For A with SVD

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

we call

$$A^+ = V \Sigma^{-1} U^T$$

the pseudo-inverse of A

Other names A^+ goes by:

- "Natural inverse"
- "Lanczos inverse"
- "Moore-Penrose inverse"

Least-squares example: simple extrapolation

$$\begin{bmatrix} y[0] \\ y[1] \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x[0] \\ x[1] \\ x[2] \\ x[3] \end{bmatrix}$$

We see the first two samples, but not the second two.

It is clear that $x[2]$ & $x[3]$ can be anything, and our observations $y[0]$, $y[1]$ will not change.

What does the least squares solution

$$\hat{x} = V\Sigma^{-1}U^T y + V_0 \kappa_0$$

look like?

Note that

$$A^TA = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

So we can take

$$V = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad V_0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(these choices are not unique, though)

$$AA^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{so} \quad U = [u_1 \; u_2] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and there is no V_0 (full row rank)

$(\Rightarrow \beta_0 = 0)$

$$\hat{x} = V\Sigma^{-1}U^T y + V_0 \kappa_0$$

$$= \begin{bmatrix} y[0] \\ y[1] \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \kappa_0[2] \\ \kappa_0[3] \end{bmatrix} \quad \text{for any } \kappa_0[2], \kappa_0[3]$$

The min-norm solution has $\kappa_0[2] = \kappa_0[3] = 0$

(usual case)

Pseudo-inverse summary

Given y, we set

$$\hat{x} = A^+ y = V\Sigma^{-1}U^T y$$

We know that

- there is no x whose residual $y - Ax$ is smaller than $y - A\hat{x}$

  In other words, there is no x such that Ax is closer to y that AA^+y.

- Of all the x with minimum residual, none has smaller norm than A^+y.

Also note that the pseudo-inverse always exists (every matrix has an SVD).

Explicit formulas for A^+ when A is full rank:

- If A is square & invertible, then

$$A^+ = A^{-1}$$

Check: $A = U\Sigma V^T$ $\quad U$ is $N \times N$, $U^TU = UU^T = I$

V is $N \times N$, $V^TV = VV^T = I$

$$A^+A = V\Sigma^{-1}\underbrace{U^TU}_{=I}\Sigma V^T = V\underbrace{\Sigma^{-1}\Sigma}_{=I}V^T = VV^T = I$$

- If A^TA is invertible (full column rank), then

$$A^+ = (A^TA)^{-1}A^T$$

This occurs when A is "tall & skinny" $(M > N)$

$$\begin{bmatrix} A \end{bmatrix}$$

and all the columns are linearly independent.
We will have $\rho = N$ non-zero singular values

Check:

$A = U\Sigma V^T$ Since $\rho = N$, V is $N \times N$ $\perp$ matrix
$VV^T = V^TV = I$
U is $M \times N$ $\perp$ matrix
$U^TU = I$

$A^TA = V\Sigma U^TU\Sigma V^T = V\Sigma^2V^T$

$(A^TA)^{-1}A^T = V\Sigma^{-2}V^TV\Sigma U^T = V\Sigma^{-1}U^T = A^+$ ✓

• If AA^T is invertible, then (full row rank)

$$A^+ = A^T(AA^T)^{-1}$$

This occurs when A is "short & fat"

$$\begin{bmatrix} & A & \end{bmatrix}$$

and all the rows are linearly independent.
We will have $\rho = M$ non-zero singular values.

<u>Check:</u>

$A = U\Sigma V^T$ U is $M \times M$, $\perp$, $UU^T = U^TU = I$
V is $N \times M$, $\perp$, $V^TV = I$

$$AA^T = U\Sigma V^T V \Sigma U^T$$
$$= U\Sigma^2 U^T$$
$$A^T(AA^T)^{-1} = V\Sigma U^T U \Sigma^{-2} U^T = V\Sigma^{-1}U^T = A^+$$

A^+ is as close to an inverse of A as possible

Left inverse:

Given $y = Ax$, we'd like $A^+y = A^+Ax \approx x$ for any x. That is, we'd like $A^+A = I$.

Of course this is not always possible.

