"Linear regression" (or Just least squares...)

Let

$x_i \in \mathbb{R}^n$, $i = 1, \dots, m$

$y_i \in \mathbb{R}$, $i = 1, \dots, m$

$$\underline{X} := \begin{pmatrix} - x_1 - \\ \vdots \\ - x_m - \end{pmatrix} \in \mathbb{R}^{m \times n}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^n$$

Want to solve

$$\underline{X}\hat{w} = y \qquad (*)$$

Case I $m < n$.

- $\underline{X}$ is fat
- few eqns, many unknowns
- $N(\underline{X}) \neq \{0\}$
  - obvious way to see this, what if $\underline{X} = e_1^T$.

To make life easier, assume for now that $\operatorname{rank}(\underline{X}) = m$. Then $\dim \operatorname{col} \underline{X} = m$, so $\underline{X}$ is surjective & $y \in \operatorname{col} \underline{X}$. Since $N(\underline{X}) \neq \{0\}$, $\exists$ many solutions to $(*)$.

~~Aside~~ ~~dim col $\underline{X}$ = dim row $\underline{X}$ follows from rank-nullity. By construction, we have m lin. indep rows~~

How can we narrow down solns?

options:

- Pick $\hat{w}$ w/ small $\|\hat{w}\|_2$
  - Jives w/ notion of regularization & "low complexity" solution
  - also, can work w/ $\|\hat{w}\|_2^2$, and use property of Hilbert spaces, which should make analysis fairly simple
- Pick "sparsest" w. Small $\|\hat{w}\|_0$ norm. ~ compressed sensing. Revisit sometime

We will go w/ option 1 for now.

If $\hat{w}, w$ both solve $(*)$, then $\hat{w} - w \in N(X)$.

If $\hat{w}$ lies in the rowspace of X, then $\hat{w} \perp (\hat{w} - w)$, since $\text{row}(X) \perp N(X)$. Suppose $w \neq \hat{w}$. In that case,

$$\|\hat{w}\|^2 = \langle w + (\hat{w} - w), w + (\hat{w} - w) \rangle$$

$$= \|w\|^2 + \underbrace{\|\hat{w} - w\|^2}_{> 0} > \|w\|^2.$$

Hence, If $\exists$ a soln in the rowspace, it has minimum norm and is unique.

To find such a solution, we need a map that brings col X back to row X. In general, X is bijective from row X to col X & the pseudo inverse provides the reverse mapping. For now, we'll cheat a little rather than use the general pseudo inverse.

Claim

$$\text{rank } X = \text{rank } X^T X = \text{rank } X X^T$$

pf

1. Claim rank X = rank X^T. dont want to show
2. Use rank-nullity. (ie show X & $X^T X$ have same Nullspace.)

- Suppose $Xw = 0$. Then $X^T X w = 0$ So, $N(X) \subset N(X^T X)$
- Suppose $X^T X w = 0$. Then $w^T X^T X w = 0 \Rightarrow |Xw|^2 = 0$
  $\Rightarrow Xw = 0 \Rightarrow N(X^T X) \subset N(X)$.

Claim follows by rank-nullity.

$X : m \times n$

Since we assumed $\operatorname{rank} X = m$ (X fat, linearly ind. rows), we get $\operatorname{rank} X X^T = m \iff (X X^T)^{-1}$ exists.

Want to solve $X \hat{w} = y$

Anything we can set $\hat{w}$ to to make this work?

Let $\hat{w} = X^T (X X^T)^{-1} y$

$$X \hat{w} = X X^T (X X^T)^{-1} y = y.$$

Note that $\hat{w} = X^T \underbrace{(X X^T)^{-1} y}_{=: z}$. This means $\hat{w} \in \operatorname{row}(X^T)$

Hence, $\hat{w}$ is our unique soln.

Note:

- $X^+ := X^T (X X^T)^{-1}$ is pseudo-inverse when rows are lin. indep.
- If X were rank deficient, what would it mean? Most importantly, $\dim \operatorname{col} X < m$. (since $\dim \operatorname{row} X = \dim \operatorname{col} X = \operatorname{rank} X$)
- Analysis would go through if we restricted to $y \in \operatorname{col}(X)$ and had a more general pseudo-inv. taking $\operatorname{col} X$ back to $\operatorname{row} X$.

Case II: $m = n$

- dont care right now since we're looking at full row/col rank matrices to isolate effects of each asymmetry separately first.

Case III $m > n$

- X is tall
- ~~linearly indep. cols~~
- $\dim \operatorname{col} X \le n < m$
  - so, X not surjective
- What about $N(X)$?
  - may or may not be empty
- If we assume $\operatorname{rank} X = n$, then
  - X has lin indep cols
  - $\dim \operatorname{col} X = n$
  - $N(X) = \{0\}$
    - rank - nullity:

$$A \in \mathbb{R}^{m \times n} \qquad \underbrace{\operatorname{rank} A}_{\substack{(= \dim \operatorname{col} A \\ = \dim \operatorname{row} A)}} + \dim N(A) = n.$$

Assume $\operatorname{rank} X = n$.

