Class Notes, Math 554, Autumn 2012

Lecture XXII: Least Squares and the Moore-Penrose inverse

We have seen last time how to solve the Least Squares Approximation Problem, of finding the minimum value of $||b - Ax||_2^2$ (given A and b), with the help of the QR factorization. The obtained minimizer x is unique, if A is full-rank ($\text{rank}(A) = n$), otherwise–trivially–and $x + y,\ y \in \mathcal{N}(A)$ will also be a minimizer.

We address now the issue of finding a specific minimizer–one that, in addition, also has minimal norm. We start by noting a fact that most undergraduates learn in linear algebra classes: that the minimizers solves the *normal equations*.

In the following, $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, and V, W are finite-dimensional spaces endowed with inner products; we can think of $A \in \mathcal{L}(V, W)$ and b as a linear transformation, and, respectively, a point in W.

Proposition 1. *(Normal equations) If $\tilde{x}$ is a minimizer, $A^*A\tilde{x} = A^*b$.*

Proof. Let us first note that $(\mathcal{R}(A))^\perp = \mathcal{N}(A^*)$. Indeed, examine $\langle y, z\rangle$ for $y \in \mathcal{N}(A^*)$ and $Z \in \mathcal{R}(A)$. For z, there must be an x such that $z = Ax$; then $\langle y, Ax\rangle = \langle A^*y, x\rangle = 0$.

Apply now the Projection Theorem with $S = \mathcal{R}(A)$ and $z = b$; the place of x in now taken by some $A\tilde{x}$ for some $\tilde{x} \in V$. This x realizes the minimum of $||z - x||_2^2$ for $x \in S$ iff $b - A\tilde{x} \in (\mathcal{R}(A))^\perp = \mathcal{N}(A^*)$. But then $A^*(A\tilde{x} - b) = 0$, so $\tilde{x}$ satisfies the nirmal equations. □

Remark 1. *The normal equations are never used in computation! As we will see, there are ways to define a condition number for matrices not necessarily invertible, or even square–but the condition number will remain multiplicative, and therefore the condition number of A^*A is potentially as high as the square of the condition number of A. Therefore, accuracy is potentially destroyed when one uses the normal equations–and why would one want to do this, when one has the QR factorization?*

As we have seen $\tilde{x}$ may not be unique. What *will* be unique will be the value $y = A\tilde{x}$; amongst all $\tilde{x}$, we will be looking for the one of minimal norm, $x^\dagger$ or "x dagger":

$$x^\dagger = \{x \in \mathbb{C}^n \ : \ A^*Ax = A^*b \ \&\ ||x||_2 \ \textit{minimal}\} = \{x \in \mathbb{C}^n \ : \ AX = y \ \&\ ||x||_2 \ \textit{minimal}\} \ .$$

To see that $x^\dagger$ is unique, consider the fact that the set $\mathcal{A} := \{x \in \mathbb{C}^n \ : \ AX = y\}$ is an *affine hyperplane* (recall notion from earlier in the quarter), that is, closed and convex. It follows then that $\mathcal{A}$ contains a unique point closest to the origin. Also, $\mathcal{A}$ is an affine translate of $\mathcal{N}(A)$, and so the closest point to 0 in $\mathcal{A}$ will necessarily be in $(\mathcal{N}(A))^\perp$.

Thus, the map $A^\dagger$ which takes b into $x^\dagger$ is a well-defined map.

Remark 2. *$A^\dagger$ is called the Moore-Penrose pseudo-inverse. Note that if A is invertible, $A^\dagger = A^{-1}$ (hence the name).*

The above considerents show that one can actually write

$$A^\dagger = \left(A\big|_{(\mathcal{N}(A))^\perp}\right)^{-1} \circ P \ ,$$

where $A\big|_{(\mathcal{N}(A))^\perp}$ is the restriction of A to $(\mathcal{N}(A))^\perp$ (which maps into $\mathcal{R}(A)$), which is trivially an isomorphism, and thus invertible, and P is the orthogonal projection onto the range $\mathcal{R}(A)$.

This makes $A^\dagger$ a linear map, which–incidentally–has a simple form in terms of the SVD of A, given in the theorem below.

Theorem 1. *If $A = U\Sigma V^*$ is a (full) SVD for A, such that the singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \sigma_{r+2} = \ldots = \sigma_n$ are thusly ordered on the diagonal of the $m \times n$ matrix Σ, then $A^\dagger = V^*\Sigma^\dagger U$ is an SVD for $A^\dagger$, where the singular values $1/\sigma_r \geq 1/\sigma_{r-1} \geq \ldots \geq 1/\sigma_1 > \sigma_{r+1} = \sigma_{r+2} = \ldots = \sigma_n = 0$ are thusly ordered on $\Sigma^\dagger$.*

Proof. The notation $\Sigma^\dagger$ is apt, since $\Sigma^\dagger$ is the Moore-Penrose pseudo-inverse for Σ.

To prove the theorem, we give the following equivalent characterization of the SVD of A.

For the matrix A of rank r, $A = U\Sigma V^*$ is a/the SVD of A if the following three facts are all true:

- $\{u_1, \ldots, u_m\}$, the columns of U, are an orthonormal basis for $\mathbb{C}^m$, with $\{u_1, \ldots, u_r\}$ being an orthonormal basis for $\mathcal{R}(A)$;
- $\{v_1, \ldots, v_n\}$, the columns of V, are an orthonormal basis for $\mathbb{C}^n$, with $\{v_{r+1}, \ldots, v_n\}$ being an orthonormal basis for $\mathcal{N}(A)$;
- $Av_i = \sigma_i u_i$, $1 \leq i \leq r$.

It will suffice to show that the above characterization holds true for $A^\dagger$.

Note that $A^\dagger = \left(A\big|_{(\mathcal{N}(A))^\perp}\right)^{-1} \circ P$ implies that $\{u_{r+1}, \ldots, u_n\}$, being a basis for $(\mathcal{R}(A))^\perp$ (which is precisely the set mapped to 0 by $A^\perp$), are a basis for $\mathcal{N}(A^\dagger)$. Also, that $\{v_1, \ldots, v_r\}$, being a basis for $(\mathcal{N}(A))^\perp$ (which is precisely the range of $A^\dagger$), are a basis for $\mathcal{R}(A^\dagger$. Both conditions follow from the fact that P is onto, and the first part of the composition is an isomorphism.

Finally, the last condition is trivially true, since the $x^\dagger$ corresponding to u_i must be of the form $\frac{1}{\sigma_i}v_i + y$, where $y \in \mathcal{N}(A)$; since $v_i \in (\mathcal{N}(A))^\perp$, $\frac{1}{\sigma_i}v_i + y$ is minimal when $y = 0$.

Therefore, conditions 1-3 are fulfilled for $A^\dagger$, and the theorem is proved. $\square$