But is there an H such that HA is closer to the identity than A^+A?

The answer is no. To see this, we consider the problem

$$(*) \qquad \min_H \|HA - I\|_F^2 \qquad \left(\|\cdot\|_F = \text{"Frobenius norm"}\right)$$

(Here the F-norm^2 of a matrix Q

$$\|Q\|_F^2 = \sum_{k,\ell} Q_{k,\ell}^2$$

= sum of the squares of the entries)

The solution to this is the H which make HA closest to the identity.

It is a fact (which comes from setting derivatives = 0) that the solution to (*) will obey

$$AA^T H^T = A$$

So, does $H = A^+ = V\Sigma^{-1}U^T$ work?

$$AA^T H^T = U\Sigma \underbrace{V^T V}_{I} \underbrace{\Sigma \underbrace{U^T U}_{I} \Sigma^{-1}}_{I} V^T = U\Sigma V^T = A$$

yes.

Right Inverse:

Our solution $\hat{x}$ should come as close to explaining the data as possible. We'd like

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

That is,

$$AA^+ y = y \qquad (\text{i.e. } AA^+ = I)$$

Of course, this might not be possible for all y. But is there an H such that AH is closer to the identity than AA^+?

Again, we look at

$$\min_H \|AH - I\|_2^2$$

After a little bit of work, we have that the solution to the above will obey

$$HH^TA^T = H$$

Does $H = V\Sigma^{-1}U^T$ work?

$$HH^TA^T = V\Sigma^{-1}\underbrace{U^TU}_{I}\underbrace{\Sigma^{-1}\underbrace{V^TV}_{I}\Sigma}_{I}U^T = V\Sigma^{-1}U^T = H$$

so yes.

Moral: $A^+ = V\Sigma^{-1}U^T$ is as close to an inverse of A as you can possibly have.

Stability

What happens when there is noise?

$$y = Ax + e \qquad A = U\Sigma V^T$$

Say we apply A^+ to y above to get $\hat{x}$:

$$\hat{x} = A^+ y = \underbrace{A^+ A x}_{= VV^T x = \tilde{x} = \text{"optimal estimate" when there is no noise.}} + A^+ e$$

How far is $\hat{x}$ from $\tilde{x}$?

(What is the reconstruction error $\|\hat{x} - \tilde{x}\|_2^2$?)

It is $$\|\hat{x} - \tilde{x}\|_2^2 = \|A^+ e\|_2^2$$

Suppose for a minute that $\|e\|_2^2 = 1$

What is the "worst case" for $\|A^+ e\|_2^2$

$$\max_{\|e\|_2^2 = 1} \|A^+ e\|_2^2 = \max_{\|e\|_2^2 = 1} \|V\Sigma^{-1} U^T e\|_2^2$$

Since $\|U^T e\|_2^2 \leq \|e\|_2^2$ (the columns of U are $\perp$)

the above is equivalent to

$$\max_{\|\beta\|_2^2 = 1} \|V\Sigma^{-1}\beta\|_2^2$$

Also, for any vector z

$$\|Vz\|_2^2 = \langle Vz, Vz \rangle = \langle z, V^T V z \rangle = \langle z, z \rangle = \|z\|_2^2$$

(the columns of V are $\perp$)

So we can further simplify to

$$\max_{\|\beta\|_2^2 = 1} \|\Sigma^{-1}\beta\|_2^2$$

The worst case β will have a 1 in the entry corresp. to the largest value on the diagonal of Σ^{-1}, and zero everywhere else.

Thus $$\max_{\|\beta\|_2^2 = 1} \|\Sigma^{-1}\beta\|_2^2 = \max_{k=1,\dots,p} 1/\sigma_k^2 = 1/\sigma_p^2$$

So $$\|\hat{x} - \tilde{x}\|_2^2 = \|V\Sigma^{-1}U^T e\|_2^2 \leq \frac{1}{\sigma_p^2} \cdot \|e\|_2^2$$

Similarly, $\|\hat{x} - \tilde{x}\|_2^2 \geq \frac{1}{\sigma_1^2} \cdot \|e\|_2^2$, so

$$\frac{1}{\sigma_1^2} \cdot \|e\|_2^2 \leq \|\hat{x} - \tilde{x}\|_2^2 \leq \frac{1}{\sigma_p^2} \cdot \|e\|_2^2$$

If σ_p is small, the worst case error can be <u>very</u> bad.

