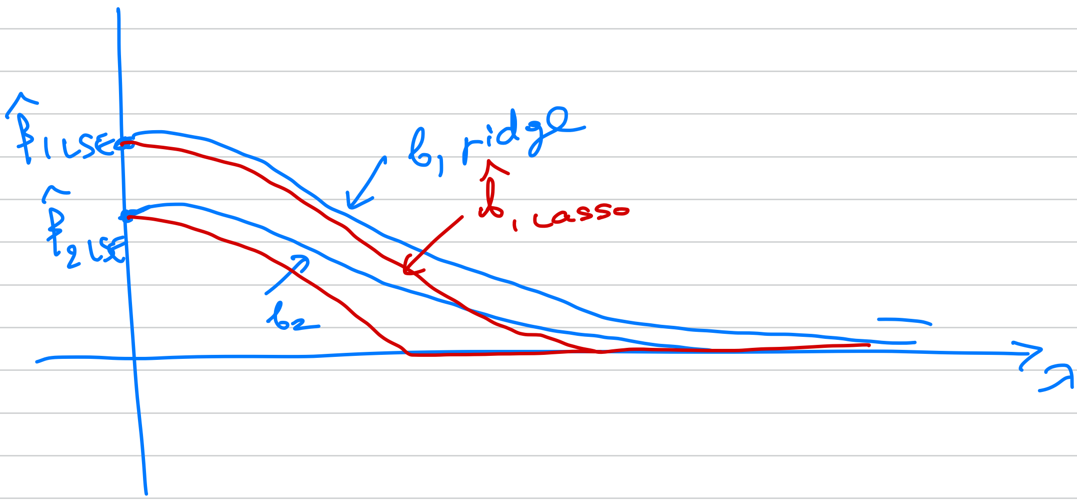
$$\min\left\{(y-X\beta)^T(y-X\beta)\ :\ \|\beta\|_1 \le s\right\}$$

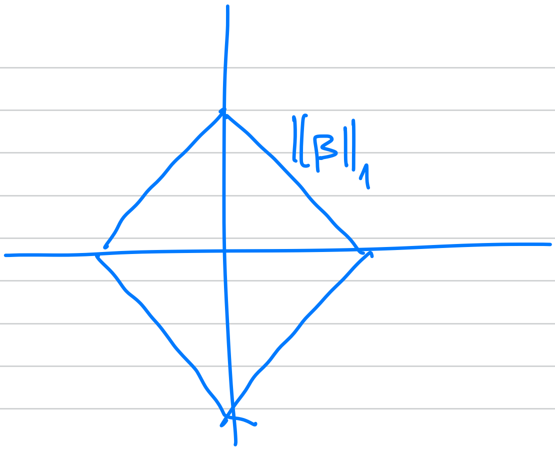


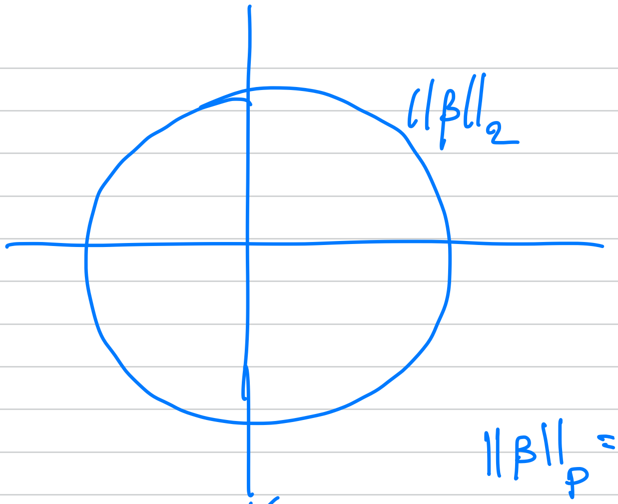


$$\min_{\beta} \left\{ (Y - X\beta)^T (Y - X\beta) + \lambda \|\beta\|_1 \right\}$$

