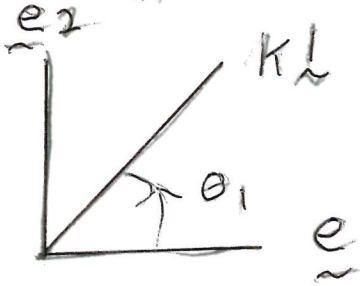
Giraud p21

17) Let $\underset{\sim}{e}_i = (0, \ldots, 0, \underset{\uparrow}{1}, 0, \ldots, 0)^T$, $i = 1, \ldots, p$ HD 16
ith position

be a basis for $\mathbb{R}^p$. Let $K\underset{\sim}{1}$ (in the span of $\underset{\sim}{1}$) be the diagonal of a hypercube or hypersphere.



a rotation does not change the angles, but makes $\cos(\theta_i)$ harder to compute

Then the angle θ_i between $\underset{\sim}{e}_i$ and $K\underset{\sim}{1}$ satisfies

$$\cos(\theta_i) = \frac{\langle \underset{\sim}{e}_i, K\underset{\sim}{1} \rangle}{\|\underset{\sim}{e}_i\| \|K\underset{\sim}{1}\|} = \frac{\sqrt{K}\, \underset{\sim}{e}_i^T \underset{\sim}{1}}{\sqrt{1\ K\ p}}$$

$$= \frac{1}{\sqrt{p}} \to 0 \text{ as } p \to \infty.$$

So the diagonal of a hypercube $\approx$ orthogonal to all edges (axes) of the hypercube in p dimensional space. (for large p.) For a vector $\underset{\sim}{v}$, $\cos(\theta_i) =$

$$\frac{\langle \underset{\sim}{e}_i, \underset{\sim}{v} \rangle}{\|\underset{\sim}{e}_i\| \|\underset{\sim}{v}\|} = \frac{v_i}{\sqrt{1\ \sum v_i^2}} = \frac{v_i}{\sqrt{\sum_{i=1}^{p} v_i^2}}$$

which often $\to 0$ as $p \to \infty$.

19} Suppose $\underset{\sim}{Y} = \mathbb{X}\underset{\sim}{\beta} + \underset{\sim}{e}$ and $(\mathbb{X}^T\mathbb{X})^{-1}$ exists (eg columns of $\mathbb{X}$ are orthonormal so $(\mathbb{X}^T\mathbb{X})^{-1} = I_p$).

Let A be a $p \times n$ matrix.

Then $E(\|A\underset{\sim}{e}\|^2) = E(\underset{\sim}{e}^T A^T A \underset{\sim}{e}) =$ ← tr(scalar) = scalar

$E[tr(\underset{\sim}{e}^T A^T A \underset{\sim}{e})] = E[tr(\underset{\sim}{e}\underset{\sim}{e}^T A A^T)]$

$tr(AB) = tr(BA)$

$= tr\left(\underbrace{E(\underset{\sim}{e}\underset{\sim}{e}^T)}_{cov(\underset{\sim}{e}) = \sigma^2 I} A A^T\right) = \sigma^2 tr(AA^T)$. (*)

Treat $\mathbb{X}$ as a constant matrix (conditional on $\mathbb{X}\underset{\sim}{\beta}$).

Then $\underset{(OLS)}{\hat{\underset{\sim}{\beta}}} - \underset{\sim}{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\underset{\sim}{Y} - \underset{\sim}{\beta} =$

$(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\underset{\sim}{\beta} + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\underset{\sim}{e} - \underset{\sim}{\beta}$. Thus

$E[\|\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}\|^2] = E[\|\underbrace{(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T}_{A}\underset{\sim}{e}\|^2]$

$= \sigma^2 tr\left[\underbrace{(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}}_{I}(\mathbb{X}^T\mathbb{X})^{-1}\right] = \sigma^2 tr[(\mathbb{X}^T\mathbb{X})^{-1}]$

↑ by (*)

If $(X^TX)^{-1} = I_p$, then $tr(I_p) = p$

and $E[\|\hat{\beta} - \beta\|^2] = p\sigma^2$.

If $n(X^TX)^{-1} \xrightarrow{P} V$, then $tr(X^TX)^{-1} \approx \frac{tr(V)}{n}$

$\geq \frac{p \min(V_{ii})}{n}$ which can be

very large if $p >> n$.

