Lecture 7 　　26/4/22

Computing LS solutions using SVD

Why?

- SVD handles rank-deficient Z and can identifies covariets causing the deficiency
- SVD is numerically stable, as opposed to solving $Z^T Z \beta = Z^T y$ using Gauss elimination
- SVD is commonly used in multivariate statistics and ML

The SVD

$$Z \in \mathbb{R}^{n \times p}$$

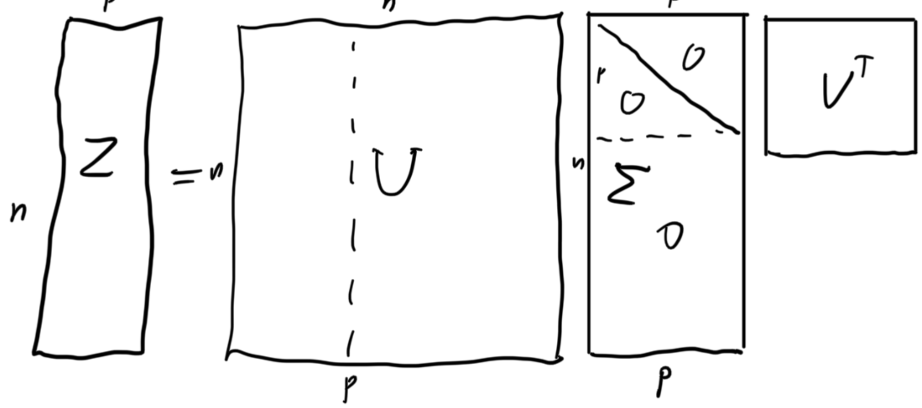
$$Z = U \Sigma V^T$$

where

$$U \in \mathbb{R}^{n \times n} \quad \Sigma \in \mathbb{R}^{n \times p} \quad V \in \mathbb{R}^{p \times p}$$

$$U^T U = U U^T = I_n \qquad V^T V = V V^T = I_p$$

$$\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_k), \quad \sigma_i \geq 0, \quad k = \min\{n, p\}$$



Standard case $n > p$

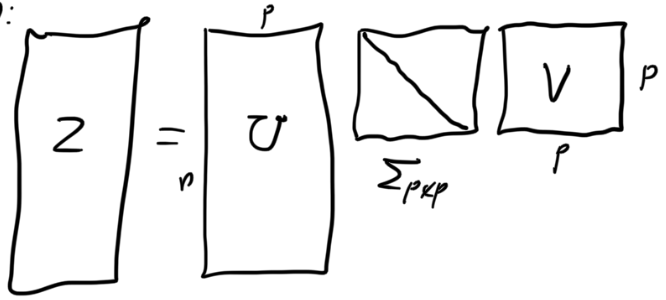
$$\tilde{y} = Z\tilde{\beta} = \underbrace{U \underbrace{\Sigma \underbrace{V^T \beta}_{I}}_{II}}_{III}$$

(I) Rotating β by muliplying it by V^T

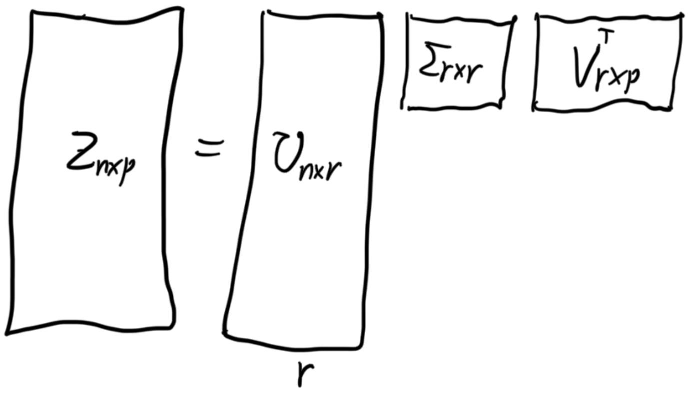
(II) Streching each coordinate of $V^T\beta$ by Σ

(III) Rotating $\Sigma V^T \beta$ by U

Skinny SVD:



In the general case where $\operatorname{rank}(Z) \leq p$, we can also eliminate coordinates corresponding to $\sigma_i = 0$.
Assume $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p$, we get



$r = \operatorname{rank}(Z)$ is the number of $\sigma_i > 0$.