Ridge

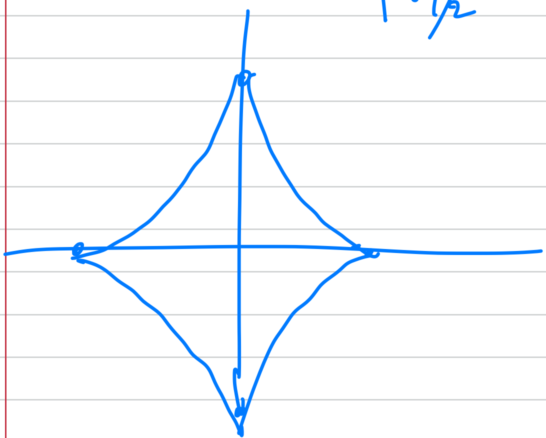


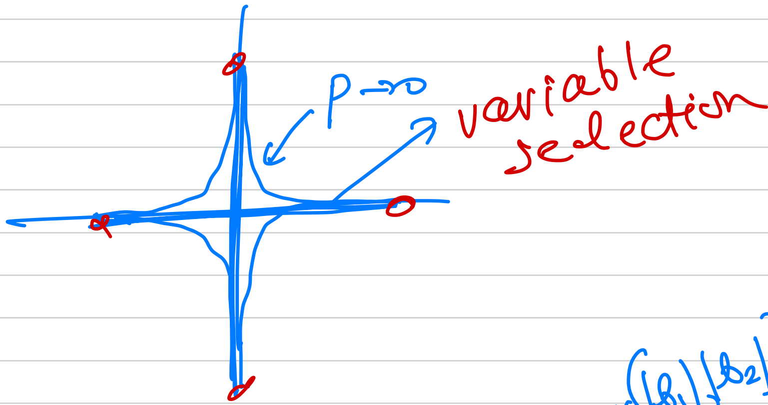




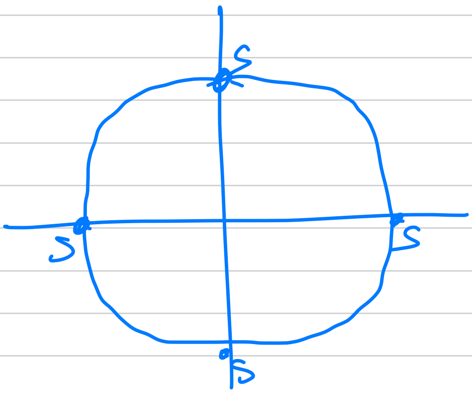
$$\|\beta\|_p = \left(\beta_1^p + \beta_2^p\right)^{1/p}$$