Errors of $\frac{1}{\sqrt{n}}$ or $\frac{1}{n}$ can add up to a big error if $p \geq n$

19] Let $\langle a, b \rangle = a^T b$ and

the Euclidean norm $\|a\| = \|a\|_2 = \sqrt{a^T a}$

$= \sqrt{\langle a, a \rangle}$.

20] The PLS literature incorrectly claims that $\beta_{OPLS} = \beta_{OLS}$ under mild conditions for p fixed and under stronger conditions for HD MLR.

$Y|x = \alpha + \beta^T x + e \Rightarrow (\alpha, \beta) = (\alpha_{OLS}, \beta_{OLS})$

under reasonable conditions.

OPLS literature assumes

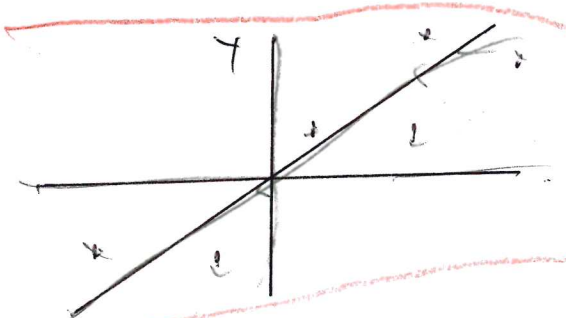
$$Y|\underset{\sim}{x} = \alpha_{OPLS} + \underset{\sim}{B}_{OPLS}^{T}\underset{\sim}{x} + e$$

which forces $\underset{\sim}{B}_{OLS} = \underset{\sim}{B}_{OPLS}$.

$$Y|\underset{\sim}{B}_{OPLS}^{T}\underset{\sim}{x} = \alpha_{OPLS} + \underset{\sim}{B}_{OPLS}^{T}\underset{\sim}{x} + e$$

often holds.

21] The conditioning error is made all the time in statistics. For OLS if $\underset{\sim}{x}$ is random, statisticians say that OLS results "hold conditional on $\underset{\sim}{x}$" $(Y|\underset{\sim}{x})$. This is often true for OLS, but not for OPLS. Statisticians also assume that if $Y = \underset{E}{\alpha} + \underset{E}{B}^{T}\underset{\sim}{x} + e$ is reasonable)



$$ESP = \hat{\underset{E}{\alpha}} + \hat{\underset{\sim}{B}}_{E}^{T}\underset{\sim}{x}$$

then $Y|\underset{\sim}{x} = \underset{E}{\alpha} + \underset{\sim}{B}_{E}^{T}\underset{\sim}{x} + e$.

Statisticians also act as if HD18
there is only one MLR model

$$Y = \alpha + \underset{\sim}{\beta}^T \underset{\sim}{x} + e.$$

22] Suppose $\begin{pmatrix} Y \\ \underset{\sim}{x} \end{pmatrix} \sim N_{p+1}\left[\begin{pmatrix} \mu_Y \\ \underset{\sim}{\mu_x} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \underset{\sim}{\Sigma_{Yx}} \\ \underset{\sim}{\Sigma_{xY}} & \underset{\sim}{\Sigma_x} \end{pmatrix}\right]$ ($\underset{\sim}{\mu}$, $\underset{\sim}{\Sigma}$)

with $\underset{\sim}{\Sigma_x}$ nonsingular. Let $\underset{k\times 1}{\underset{\sim}{w}} = \underset{k\times p}{A}\,\underset{\sim}{x} = \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix}$ ← k linear combos

where A is a full rank constant matrix.

Then $Y \mid \underset{\sim}{w} = Y \mid A\underset{\sim}{x}$ follows the pop OLS model of regressing Y on $A\underset{\sim}{x} = \underset{\sim}{w}$.