"Average case" error can also be related to the singular values.

Say e is iid Gaussian noise (AWGN)

$$e_k \sim \text{Normal}(0, \sigma^2)$$

The average measurement error is

$$E\|e\|_2^2 = M\sigma^2$$

The average reconstruction error is

$$E\|A^+ e\|_2^2 = \underbrace{\frac{1}{M}\cdot\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \cdots + \frac{1}{\sigma_p^2}\right)}_{\text{error magnification} \sim \text{average of the singular values}} \cdot M\sigma^2$$

If σ_p is tiny, $\frac{1}{\sigma_p^2}$ will dominate the rest of the terms.

Moral: Applying the pseudo-inverse A^+ can be disasterous in the presence of noise.

One way to protect against this is to truncate the SVD.

We have

$$A = U\Sigma V^T = \sum_{k=1}^{p} \sigma_k u_k v_k^T$$

Set a threshold σ_{min}. Find $p' \leq p$ so that $\sigma_{p'} \geq \sigma_{min}$, $\sigma_{p'+1} < \sigma_{min}$

i.e. $\sigma_1, \sigma_2, \ldots, \sigma_{p'}$ are above σ_{min}

$\sigma_{p'+1}, \ldots, \sigma_p$ are below σ_{min}

The truncated sum

$$A' = \sum_{k=1}^{p'} \sigma_k u_k v_k^T$$

is a good approximation to A (especially if σ_{min} is small compared to σ_1).

Then, instead of applying the pseudo inverse of A, we use $A'^{\dagger}$.

$$\hat{x} = A'^{\dagger} y = A'^{\dagger} A x + A'^{\dagger} e$$

The reconstruction error

$$\hat{x} - \tilde{x} = A'^{+}Ax - A^{+}Ax + A'^{+}e$$

$$= \underbrace{(A'^{+} - A^{+})Ax}_{\text{truncation error}} + \underbrace{A'^{+}e}_{\text{noise error}}$$

now has two parts

To see what the truncation error is, write

$$A'^{+} - A^{+} = \sum_{k=p'+1}^{p} \frac{1}{\sigma_k} v_k u_k^T$$

then

$$(A'^{+} - A^{+})Ax = \left(\sum_{k=p'+1}^{p} \frac{1}{\sigma_k} v_k u_k^T \right) \cdot \left(\sum_{\ell=1}^{p} \sigma_\ell \cdot u_\ell v_\ell^T x \right)$$

$$= \sum_{k=p'+1}^{p} v_k \cdot \langle v_k, x \rangle$$

= portion of x in the space we are cutting out

$$\| (A'^{+} - A^{+})Ax \|_2^2 =$$

Also,

$$A'^{+}e = \sum_{k=1}^{p'} 1/\sigma_k \langle u_k, e \rangle v_k$$

and

$$\|A'^{+}e\|_2^2 = \sum_{k=1}^{p'} \frac{1}{\sigma_k^2} |\langle u_k, e \rangle|^2$$

Note that

$$\underbrace{(A'^{+} - A^{+})Ax}_{\text{truncation error}} \perp \underbrace{A'^{+}e}_{\text{noise error}}$$

since $A'^{+}e \in \text{span}\{v_1, \ldots, v_{p'}\}$

$$(A'^{+} - A^{+})Ax \in \text{span}\{v_{p'+1}, \ldots, v_p\}$$

So the total error is

$$\|\hat{x} - \tilde{x}\|_2^2 = \sum_{k=p'+1}^{p} |\langle v_k, x \rangle|^2 + \sum_{k=1}^{p'} \frac{1}{\sigma_k^2} |\langle u_k, e \rangle|^2$$

Tikhonov Regularization

There is another way (other than truncation) to get stable solutions to least squares problems.