$$\|\beta\|_{1/2} \quad \left(\beta_1^{1/2} + \beta_2^{1/2}\right)^{1/2}$$





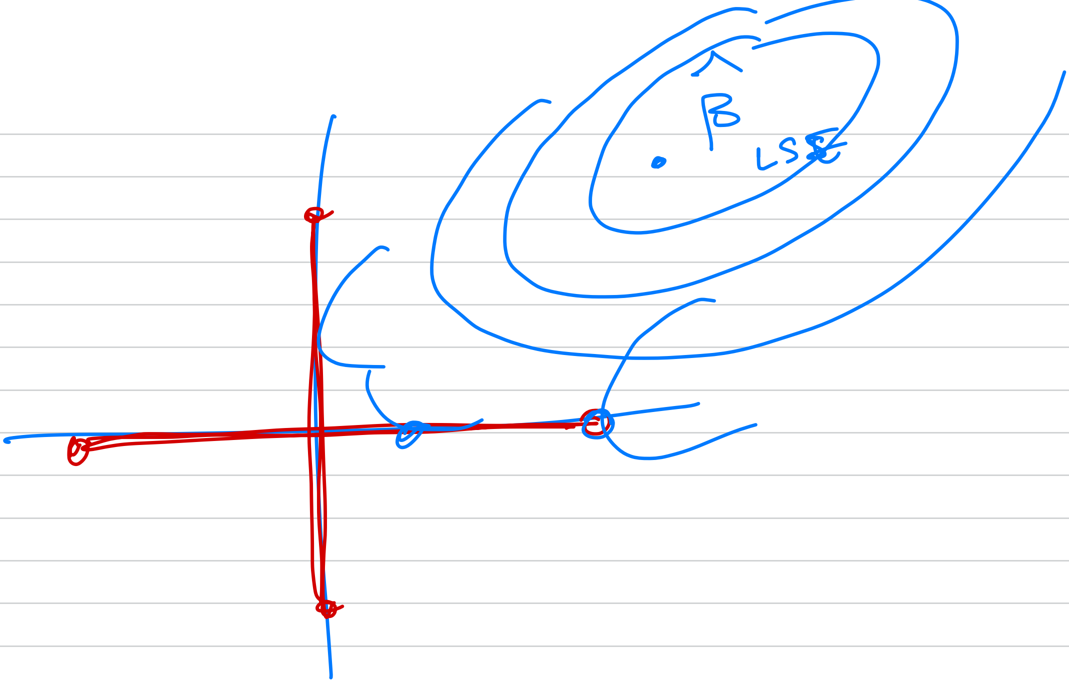
$\|\beta\|_3 \leq s$



$\|\beta\|_\infty \leq s \Rightarrow \max(|\beta_1|, |\beta_2|) \leq s$

$\Rightarrow |\beta_1| \leq s$

$|\beta_2| \leq s$



Elastic net ($\Rightarrow$ glmnet library R)

$$\min \left\{ (Y - X\beta)^T (Y - X\beta) + \alpha \lambda_1 \|b\|_1 + (1-\alpha) \lambda_2 \|b\|_2^2 \right\}$$

Ridge-Lasso εφαρμόσιμες κ' για $p>n$

Μέθοδοι μείωσης διάστασης (του X)

$$Y = b_0 + b_1 X_1 + \dots + b_p X_p$$

(π.χ. $X_1, \dots, X_p$ βαθμοί σε 36 μαθήματα αποφοίτου

y = μισθός 5 έτη μετά την αποφοίτηση)

$X_1, \dots, X_p$ <u>έχουν συσχέτιση</u>

Αν θέλαμε να πάρουμε <u>ένα σκορ</u> (συνάρτηση των $X_1, \dots, X_p$)