Proof] Let $B\begin{pmatrix} Y \\ \underset{\sim}{x} \end{pmatrix} = \begin{pmatrix} 1 & \underset{\sim}{0}^T \\ \underset{\sim}{0} & A \end{pmatrix}\begin{pmatrix} Y \\ \underset{\sim}{x} \end{pmatrix} =$

$$\begin{pmatrix} Y \\ A\underset{\sim}{x} \end{pmatrix} \sim N_{k+1}\left(B\underset{\sim}{\mu}, B\underset{\sim}{\Sigma}B^T\right)$$

$$\sim N_{k+1}\left[\begin{pmatrix} \mu_Y \\ A\underset{\sim}{\mu_x} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \underset{\sim}{\Sigma_{Yx}}A^T \\ A\underset{\sim}{\Sigma_{xY}} & A\underset{\sim}{\Sigma_x}A^T \end{pmatrix}\right]$$

$\underset{\sim}{\Sigma_{Yx}}A^T$ ← $\mathrm{cov}(Y, A\underset{\sim}{x})$

$A\underset{\sim}{\Sigma_x}A^T$ ← $\mathrm{cov}(A\underset{\sim}{x})$

$A\underset{\sim}{\Sigma_{xY}}$ ↑ $\mathrm{cov}(A\underset{\sim}{x}, Y)$

$A\underset{\sim}{\mu_x}$ ↑ $E(A\underset{\sim}{x})$

MVN determined by mean and cov matrix

Then $Y \mid A\underset{\sim}{X} = Y \mid \underset{\sim}{W} \sim N(\mu_{Y|\underset{\sim}{W}}, \Sigma_{Y|\underset{\sim}{W}})$.

$$\mu_{Y|\underset{\sim}{W}} = \mu_Y + \Sigma_{Y\underset{\sim}{X}} A^T (A\Sigma_{\underset{\sim}{X}} A^T)^{-1} (A\underset{\sim}{X} - A\underset{\sim}{\mu_X})$$

$$= \mu_Y + \Sigma_{Y\underset{\sim}{W}} \Sigma_{\underset{\sim}{W}}^{-1} (A\underset{\sim}{X} - A\underset{\sim}{\mu_X})$$

$$= \underbrace{\mu_Y - \Sigma_{Y\underset{\sim}{W}} \Sigma_{\underset{\sim}{W}}^{-1} E(\underset{\sim}{W})}_{E(Y) - \underset{\sim}{\beta}_{\underset{\sim}{W},OLS}^T E(\underset{\sim}{W})} + \underbrace{\Sigma_{Y\underset{\sim}{W}} \Sigma_{\underset{\sim}{W}}^{-1}}_{\underset{\sim}{\beta}_{\underset{\sim}{W},OLS}^T} \underset{\sim}{W}$$

$$\alpha_{\underset{\sim}{W},OLS}$$

$$\Sigma_{Y|\underset{\sim}{W}} = \sigma^2_{Y|\underset{\sim}{W}} = \Sigma_Y - \Sigma_{Y\underset{\sim}{X}} A^T (A\Sigma_{\underset{\sim}{X}} A^T)^{-1} A\Sigma_{\underset{\sim}{X}Y}$$

$$= \Sigma_Y - \Sigma_{Y\underset{\sim}{W}} \Sigma_{\underset{\sim}{W}}^{-1} \Sigma_{\underset{\sim}{W}Y}.$$

So $Y \mid \underset{\sim}{W} = \alpha_{\underset{\sim}{W},OLS} + \underset{\sim}{W}^T \underset{\sim}{\beta}_{\underset{\sim}{W},OLS} + e$

where $e \sim N(0, \sigma^2_{Y|\underset{\sim}{W}})$. □

23} In particular, let $W = \underset{\sim}{n}^T \underset{\sim}{X}$ with $A = \underset{\sim}{n}^T$.
(W: R.V., A: $1 \times p$)

Then under the conditions of 22},

$$Y | \underset{\sim}{n}^T \underset{\sim}{X} \sim N\left(\alpha_{\underset{\sim}{n}} + \underset{\sim}{\beta}_{\underset{\sim}{n}}^T \underset{\sim}{X}, \sigma_{\underset{\sim}{n}}^2\right) \text{ where}$$

$$\alpha_{\underset{\sim}{n}} = \mu_Y - \underset{\sim}{\beta}_{\underset{\sim}{n}}^T \underset{\sim}{\mu}_X, \quad \underset{\sim}{\beta}_{\underset{\sim}{n}} = \lambda \underset{\sim}{n}, \quad \sigma_{\underset{\sim}{n}}^2 = \underset{(=\sigma_Y^2)}{\sigma_Y^2} - \underset{\sim}{\beta}_{\underset{\sim}{n}}^T \underset{\sim}{\Sigma}_{XY}$$

$$\text{and } \lambda = \frac{\underset{\sim}{\Sigma}_{XY}^T \underset{\sim}{n}}{\underset{\sim}{n}^T \underset{\sim}{\Sigma}_X \underset{\sim}{n}}.$$