- Buys you: $N(X) = \{0\}$.

Suppose $\hat{y} \in \operatorname{col} X$. claim: $\exists$ a unique $\hat{w}$ solns $X\hat{w} = \hat{y}$ $(**)$

Pf $\operatorname{rank} X^T X = n \implies (X^T X)^{-1}$ exists. Also, since $\hat{y} \in \operatorname{col} X$, a $\hat{w}$ solns $(**)$ $\exists$. observe

$$X\hat{w} = \hat{y} \iff (X^T X)^{-1} X^T X \hat{w} = (X^T X)^{-1} X^T y$$

$$\iff \hat{w} = (X^T X)^{-1} X^T y$$

Also, $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ has nullspace $= \{0\}$

└ why? TODO

This proves claim.

Suppose y possibly outside col $\mathbf{X}$. Want $\hat{w}$ s.t.

$$\hat{w} \in \operatorname*{argmin}_{w} \|\mathbf{X}w - y\|_2$$

But we're in a Hilbert space, so soln to this is just projection

i.e., $\|\hat{y} - y\|_2^2 \leq \|y' - y\|^2 \quad \forall y' \in \text{col } \mathbf{X}$

for $\hat{y}$ in col $\mathbf{X}$ if $\hat{y}$ is the unique vector satisfying

$(\hat{y} - y) \perp \text{col } \mathbf{X}$.

<u>TODO:</u> Formally, recall Hilbert proj thm.

Hence, the optimal $\hat{w}$ satisfies

$$(\mathbf{X}\hat{w} - y) \perp \text{each col of } \mathbf{X}$$

$$\Updownarrow$$

$$(\mathbf{X}\hat{w} - y)^T\mathbf{X} = \vec{0}$$

$$\iff \hat{w}^T\mathbf{X}^T\mathbf{X} = y^T\mathbf{X} \iff \boxed{\mathbf{X}^T\mathbf{X}\hat{w} = \mathbf{X}^Ty.}$$

Know $(\mathbf{X}^T\mathbf{X})^{-1}$ exists.

$$\hat{w} = (\mathbf{X}^T\mathbf{X})\mathbf{X}^Ty$$

Observations

- $\underline{X}^+ := (\underline{X}^T \underline{X})^{-1} \underline{X}^T$ is pseudo-inv. when cols are lin indep.

- If $\underline{X}$ were rank deficient, what would it mean?
  $N(\underline{X}) \neq 0$. So, many $\hat{w}$ satisfy $\underline{X}\hat{w} = \hat{y}$
  where $\hat{y} := \text{proj}_{col(X)}(y)$. So, many least squares soln.
  Which to pick?

Computational Consideration: Have to invert $(X^T\underline{X}) \in \mathbb{R}^{n \times n}$.
What if n (dim of feature space) is big?

Can solve optimization problem directly.

$$\min_w \underbrace{\| \underline{X} w - y \|^2}_{=: J(w)}$$

$$\nabla J(w) = (\underline{X} w - y)^T \underline{X}$$

double check. Can confirm by computing $\frac{\partial J}{\partial w_i}$ and seeing if this is i-th component of $(\underline{X}w - y)^T \underline{X}$.

Use GD:

$$w_{t+1} = w_t - (\underline{X} w - y)^T \underline{X}$$

- GD: $O(mn)$ flops per iteration
- Normal eq. / pseudo-inverse: $\approx O(n^3)$
- If $m << n$, GD seems probably better.