Recall that the pseudo-inverse will give us a solution that solves

$$(*) \qquad \min_x \|y - Ax\|_2^2$$

If A has singular values which are very small, $(*)$ could make x huge (in the direction of the v_k corresponding to tiny σ_k) in its effort to match y as closely as possible.

We can modify $(*)$ by adding a regularization term that penalizes x with large norm.

Instead of $(*)$, we solve

$$(**) \qquad \min_x \|y - Ax\|_2^2 + \delta \cdot \|x\|_2^2$$

keeps $\|x\|_2^2$ from getting too big

for some small $\delta > 0$.

It is not terribly hard to show (on homework) that (**) has a unique solution given by

$$\hat{X}_{Tik} = (A^TA + \delta I)^{-1}A^Ty$$

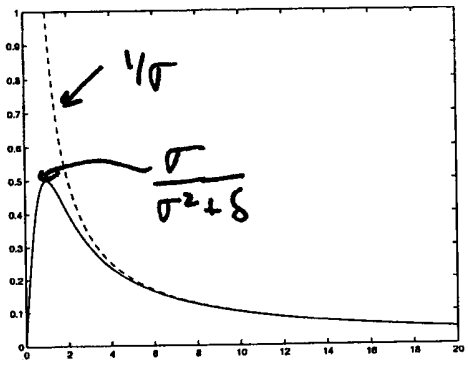
For small δ, $A^TA + \delta I$ will be close to A^TA, but it is <u>guaranteed</u> to be invertible.

In terms of the SVD of A, we can write

$$\hat{X}_{Tik} = \sum_{k=1}^{p} \frac{\sigma_k}{\sigma_k^2 + \delta} \cdot \langle u_k, y \rangle \cdot v_k$$

↳ this looks like $1/\sigma_k$ for $\sigma_k >> \delta$ and goes to 0 (dampens) for $\sigma_k << \delta$.

$\delta = 1$



$\sigma \longrightarrow$

Summary:

Pseudo-Inverse (least-squares):

- $\hat{X}_{LS} = V\Sigma^{-1}U^T y = \sum_{k=1}^{p} \frac{1}{\sigma_k} \langle u_k, y \rangle v_k$

- If A has full rank, we can compute without SVD using

$$\hat{X}_{LS} = (A^TA)^{-1}A^T y \qquad (M \geq N)$$

or

$$\hat{X}_{LS} = A^T(AA^T)^{-1} y \qquad (M \leq N)$$

- The solution can "blow up" (and probably will) if σ_p is very small

Truncated Pseudo-Inverse

- $\hat{X}_{Trunc} = \sum_{k=1}^{p'} \frac{1}{\sigma_k} \langle u_k, y \rangle v_k$

- simply throw away the terms in the sum for $\hat{X}_{LS}$ that correspond to small singular values
- need the SVD to compute this
- stable if p' is chosen appriately

Tikhonov Regularized Least-Squares

- $$\hat{X}_{Tik} = V\Sigma(\Sigma^2 + \delta I)^{-1}U^T y$$
$$= \sum_{k=1}^{p} \frac{\sigma_k}{\sigma_k^2 + \delta} \langle u_k, y \rangle v_k$$

- gradually "dampens" the components corresponding to very small singular values

- also has the advantage that we don't need the SVD to compute it:

$$\hat{X}_{Tik} = (A^T A + \delta I)^{-1} A^T y$$

Extensions

There are natural generalizations to both Tikhonov regularization and the truncated SVD.

For Tikhonov, in adding the regularization term $\delta\|x\|_2^2$ we are implicity assuming some type of structure on x — that is, that its energy is not too large. (**) on page 39 favors signals with smaller energy.

We can encourage different types of structure as well. For any matrix D w/ N columns, we can solve

$$\min_x \|y - Ax\|_2^2 + \delta\|Dx\|_2^2$$

this encourages Dx to have energy which is not too large.

A typical example is if D is an approximation to a derivative.

i.e. $D = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & 1 & -1 & \\ & & & \ddots & \end{bmatrix}$

Roughly speaking, this keeps the solution from "wiggling around too much". Another model which has a similar effect is

$$D = WF,$$

where F is a Fourier transform matrix and W is a diagonal matrix with positive entries which are larger in places (i.e. rows of F) corresponding to high frequencies and smaller at indices corresponding to low frequencies. This penalizes high-frequency components of the solution more than low-frequency components.