Proof} $\begin{pmatrix} 1 & \underset{\sim}{0}^T \\ \underset{\sim}{0} & \underset{\sim}{n}^T \end{pmatrix} \begin{pmatrix} Y \\ \underset{\sim}{X} \end{pmatrix} \sim \begin{pmatrix} Y \\ \underbrace{\underset{\sim}{n}^T \underset{\sim}{X}}_{W} \end{pmatrix} \sim N_2\left[\begin{pmatrix} \mu_Y \\ \underset{\sim}{n}^T \underset{\sim}{\mu}_X \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \underset{\sim}{\Sigma}_{XY}^T \underset{\sim}{n} \\ \underset{\sim}{n}^T \underset{\sim}{\Sigma}_{XY} & \underset{\sim}{n}^T \underset{\sim}{\Sigma}_X \underset{\sim}{n} \end{pmatrix}\right]$

$$E(Y|W) = \mu_Y + \frac{\underset{\sim}{\Sigma}_{XY}^T \underset{\sim}{n}}{\underset{\sim}{n}^T \underset{\sim}{\Sigma}_X \underset{\sim}{n}} \left(\underset{\sim}{n}^T \underset{\sim}{X} - \underset{\sim}{n}^T \underset{\sim}{\mu}_X\right) =$$

$$\underbrace{\mu_Y - \lambda \underset{\sim}{n}^T \underset{\sim}{\mu}_X}_{\alpha_{\underset{\sim}{n}}} + \underbrace{\lambda \underset{\sim}{n}^T}_{\underset{\sim}{\beta}_{\underset{\sim}{n}}^T} \underset{\sim}{X}$$

$$\sigma_{\underset{\sim}{n}}^2 = \sigma_Y^2 - \frac{\underset{\sim}{\Sigma}_{XY}^T \underset{\sim}{n}\, \underset{\sim}{n}^T \underset{\sim}{\Sigma}_{XY}}{\underset{\sim}{n}^T \underset{\sim}{\Sigma}_X \underset{\sim}{n}} = \sigma_Y^2 - \frac{\left(\underset{\sim}{\Sigma}_{XY}^T \underset{\sim}{n}\right)^2}{\underset{\sim}{n}^T \underset{\sim}{\Sigma}_X \underset{\sim}{n}} = \sigma_Y^2 - \underset{\sim}{n}^T \underset{\sim}{\Sigma}_{XY}.$$

□

Note that $\sigma_{\underset{\sim}{n}}^2 < \sigma_Y^2$ unless $\underset{\sim}{n}^T \underset{\sim}{\Sigma}_{\underset{\sim}{x}Y} = 0$. OPLS uses $\underset{\sim}{n} = \underset{\sim}{\Sigma}_{\underset{\sim}{x}Y}$.

24] If $\begin{pmatrix} Y \\ \underset{\sim}{x} \end{pmatrix} \sim N_{K+1}(\underset{\sim}{\mu}, \underset{\sim}{\Sigma})$, then there are a multitude of MLR models. Every linear combination $\underset{\sim}{n}^T \underset{\sim}{x}$ gives an MLR model.

($\begin{pmatrix} Y \\ \underset{\sim}{n}^T \underset{\sim}{x} \end{pmatrix} \sim N_2(\tilde{\underset{\sim}{\mu}}, \tilde{\underset{\sim}{\Sigma}})$, so highest density regions are ellipsoids, which are linear.)

cases: $(\underset{\sim}{x}_i^T, Y_i)^T$ i.i.d $N_{p+1}(\underset{\sim}{\mu}, \underset{\sim}{\Sigma})$ is a very common lit. assumption

25] under 22] $Y | \underset{\sim}{x}$ and $Y | \underset{\sim}{x}_I$ follow OLS MLR models, which was known. (Hence $Y | \underset{\sim}{w}$ follows an OLS MLR model.)

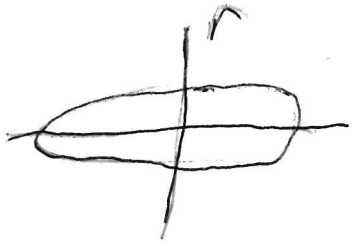
26] A response plot is a plot of ESP vs Y. For MLR add the identity line with unit slope and 0 intercept.