Also note that eigenvalues of

$$Z^T Z = V \underbrace{\Sigma^T \Sigma} V^T \text{ are } \sigma_1^2, \dots, \sigma_p^2.$$

$\tilde{y} = Z\tilde{\beta}$

$$\|y - \tilde{y}\|^2 = \|y - Z\tilde{\beta}\|^2 = \|y - U\Sigma V^T \tilde{\beta}\|^2$$

$$= \| \underbrace{U^T y}_{y^*} \; U^T U \Sigma \underbrace{V^T \tilde{\beta}}_{\beta^*} \|^2 \qquad \sigma_i = 0 \text{ if } i \geq r, \; i = 1, \dots, n$$

$$= \| y^* - \Sigma \beta^* \|^2 = \sum_{i=1}^{n} (y_i^* - \sigma_i \beta_i^*)^2$$

$$= \sum_{i=1}^{r} (y_i^* - \sigma_i \beta_i^*)^2 + \sum_{i=r+1}^{p} (y_i^* - 0\beta_i^*)^2 + \sum_{i=p+1}^{n} (y_i^* - 0 \cdot \beta_i^*)^2$$

$$= \underline{\sum_{i=1}^{r} (y_i^* - \sigma_i \beta_i^*)} + \underline{\sum_{i=r+1}^{n} (y_i^*)^2}$$

The LS solution satisfies:

$$\beta_i^* = \begin{cases} y_i^* / \sigma_i & i = 1, \dots, r \\ \text{anything} & i = r+1, \dots, p \end{cases}$$

usually we take $\beta_i^* = 0$, $i = r+1, \dots, p$

to get the shortest solution

overall:

(I) $Z = U \Sigma V^T$

(II) $y^* := U^T y$

(III) $\beta_j^* = \begin{cases} y_i^* / \sigma_j & \sigma_j \neq 0 \\ 0 & \sigma_j = 0 \end{cases}$

(IV) $\hat{\beta} = V \beta^*$

computing SVD is done in $O(np^2)$

The same SVD can be used for multiple y's

$\boxed{\text{ANOVA}}$

Recap:

- The main idea: test if there exists a significant difference in the marginal means between subgroups required means between subgroups

- The cell means model:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

$$i = 1, \dots, k \quad j = 1, \dots, n_i$$

- When $k=2 \Rightarrow$ use t-test

- The effect means model,

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\sum \alpha_i n_i = 0$$

The main statistical problem is testing: $H_0: \mu_1 = \dots = \mu_k$

or $H_0: \alpha_1 = \dots = \alpha_k = 0$

$$Y \sim N(\beta Z, \sigma^2 I)$$

$$\beta = (\mu_1, \dots, \mu_k)^T$$

$$y = (Y_{11}, Y_{12}, \dots, Y_{1n_1}, Y_{21}, \dots, Y_{k1}, \dots Y_{kn_k})^T$$

$$Z = \begin{array}{c} \\ \\ n_1 \\ \\ \\ \\ n_2 \\ \\ \\ n \end{array} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & & & \vdots \\ \vdots & & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & \vdots & & \\ 0 & 1 & & \\ \vdots & & \vdots & 0 \\ \vdots & & 0 & 1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \end{bmatrix}$$

$$\hat{\beta} = (Z^T Z)^{-1} Z^T y = \begin{bmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \end{bmatrix}$$

$$\bar{y}_{i\cdot} := \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \qquad \begin{bmatrix} \vdots \\ \bar{y}_{k\cdot} \end{bmatrix}$$

- For the effect model:

$$\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}$$

$$\hat{\mu} = \bar{y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{k} \bar{y}_{i\cdot} \cdot n_i = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$$

- The SS for the cell means:

$$SS_{within} := SS_{res} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$$

- The total SS:

$$SS_{tot} = \sum_{i=1}^{k}\sum_{j=1}^{n_j} (y_{ij} - \bar{y}_{\cdot\cdot})^2$$

- We can test H_0 using the extra SS principle:

$$F = \frac{\frac{1}{k-1}(SS_{tot} - SS_{within})}{\frac{1}{n-k} SS_{within}}$$

reject H_0 at level α if

$$F > F^{1-\alpha}_{k-1, n-k}$$

- we also define

$$SS_{between} := SS_{fit} = SS_{tot} - SS_{within}$$

$$= \sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

$$SS_{tot} = SS_{between} + SS_{within}$$

- We usually summarize values in so-called ANOVA table:

S	DF	SS	MS	F

Source	DoF	SS	MS	F
Groups	$k-1$	$SS_{between}$	$SS_{between}/k-1$	$\frac{MS_{between}}{MS_{within}}$
Error	$n-k$	SS_{within}	$SS_{within}/n-k$	
Total	$n-1$	SS_{tot}		

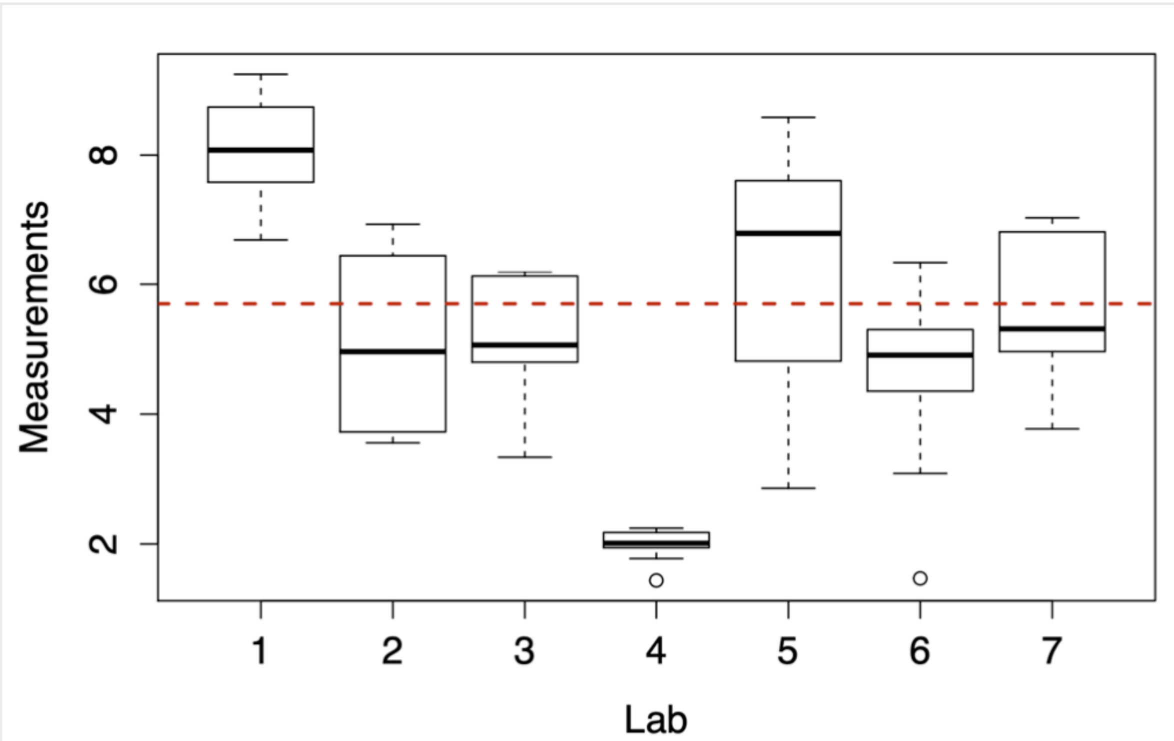
$$MS_{within} = s^2$$

Example: we send 10 pills to each of 7 different labs:

10 pills

Lab 1 . . . . Lab 7

H_0: there is no significant difference in measurements accross all the labs



We have the ANOVA table:

$n = 7 \times 10 = 70$

	DoF	SS	MS	F	p
Labs	6	[illegible]	[illegible]	[illegible]	[illegible]

[illegible]	6	0.125	.021	[illegible]	1.45×10[illegible]
Error	63	0.231	.0037		
Total	69	0.356			

We reject H_0

Contrasts

- suppose that we detected some effect, which are the groups causing it?