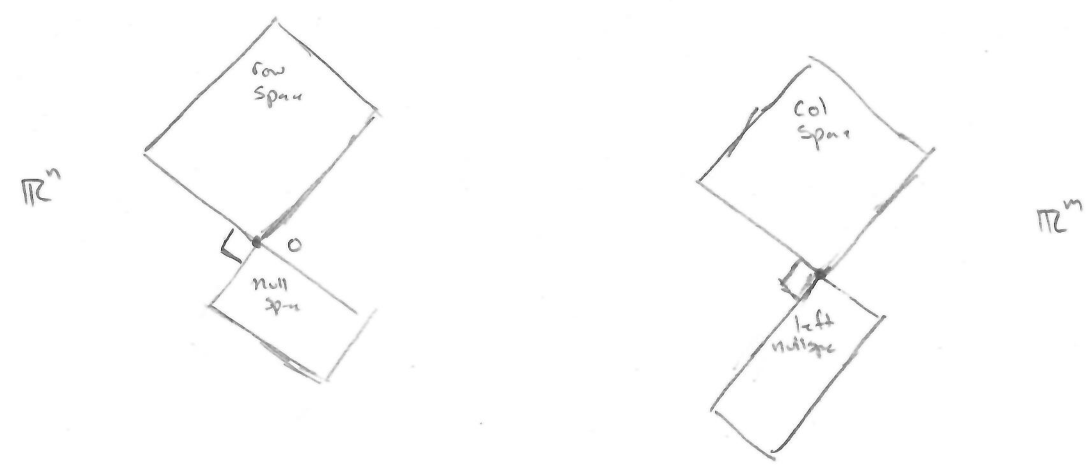
Omissions / taken on faith so far

- rank A = rank A^T
- rank-nullity thm
- Hilbert proj. thm.
- form of ∇J

SVD

Diagonalization makes life incredibly easy, when possible. Life is particularly easy when eigenvectors are orthogonal. The SVD provides an orthonormal diagonal decomposition for arbitrary (non-square) matrices. The catch: The basis vectors used for domain / range can be different. So, SVD not necessarily suitable for applications w/ powers of matrices since you don't get $A^2 = S\Lambda S^T S \Lambda S^T = S \Lambda^2 S^T$. But useful when this doesn't matter.

$$A \in \mathbb{R}^{m \times n}$$



Let $r = \operatorname{rank} A$

We want an orthonormal basis $v_1, \ldots, v_r$ & $u_1, \ldots, u_r$ for col space s.t.

$$A v_i = \sigma_i u_i$$

What's a good candidate for $(v_i)_{i=1}^r$? How about eigenvectors of

$A^TA \in \mathbb{R}^{n \times n}$ basis for rowspace $\quad AA^T \in \mathbb{R}^{m \times m}$

Note: A^TA & AA^T are positive semi def. Why? b/c $x^TA^TAx = \|Ax\|^2 \geq 0 \ \forall x$.

Have, $A^TAv_i = \sigma_i^2 v_i \Rightarrow v_i^T A^TAv_i = \sigma_i^2 v_i^T v_i \Rightarrow \|Av_i\| = \sigma_i$

~~Let $u_i = Av_i$~~

Let $\hat{u}_i = Av_i$. Lets see what happens.

Have

$$A^T A v_i = \sigma_i^2 v_i \iff A A^T A v_i = \sigma_i^2 A v_i$$

$\iff A A^T \hat{u}_i = \sigma_i^2 \hat{u}_i$. so, $u_i := \frac{1}{\sigma_i} A v_i$ is a unit eigvector of AA^T

Have

- $(u_i)_{i=1}^r$ orthonormal basis for col space
- $(v_i)_{i=1}^r$ orthonormal basis for row space
- $A v_i = \sigma_i u_i$
- $\sigma_i > 0$, $\sigma_i = \sqrt{\lambda_i}$, where $\lambda_i \in \sigma(A^T A) \setminus \{0\}$

Complete each basis w/ orthonormal basis for nullspace / left nullspace to get $(v_1, \dots, v_r, \dots v_n)$, ..., $(u_1, \dots, u_r, \dots u_m)$

- V is $n \times n$ matrix w/ cols (v_i)
- U is $m \times m$ w/ cols (u_i)
- Σ is ? w/ diagonal entries σ_i
  $n \times n$ = dim A. see

$A v_i = \sigma u_i \iff AV = U\Sigma \iff \boxed{A = U \Sigma V^T}$

SVD

Pseudo - inverse

We know A is an injective map from row span to col. space.

Want to construct an inverse that will bring any $y \in col(A)$ back to unique x s.t. $Ax = y$. This is what pseudo-inv. does.

This all becomes easy if we can map $\sigma_i u_i$ back to v_i

i.e., want A^+ s.t.

$$A^+(\sigma_i u_i) = v_i$$

$$\Updownarrow$$

$$A^+ u_i = \frac{1}{\sigma_i} v_i$$

$$\Updownarrow$$

$$A^+ U = V\Sigma^+$$, where Σ^+ is same as Σ but w/ $1/\sigma_i$ on diag. when $\underbrace{\sigma_i \neq 0}$

holds by def of σ_i, since only defined up to $i = \text{rank}(A)$

$$\Updownarrow$$

$$\boxed{A^+ := V\Sigma^+ U^T}$$

As a sanity check, say $Ax = y$ w/ $x \in row(A)$, $x = \sum_{i=1}^{r} \alpha_i v_i$.

$y = \sum_{i=1}^{r} \alpha_i \sigma_i u_i \Rightarrow A^+ y = \sum_{i=1}^{r} \alpha_i A^+ \sigma_i u_i = \sum_{i=1}^{r} \alpha_i v_i = x$.

So,

- $A^+ A x = x$, for $x \in row(x)$.

other implications:

More generally,

- If $x \in \mathbb{R}^n$, $A^+ A x = \text{proj}_{\text{row}(A)}(x)$

- If $y \in \mathbb{R}^m$, $A^+ y = A^+ (\hat{y} + z)$ for $\hat{y} = \text{proj}_{\text{col}(A)}(y)$,

and any $z \in N(A^T)$. Continuing,

$$A y = \hat{x}$$

where $\hat{x}$ is unique element of $\text{row}(A)$ solving $A\hat{x} = \hat{y}$, w/ $\hat{y} = \text{proj}_{\text{col} A}(y)$.