Kyle Kloster
October 19, 2014

Intro to Least Squares Problems (LSQ)

Suppose we want to find a solution to a system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$, (with $\boldsymbol{A}$ $m \times n$) that doesn't have an exact solution – maybe $\boldsymbol{A}$ isn't full rank, or $m > n$, or for whatever reason $\mathbf{b}$ isn't in the column-space of $\boldsymbol{A}$, i.e. $\mathbf{b} \notin \text{col}(\boldsymbol{A})$.

Then an alternative kind of "solution" is a vector $\mathbf{x}$ that minimizes the 2-norm of the residual error, i.e. $\text{argmin}_{\mathbf{x}} \|\mathbf{b} - \boldsymbol{A}\mathbf{x}\|_2$. Here is a summary of approaches to finding such a solution. There are different cases, depending on the dimensions and rank of $\boldsymbol{A}$.

Full rank, $m = n$

Then $\boldsymbol{A}$ is invertible, and so $\mathbf{x} = \boldsymbol{A}^{-1}\mathbf{b}$ gives a unique solution: $\boldsymbol{A}^{-1}\mathbf{b}$ is unique and $\|\mathbf{b} - \boldsymbol{A}\mathbf{x}\|_2 = 0$ is minimal since norms must be nonnegative.

Full rank, $m > n$

We explore a few approaches to this case. To motivate learning more about QR decompositions, we first explored using a QR factorization to solve the LSQ problem in this case.

Using QR decomposition

Let $\boldsymbol{A} = \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{R}} = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{U} \end{bmatrix} \begin{bmatrix} \boldsymbol{R} \\ 0 \end{bmatrix} = \boldsymbol{Q}\boldsymbol{R}$ for $\boldsymbol{Q}$ an $m \times n$ matrix of orthonormal columns and $\boldsymbol{R}$ an $n \times n$ upper-triangular matrix. Since $\boldsymbol{A}$ is full rank, $\boldsymbol{R}$ is invertible. Now let's use the orthogonal invariance of the 2-norm!

$$\begin{aligned} \|\mathbf{b} - \boldsymbol{A}\mathbf{x}\|_2 &= \|\tilde{\boldsymbol{Q}}^T \left(\mathbf{b} - \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{R}}\mathbf{x}\right)\|_2 \quad \text{orthogonal invariance of 2-norm, plugging in QR} \\ &= \left\| \begin{bmatrix} \boldsymbol{Q}^T\mathbf{b} \\ \boldsymbol{U}^T\mathbf{b} \end{bmatrix} - \begin{bmatrix} \boldsymbol{R}\mathbf{x} \\ 0 \end{bmatrix} \right\|_2 \\ &= \left\| \begin{bmatrix} \boldsymbol{Q}^T\mathbf{b} - \boldsymbol{R}\mathbf{x} \\ \boldsymbol{U}^T\mathbf{b} \end{bmatrix} \right\|_2 \end{aligned}$$

which leads to choosing $\mathbf{x} = \boldsymbol{R}^{-1}$ so cancel out the upper partition in the expression, yielding $\|\boldsymbol{U}^T\mathbf{b}\|_2$ as the minimized error.

Uniqueness of solution In class we touched on the question "is QR unique, or could we improve this error by choosing a different QR factorization that would yield smaller $\|\boldsymbol{U}^T\mathbf{b}\|_2$?." This is where I made the mistake of thinking $\boldsymbol{U}^T$ was a tall narrow set of orthogonal columns, and I *incorrectly* stated $\|\boldsymbol{U}^T\mathbf{b}\|_2 = \|\mathbf{b}\|_2$ via the orthogonal invariance of the 2-norm.

Despite my mistake, the error *is* uniquely minimized: looking at the uniqueness of the QR decomposition (described in the QR uniqueness Lecture Appendix) we have that any other QR decomposition would yield a factor $\boldsymbol{U}_2$ related to our original $\boldsymbol{U}$ by some square orthogonal matrix $\boldsymbol{T}$ (which depends on $\boldsymbol{U}_2$) as follows: $\boldsymbol{U}_2 = \boldsymbol{U}\boldsymbol{T}$. This allows us to write, for any choice of QR decomposition we might make, $\|\boldsymbol{U}_2^T\mathbf{b}\|_2 = \|(\boldsymbol{U}\boldsymbol{T})^T\mathbf{b}\|_2 = \|\boldsymbol{T}^T\boldsymbol{U}^T\mathbf{b}\|_2 = \|\boldsymbol{U}^T\mathbf{b}\|_2$ using

the orthogonal invariance of the 2-norm – which this time we *can* use since $\boldsymbol{T}$ is *square* orthogonal.

This proves that the solution $\mathbf{x} = \boldsymbol{R}^{-1}\boldsymbol{Q}^T\mathbf{b}$ is the unique 2-norm minimizer.