- We can check differences between individual groups:

$$\left(\frac{\bar{y}_{i.} - \bar{y}_{j.}}{s\sqrt{\frac{1}{n_j} + \frac{1}{n_i}}} \right. \sim t_{n-k} \qquad s^2 = \frac{SS_{res}}{n-k}$$

there are $\binom{k}{2} = \frac{k(k-1)}{2}$ such comparisons

- Is it possible that we fail to reject $H_0: \mu_1 = .. = \mu_k$ but find some of the group differences significant? Yes

- Is it possible that we reject H_0 but do not find any of the differences significant? Yes

- We can also look at <u>contrast</u>:

suppose that we want to test the effectiveness of three detergents with phosphates against four detergents without.

H_0: phosphates detergents are no different from non-phosphates

We use

we use:

$$t = \frac{\underbrace{\bar{y}_{1\cdot} + \bar{y}_{2\cdot} + \bar{y}_{3\cdot}}_{3} - \underbrace{\bar{y}_{4\cdot} + \bar{y}_{5\cdot} + \bar{y}_{6\cdot} + \bar{y}_{7\cdot}}_{4}}{s\sqrt{\frac{1}{9}\left(\frac{1}{n_1}+\frac{1}{n_2}+\frac{1}{n_3}\right)+\frac{1}{16}\left(\frac{1}{n_4}+\dots+\frac{1}{n_7}\right)}}$$

$s^2 = MSE_{within}$

$$t \sim t_{n-7}$$

- Another option: a single product with some "treatment" against four other product without:

$$t = \frac{\bar{y}_{1\cdot} - \frac{1}{4}\left(\bar{y}_{2\cdot} + \dots + \bar{y}_{5\cdot}\right)}{s\sqrt{\frac{1}{n_1} + \frac{1}{16}\left(\frac{1}{n_2} + \dots + \frac{1}{n_5}\right)}}$$

- In general:

$\sum_{i=1}^{k} \lambda_i = 0 \quad \sum \lambda_i^2 > 0$

$$t = \frac{\sum_{i=1}^{k} \lambda_i \bar{y}_{i\cdot}}{s\sqrt{\sum_{i=1}^{k} \frac{\lambda_i^2}{n_i^2}}} \sim t_{n-k}$$

Example:

	Potassium	No Potassium
sulfur	$\bar{y}_{1\cdot}$	$\bar{y}_{2\cdot}$
No sulfur	$\bar{y}_{3\cdot}$	$\bar{y}_{4\cdot}$

We can examine several effects:

- Sulfur: $\bar{y}_{1\cdot} + \bar{y}_{2\cdot} - \bar{y}_{3\cdot} - \bar{y}_{4\cdot}$

- Pottassium: $\bar{y}_{1.} + \bar{y}_{3.} - \bar{y}_{2.} - \bar{y}_{4.}$
- combined effect: $\bar{y}_{1.} - \frac{\bar{y}_{4.} - \bar{y}_{2.} - \bar{y}_{3.}}{3}$

Multiple Comparisons

- How can we find out which groups are motivating the rejection of the null?
- We can all pairs using $\binom{k}{2}$ t-tests However, these additional tests inflate the prob. of making Type I error (falsely rejecting the null)

  if $k=10$, then we do 45 tests and reject each one under the null w.p. α, so $Pr(\text{reject}) > \alpha$

Muliple Hypothesis Testing

- multiple comparisons is a special case of muliple hypothesis testing.
- We formulate a family of m null hypotheses $\{H_{0,i}\}_{i=1}^{m}$ and make up a test with a family-wise type I error rate (FWER) of at most α:

$$H_0 = \bigcap_{i=1}^{n} H_{0,i} \qquad Pr(\text{reject } H_0 \mid H_0 \text{ is true}) \le$$

Bonferroni's Union Bound

- We have m tests (e.g. $m = \binom{k}{2}$)
- We conduct each test at level α/m (e.g. for t-tests, we rejet

based on $t_{n-k}^{1-\frac{\alpha}{2m}}$)

- Under this procedure:

  $Pr(\text{reject something} \mid H_0 \text{ is true})$

  $$\leq \sum_{l=1}^{m} Pr(\text{reject } l\text{-th test} \mid H_0 \text{ is true})$$

  $$\leq \sum_{l=1}^{m} \frac{\alpha}{m} = m \cdot \frac{\alpha}{m} = \alpha$$