Condition Number (of a matrix)

Rel. stability of solving $Az = b$

A is $N \times N$

want to find z.

$$M = \max_{x} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

max. amt that A stretches a vector

why $\|x\|$ a number

$$\frac{\|Ax\|}{\|x\|} = \|A \underbrace{x/\|x\|}_{y \to \|y\|=1}\|$$

$$m = \min_{x} \frac{\|Ax\|}{\|x\|} = \min_{\|x\|=1} \|Ax\|$$

min amt A shrinks any vector

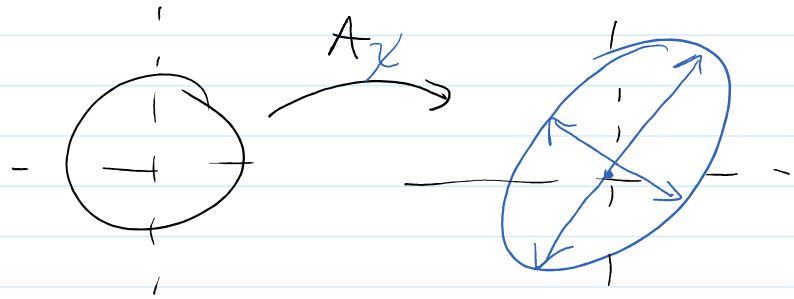
<u>notice</u>: if A is singular then for some x we have $Ax = 0$

so $m = 0 \iff A$ singular.

Defn: Condition Number

$$K(A) = M/m = \frac{\text{max stretching}}{\text{min stretching}}$$

$$K(A) = M/m = \frac{\text{max stretching}}{\text{min stretching}}$$



<u>Imagine</u> $Az = b$ and $A\delta_z = \delta_b$

then

$$A(z+\delta_z) = b + \delta_b$$

By defn

$\forall x$ $\|Ax\| \le M\|x\|$

and $\|Ax\| \ge m\|x\|$

So

$$\boxed{\|b\| = \|Az\| \le M\|z\|} \quad (1)$$

and

$$\boxed{\|\delta_b\| \; \|A\delta_z\| \ge m\|\delta_z\|} \quad (2)$$

re-arrange

$$\boxed{}$$

by 2 $$\boxed{\frac{\|\delta_b\|}{m} \ge \|\delta_z\|}$$

<u>multiply:</u>

$$\boxed{\underbrace{\frac{\|\delta_b\|}{\|b\|}}_{\kappa(A)} \underbrace{\frac{M}{m}} \ge \underbrace{\frac{\|\delta_z\|}{\|z\|}}}$$

rel. change in b $\quad K(A) \quad$ rel. change in z

<u>interp</u>: rel change in b $\rightsquigarrow$ $K(A)$ to change in z.

If $K(A)$ is small then system is stable b/c rel. large changes in b propogate to small changes in soln.

if $K(A)$ is large then small changes to b lead to LARGE changes in soln.

If A is singular (not invertible)

then $K(A) = \dfrac{M}{m} \leftarrow m = 0 \quad = \infty$

in which case $Az = b$

(∞ change in z — no change in b).

Relate back to SVD

<u>Note</u>: If Q is orthogonal then $\|Qx\| = \|x\|$

<u>pf.</u>

$$\|Qx\|^2 = (Qx)^T(Qx) = x^TQ^TQx = x^Tx$$

interp: orthog. transf.

rotate space - def stretch $= \|x\|^2$

Let $A = UDV^T$

then

$$M = \max_{\|x\|=1} \|Ax\| = \max_{\|x\|=1} \|UDV^Tx\|$$

$$= \max_{\|x\|=1} \|D\underbrace{V^Tx}_{y}\| \quad \text{where } \|y\|=1$$

$$= \max_{\|y\|=1} \|Dy\| \qquad D \to \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_N)$$

$$= \max_{\|y\|=1} \sqrt{(\sigma_1 \underset{1}{y_1})^2 + (\sigma_2 \underset{0}{y_2})^2 + \cdots + (\sigma_N \underset{0}{y_n})^2}$$

set $y_1 = 1$, others to 0 $\qquad \sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots \sigma_N$

$$= \sigma_1$$

Similarly $m = \sigma_N$

hence $\boxed{K(A) = \sigma_{max} / \sigma_{min}}$

<u>Why do we care?</u>

Recall for LS regression, we said

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

this found b/c if $\boxed{\frac{\partial L}{\partial \beta} = 0}$ squared loss

this leads to system of eqns

$$\boxed{\underbrace{(X^T X)}_{A}\underbrace{\beta}_{z} = \underbrace{X^T Y}_{b}}$$

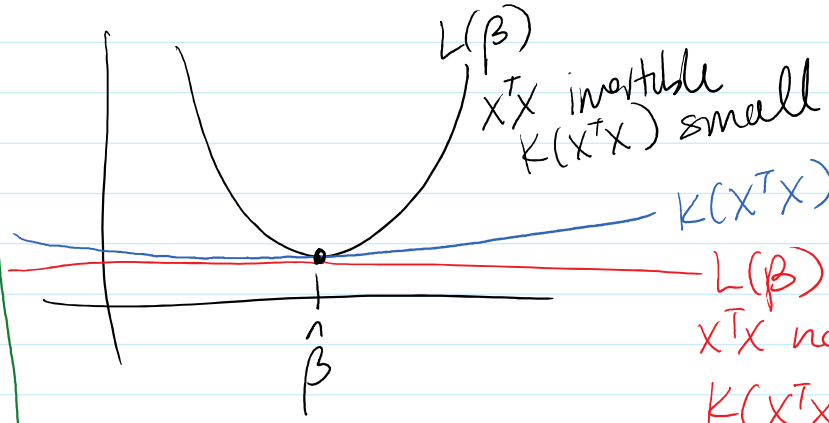
Normal equations

the stability of $\hat{\beta}$ (my soln) depends on $K(X^T X)$.