Using the Normal Equations

We want to find $\mathbf{x}$ that minimizes $\|\mathbf{b} - \boldsymbol{A}\mathbf{x}\|_2$; the same $\mathbf{x}$ minimizes

$$\begin{aligned} f(\mathbf{x}) &= \|\mathbf{b} - \boldsymbol{A}\mathbf{x}\|_2^2 \\ &= (\mathbf{b} - \boldsymbol{A}\mathbf{x})^T(\mathbf{b} - \boldsymbol{A}\mathbf{x}) \\ &= \mathbf{b}^T\mathbf{b} - \mathbf{x}^T\boldsymbol{A}^T\mathbf{b} - \mathbf{b}^T\boldsymbol{A}\mathbf{x} + \mathbf{x}^T\boldsymbol{A}^T\boldsymbol{A}\mathbf{x} \\ &= \mathbf{b}^T\mathbf{b} - 2\mathbf{b}^T\boldsymbol{A}\mathbf{x} + \mathbf{x}^T\boldsymbol{A}^T\boldsymbol{A}\mathbf{x}. \end{aligned}$$

The $\mathbf{x}$ that minimizes $f(\mathbf{x})$ is the same $\mathbf{x}$ that minimizes $h(\mathbf{x}) = -2\mathbf{b}^T\boldsymbol{A}\mathbf{x} + \mathbf{x}^T\boldsymbol{A}^T\boldsymbol{A}\mathbf{x}$ because $\mathbf{b}^T\mathbf{b}$ is a constant independent of $\mathbf{x}$. The minimum of the convex function $h(\mathbf{x})$ is where its gradient, $\nabla(\mathbf{x})$, is 0. we have

$$\frac{\partial}{\partial \mathbf{x}}\left(-2\mathbf{b}^T\boldsymbol{A}\mathbf{x}\right) = -2\boldsymbol{A}^T\mathbf{b} \quad \text{and} \quad \frac{\partial}{\partial \mathbf{x}}\left(\mathbf{x}^T\boldsymbol{A}^T\boldsymbol{A}\mathbf{x}\right) = 2\boldsymbol{A}^T\boldsymbol{A}\mathbf{x}$$

so $\nabla(\mathbf{x}) = -2\boldsymbol{A}^T\mathbf{b} + 2\boldsymbol{A}^T\boldsymbol{A}\mathbf{x}$. Setting $\nabla(\mathbf{x}) = 0$ yields the **normal equations**:

$$\boldsymbol{A}^T\boldsymbol{A}\mathbf{x} = \boldsymbol{A}^T\mathbf{b}. \tag{1}$$

Since we are in the case that $\boldsymbol{A}$ is full rank and $m \times n$ with $m > n$, we have that $\boldsymbol{A}^T\boldsymbol{A}$ is invertible. Hence, $\mathbf{x} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\mathbf{b}$ is the unique minimizer of $h(\mathbf{x})$, thus the unique minimizer of $f(\mathbf{x})$, and hence of $\|\mathbf{b} - \boldsymbol{A}\mathbf{x}\|_2$.

(Side note: to see more on *why* a zero of the gradient of $h(\mathbf{x})$ is a minimizer of the function $h(\mathbf{x})$, take CS 520 on optimization. The intuition is the same reason that minima (and maxima) of a function of *one* variable, $f(x)$, are where the derivative is 0, i.e. $f'(x) = 0$.)

Using the SVD

Let $\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$ be the SVD for $\boldsymbol{A}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are square, orthogonal. Then using the orthogonal invariance of the 2-norm we have

$$\begin{aligned} f(\mathbf{x}) &= \|\mathbf{b} - \boldsymbol{A}\mathbf{x}\|_2^2 \\ &= \|\boldsymbol{U}^T(\mathbf{b} - \boldsymbol{U}\Sigma\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{V}^T)\mathbf{x}\|_2^2 \\ &= \|(\boldsymbol{U}^T\mathbf{b}) - \Sigma(\boldsymbol{V}^T\mathbf{x})\|_2^2. \end{aligned}$$

Since $\boldsymbol{A}$ is $m \times n$ with $m > n$, we know $\Sigma = \begin{bmatrix}\boldsymbol{D} \\ 0\end{bmatrix}$ where $\boldsymbol{D}$ is $n \times n$ diagonal, $\boldsymbol{D} = \text{diag}(\sigma_1, \cdots, \sigma_n)$. From this, we can write

$$\begin{aligned} f(\mathbf{x}) &= \|(\boldsymbol{U}^T\mathbf{b}) - \Sigma(\boldsymbol{V}^T\mathbf{x})\|_2^2 \\ &= \left\| \begin{bmatrix}\mathbf{c}_1 \\ \mathbf{c}_2\end{bmatrix} - \begin{bmatrix}\boldsymbol{D} \\ 0\end{bmatrix}(\boldsymbol{V}^T\mathbf{x}) \right\|_2^2 \end{aligned}$$