που να σχετίζεται (δούλευαν) με την y

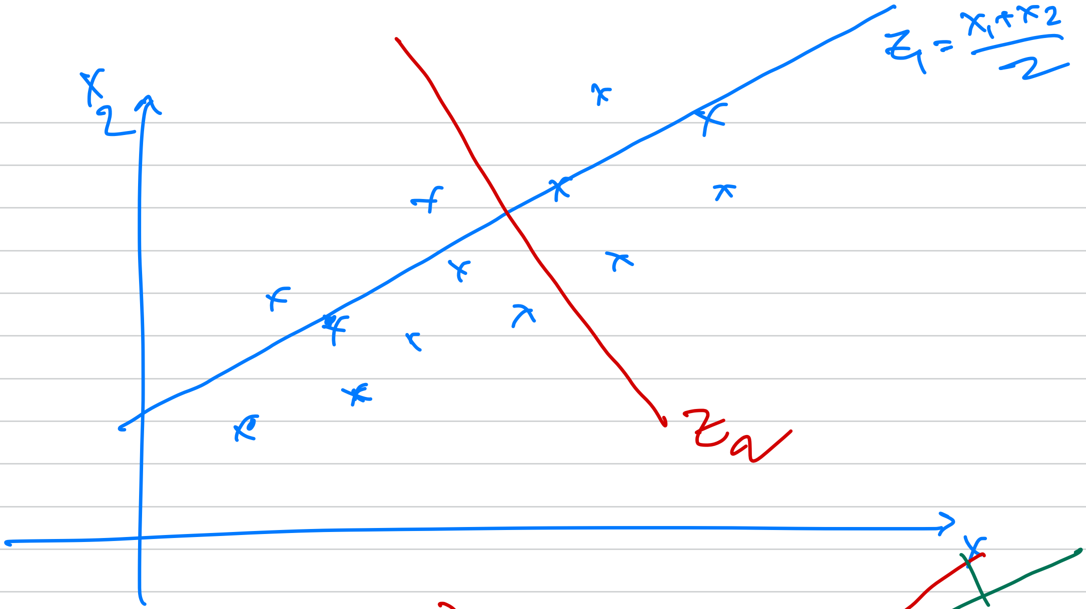
(π.χ. μέσος όρος πτυχίου) $= Z$

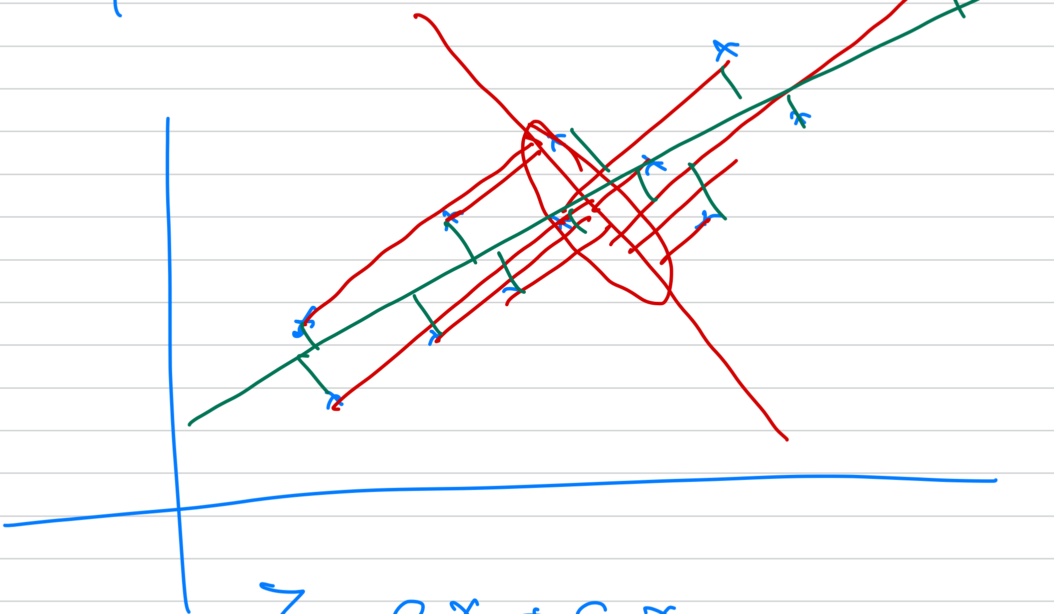
$$Z = \varphi_1 X_1 + \dots + \varphi_p X_p$$

$$y = \theta_0 + \theta_1 \cdot Z$$

Αν προσθέταμε κ' μια δεύτερη Z_2

(θα έπρεπε να είναι <u>ασυσχέτιστη</u> με την Z_1)