Q: In reality when do we get an "ill-conditioned" regression problem i.e. $K(X^T X)$ large?

We said previously that $\hat{\beta}$ unique $\Longleftrightarrow$ $X^T X$ invertible

$(\text{rank}(X) = N)$



$K(X^T X)$ large: $X^T X$ "close" to not invertible

$L(\beta)$

$X^T X$ not invertible

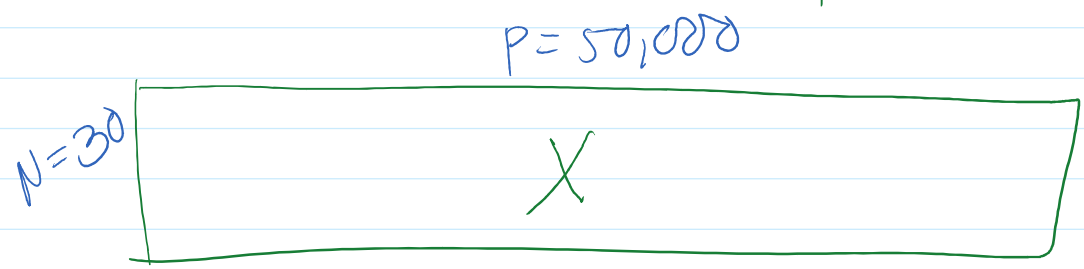
$K(X^T X) = \infty$

Ex. X^TX not invertible if any cols of X are a LC of any other

Ex. more likely X^TX close to not invertible one variable $\approx$ LC of others

Ex. If $P > N$ then X^TX is not invertible.

e.g. X is gene expression for $P = 50{,}000$ genes of $N = 30$ people.



How do we deal w/ this?

(*) (1) Variable selection (remove variables to better condition our problem)

(2) Shrinkage (Ridge / LASSO)

(3) Dimensionality Reduction (PCA regression)

③ Dimensionality Reduction (PCA regression)

↳ unsupervised analysis.

Variable Selection

Goal: pick a subset of important variables and just use those.

Q: How do we define "important" variables,
↳ How do I define a good set of vars?

Two approaches:

① individual metrics

e.g. p-value for each variable, choose vars w/ smallest p-values

problem: perfomance of one var might be affected by inclusion/exclusion of others

② metrics for sets of variables

idea: build a model w/ diffent collections of vars and use "best" perfomy group.

Careful: don't look at training metrics
b/c training metrics ↑ as p ↑
(more flex as p ↑)

Solns: ① train/val/test split.

② penalize train metrics by p.
(classic)

Ex. Adjusted R^2 (↑ R^2 = better)

$$R^2_{adj} = 1 - \frac{N-1}{N-p-1}(1-R^2)$$

as p ↑ R^2_{adj} ↓

RSS-based metrics (↓ RSS = better)

① Mallow's C_p

$$C_p = \frac{1}{N}\left(RSS + 2p\hat{\sigma}^2\right)$$

est. of error var

increases w/ p

penalty

② AIC: Akaike's Information Criterion

$$AIC = \frac{1}{N\hat{\sigma}^2}\left(RSS + \underbrace{2p\hat{\sigma}^2}_{\text{penalty for large } p}\right)$$

③ BIC : Bayesian Inform. Criterion

$$BIC = \frac{1}{N}\left(RSS + \underbrace{\log(N)\, p\, \hat{\sigma}^2}_{\text{penalty}}\right)$$

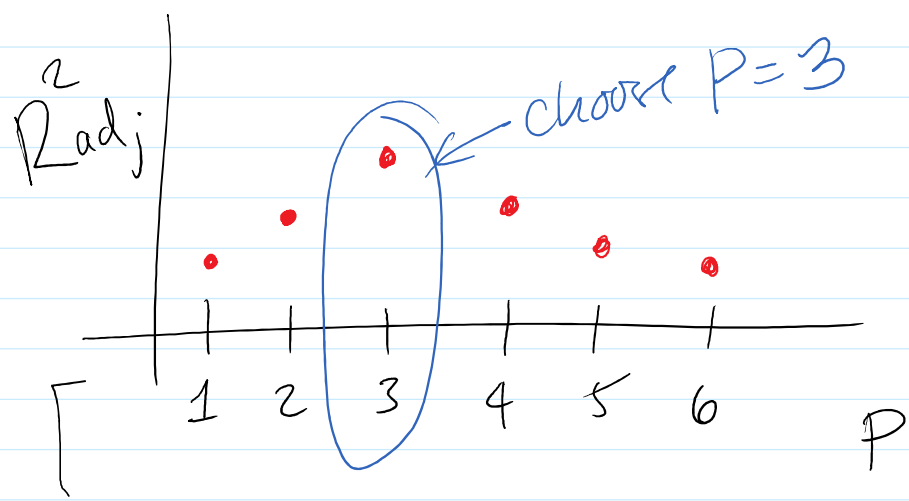
Systematic Search: (greedy heuristics)

① Forward stepwise Selection

→ start w/ no covs

→ add variables 1 at a time to regr. model

→ add var that gives best (inc/dec) in metric



② Backward stepwise Selection

→ start w/ all covs.

→ remove one at a time
remove var. that hurts model metric
the least.

<u>Ideally</u>: test all subsets of vars.
(not computation possible for large P)