For MLR, $ESP = \hat{\alpha}_E + \underset{\sim}{x}^T \hat{\underset{\sim}{\beta}}_E = \hat{Y}$.

Want scatter about the identity line with no other pattern.

$E(Y | \underset{\sim}{x}_I^T \hat{\underset{\sim}{\beta}}_E) = 0 + 1 (\hat{\alpha}_E + \underset{\sim}{x}^T \hat{\underset{\sim}{\beta}}_E)$

27] For MLR, a residual plot is a plot of $ESP = \hat{Y}$ vs r $= Y - \hat{Y}$.



want scatter about the $r = 0$ line with no other pattern.

28] Arrange a data set $Y_1, \ldots, Y_n$ $Z_1, \ldots, Z_n$ in ascending order.

$$\underbrace{Y_{(1)}}_{\min} \leq Y_{(2)} \leq \cdots \leq Y_{(n-1)} \leq \underbrace{Y_{(n)}}_{\max}$$

are the order statistics.

29] The sample median (ave of middle order stat or stats)

$$MED(n) = \begin{cases} Y_{(\frac{n+1}{2})} & n \text{ odd} \\ \dfrac{Y_{(\frac{n}{2})} + Y_{(\frac{n}{2}+1)}}{2} & n \text{ even} \end{cases}$$

30] The sample median absolute deviation

$MAD(Z_1, \ldots, Z_n) = MAD(n) = MED(|Z_i - MED(n)|, i = 1, \ldots, n)$

is a measure of spread like the standard

deviation.

ex) 1, 2, 3, 4, 5, 6, 7, 8, 9

↑

MED(n) = 5

the median is the middle ordered value or the average of 2 middle values.

$|Y_i - MED(n)|$: +4, +3, +2, −1, 0, 1, 2, 3, 4

ordered 0, 1, 1, 2, 2, 3, 3, 4, 4

↑

MAD(n) = 2

31) data $W = \underset{n \times p}{X} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = (\underset{x_1}{V_1}, \ldots, \underset{x_p}{V_p})$ (pth variable)

(X_1 for MLR)

↑ nth "case"

$$= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$ ← 1st case

2nd variable.

32) The coordinatewise median $MED(W)$

$= (MED(x_1), \ldots, MED(x_p))^T$ where

MED(X_i) is the sample median of the ith column.

33] Know how to get $\bar{x}$ and the coordinatewise median for a small data set.

ex] 4 measurements on 5 trees W^T or x^T

N	E	S	W
72	66	76	77
60	53	66	63
56	57	64	58
41	29	36	38
32	32	35	36
261	237	277	272
			Σ sum

ordered (columns → rows)

32	41	56	60	72 N
29	32	53	57	66 E
35	36	64	66	76 S
36	38	58	63	77 W
		MED(W^T)		

show work

$$\frac{1}{5}\begin{pmatrix} 261 \\ 237 \\ 277 \\ 272 \end{pmatrix} = \begin{pmatrix} 52.2 \\ 47.4 \\ 55.4 \\ 54.4 \end{pmatrix} = \bar{x}$$

See HW 3

34] The ith Mahalanobis distance

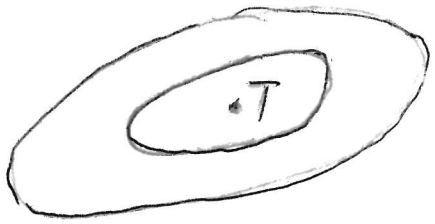
$D_i = \sqrt{D_i^2}$ where $D_i^2 = D_i^2(\underset{\text{vector}}{T(W)}, C(W))$

$= (x_i - T(W))^T (C(W))^{-1} (x_i - T(W))$

$= (x_i - T)^T C^{-1} (x_i - T) = D_{x_i}^2(T, C)$.

$D_x^2(T, C) = (x - T)^T C^{-1} (x - T)$ ← eigenvectors determine the axes of the hyperellipsoid

T ← center of hyperellipsoid



contours of constant distance of x from T.

35] $D_T^2(x, C) = D_x^2(T, C)$.

(distance of x from T = distance of T from x)

36] The pop squared Mahalanobis distance

$D_x^2(\mu, \Sigma) = (x - \mu)\Sigma^{-1}(x - \mu)$ where

μ = pop location vector and Σ = pop dispersion matrix, often the pop. cov matrix in low dimensions.