The program

$$\min_x \|y - Ax\|_2^2 + \delta \|Dx\|_2^2$$

has closed-form solution

$$\hat{x} = (A^TA + \delta D^TD)^{-1}A^Ty$$

when $A^TA + \delta D^TD$ is invertible.

Least-Squares with Linear Constraints

Another way to incorporate a model for x is to force the solution to lie in a particular subspace of $\mathbb{R}^N$.
If this d-dimensional subspace is spanned by the columns of the $N \times d$ matrix G, we solve

$$\min_{x \in \mathbb{R}^N} \|y - Ax\|_2^2 \quad \text{subject to } x = G\alpha \text{ for some } \alpha \in \mathbb{R}^d$$

$$= \min_{\alpha \in \mathbb{R}^d} \|y - AG\alpha\|_2^2$$

If AG has full column rank, the the solution is

$$\hat{\alpha} = (G^TA^TAG)^{-1}G^TA^Ty$$

and the corresponding estimate of x is

$$\hat{x} = G\hat{\alpha} = G(G^T A^T A G)^{-1} G^T A^T y.$$

(If AG is not full column-rank, then we can take $\hat{\alpha} = (AG)^+ y$.)

The basic idea is that while the $N \times N$ matrix $A^T A$ may not be well conditioned, the $d \times d$ matrix $G^T A^T A G$ may be — of course, establishing this requires a careful study of the relationship between G & A.

The truncated SVD is actually an example of constrained least-squares. There we implicitly took

$$G = \left[v_1 \middle| v_2 \middle| \cdots \middle| v_{p'} \right]$$

where the $\{v_i\}$ are the "right singular vectors" of A.

Example: circular deconvolution

Let $h \in \mathbb{R}^{1024}$ be defined by

$$h_n = \begin{cases} 1 & 0 \le n \le 31 \\ 0 & 32 \le n \le 1023 \end{cases},$$

and let H be the corresponding matrix generated by H:

$$H = \begin{bmatrix} h_0 & h_{N-1} & h_{N-2} & \cdots & h_1 \\ h_1 & h_0 & h_{N-1} & \cdots & h_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ h_{N-1} & h_{N-2} & h_{N-3} & \cdots & h_0 \end{bmatrix}.$$

(Here $N = 1024$.) As we saw in the previous homework, we can write

$$H = FDF^H = \sum_{k=0}^{N-1} d_k f_k f_k^H$$

where the matrix F has columns

$$f_k = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 \\ e^{j2\pi k/N} \\ e^{j2\pi 2k/N} \\ \vdots \\ e^{j2\pi(N-1)k/N} \end{bmatrix}, \quad k = 0, \ldots, N-1$$

and the d_k are the DFT coefficients of h. This is not quite an SVD of H (since H is real-values, and both F and D are complex-valued), but it has the same essential properties — just be careful below about conjugates and the difference between z^2 and $|z|^2$.

1. In MATLAB, construct H. If you like, you can use the diagonalization formula above to do this; it is easy to construct F using `F = 1/sqrt(N)*fft(eye(N));`. Turn in your code and a plot of `imagesc(H)`.

2. Suppose we observe

$$y = Hx + e,$$

and attempt to recover x using Tikhonov regularization

$$\min_x \|y - Hx\|_2^2 + \delta\|x\|_2^2.$$

Show that we can write $\hat{x} = Gy$, where G is also a circulant matrix. What are the Fourier coefficients for the $g \in \mathbb{R}^N$ whose circular shifts generate G?

3. The file `hw7prob5.mat` contains a vector x of length 1024 and a noisy observation through H, $y = Hx+e$, where $e_n \sim$ Normal$(0, 1)$. Estimate x using Tikhonov regulatization for $\delta = 1e-4, 1e-2, 1$ and 5. For each value of δ, plot x and $\hat{x}$ overlayed on one another. Comment on how the estimate changes as δ gets larger. Turn in your plots and your code.