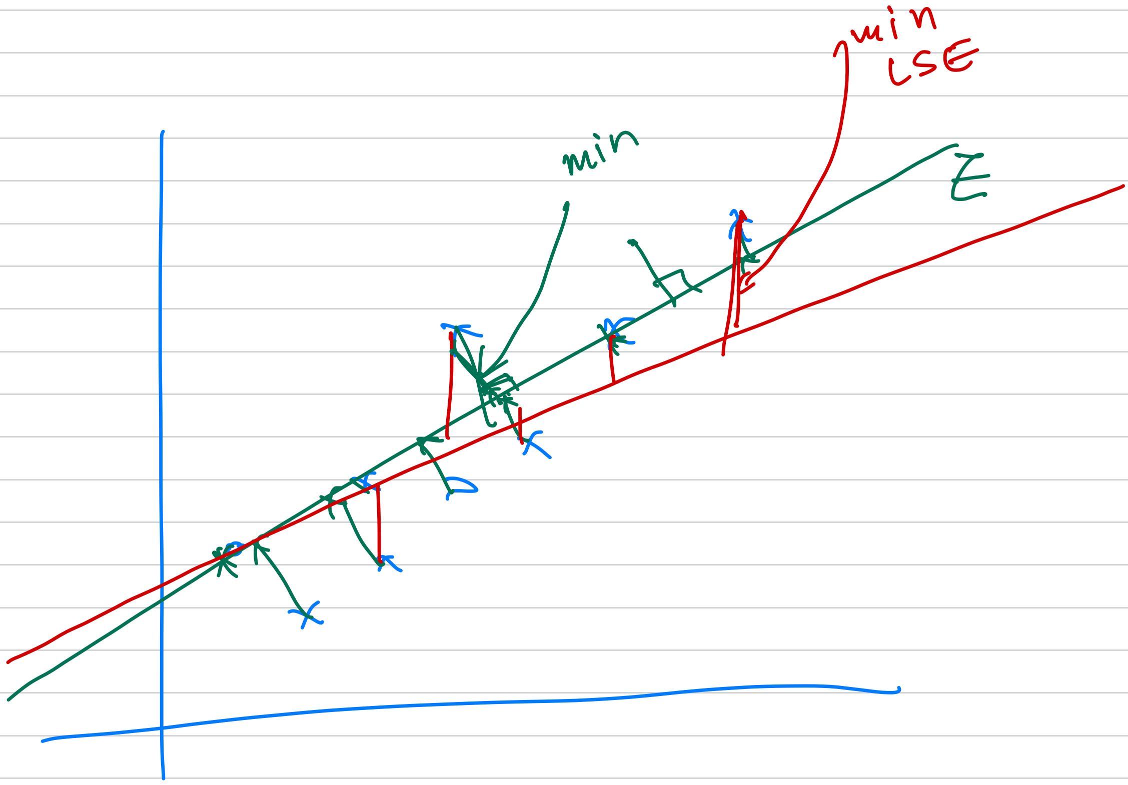


$$Z = \varphi_1 x_1 + \varphi_2 x_2$$

Ποια είναι τα φ_1, φ_2 που δείχνουν με τον καλύτερο τρόπο τα (x_1, x_2) $\forall$ παρατήρηση;

$$(x_1, x_2) \rightarrow Z = \varphi_1 x_1 + \varphi_2 x_2$$

x_1	x_2	Z
x_{11}	x_{12}	Z_1
x_{21}	x_{22}	$\vdots$
$\vdots$		$\vdots$
x_{n1}	x_{n2}	Z_n



Z : οι προβολές των παρατηρήσεων να έχουν τη μέγιστη δυνατή διασπορά

$\Updownarrow$

ελαχιστοποίηση των αποστάσεων (μέσω προβολής) από την ευθεία

$\Downarrow$

1^{η} κύρια συνιστώσα των $(X_1, X_2, \dots, X_p)$

$$Z_1, Z_2, \dots, Z_M \qquad (M \leq p)$$

$$Z_m = \sum_{j=1}^{p} \varphi_{jm} X_j \qquad m = 1, \dots, M$$

φ_{jm} : loadings

① $y = b_0 + b_1 X_1 + \dots + b_p X_p$ ⇐ αρχικό

$\quad = b_0 + \sum_{j=1}^{p} b_j X_j$

② $y = \theta_0 + \theta_1 Z_1 + \dots + \theta_M Z_M$ ⇐ μειωμένη διάσταση

$$= \theta_0 + \sum_{m=1}^{M} \theta_m Z_n$$

$$= \theta_0 + \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \varphi_{jm} X_j =$$

$$= \theta_0 + \sum_{j=1}^{p} \underbrace{\sum_{m=1}^{M} \theta_m \varphi_{jm}}_{b_j} X_j$$

$$b_j = \sum_{m=1}^{M} \theta_m \varphi_{jm} \qquad \forall\ j = 1, \dots, p.$$

$(2) \Leftrightarrow \quad y = b_0 + b_1 x_1 + \cdots + b_p x_p$

ν.δ. $\quad b_j = \sum_{m=1}^{M} \theta_m \varphi_{jm} \quad \forall j = 1, \ldots, p$

↑ δοσμένα