and so minimizing $f(\mathbf{x})$ is achieved by $\mathbf{x}$ satisfying $\boldsymbol{V}^T\mathbf{x} = \boldsymbol{D}^{-1}\mathbf{c}_1$ where $\begin{bmatrix}\mathbf{c}_1 \\ \mathbf{c}_2\end{bmatrix} = \boldsymbol{U}^T\mathbf{b}$. Multiplying by $\boldsymbol{V}$ yields $\mathbf{x} = \boldsymbol{V}\boldsymbol{D}^{-1}\mathbf{c}_1 = \sum_{i=1}^n \frac{1}{\sigma_i}\mathbf{v}_i(\mathbf{u}_i^T\mathbf{b})$.

Since $\boldsymbol{A}$ is full rank, we have $\sigma_i \neq 0$ for $i = 1 : n$, and so this all makes sense. Further note that, since the SVD is unique (up to factors of ± 1 on the vectors $\mathbf{u}_i$ and $\mathbf{v}_i$), this expression of $\mathbf{x}$ is unique.

In fact, using the expression below for $\boldsymbol{A}^+$ we have that

$$\begin{aligned}\mathbf{x} &= \boldsymbol{V}\boldsymbol{D}^{-1}\mathbf{c}_1 \\ &= \sum_{i=1}^{n} \tfrac{1}{\sigma_i}\mathbf{v}_i(\mathbf{u}_i^T\mathbf{b}) \\ &= \boldsymbol{A}^+\mathbf{b} \\ &= \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\mathbf{b}\end{aligned}$$

Generalized Inverses and the Rank-deficient LSQ problem

In the SVD approach we took $\Sigma = \begin{bmatrix}\boldsymbol{D} \\ 0\end{bmatrix}$ where $\boldsymbol{D} = \text{diag}(\sigma_1, \cdots, \sigma_n)$ is an $n \times n$ diagonal matrix, and since we were in the case that $\boldsymbol{A}$ had full rank, $\sigma_n \neq 0$, so $\boldsymbol{D}$ was invertible. In class we discussed using this to define a "pseudo-inverse" for $\boldsymbol{A}$ as follows:

$$\boldsymbol{A}^+ := \boldsymbol{V}\Sigma^+\boldsymbol{U}^T = \boldsymbol{V}\begin{bmatrix}\boldsymbol{D}^{-1} & 0\end{bmatrix}\boldsymbol{U}^T$$

where $\boldsymbol{D}^{-1} = \text{diag}(\frac{1}{\sigma_1}, \cdots, \frac{1}{\sigma_n})$ exists because we had assumed $\boldsymbol{A}$ was full rank.

Now we abandon the assumption that $\boldsymbol{A}$ is full rank. To solve this case we expand our definition of $\boldsymbol{A}^+$ to the case that $\sigma_i = 0$ for $i > r$, where $r = \text{rank}(\boldsymbol{A}) < n$. In this situation, $\boldsymbol{D} = \text{diag}(\sigma_1, \cdots, \sigma_r, 0, \cdots, 0)$, and we define $\boldsymbol{D}^+ = \text{diag}(\frac{1}{\sigma_1}, \cdots, \frac{1}{\sigma_r}, 0, \cdots, 0)$. Then we can define the **generalized inverse** of $\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$ to be

$$\boldsymbol{A}^+ := \boldsymbol{V}\Sigma^+\boldsymbol{U}^T = \boldsymbol{V}\begin{bmatrix}\frac{1}{\sigma_1} & & & 0 & \cdots & 0 \\ & \ddots & & 0 & \cdots & 0 \\ & & \frac{1}{\sigma_r} & 0 & \cdots & 0 \\ & & & 0 & \cdots & 0\end{bmatrix}\boldsymbol{U}^T = \sum_{j=1}^{r}\tfrac{1}{\sigma_j}\mathbf{v}_j\mathbf{u}_j^T.$$

This is one approach to the generalized inverse. Another approach is to define the generalized inverse to be the unique matrix satisfying certain properties:

Moore Penrose generalized inverse of an arbitrary $m \times n$ matrix $\boldsymbol{A}$ is the unique $n \times m$ matrix $\boldsymbol{X}$ satisfying

1. $\boldsymbol{A}\boldsymbol{X}\boldsymbol{A} = \boldsymbol{A}$
2. $\boldsymbol{X}\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}$
3. $(\boldsymbol{A}\boldsymbol{X})^T = \boldsymbol{A}\boldsymbol{X}$
4. $(\boldsymbol{X}\boldsymbol{A})^T = \boldsymbol{X}\boldsymbol{A}$

Remarks If $\boldsymbol{A}^{-1}$ exists, then $\boldsymbol{X} = \boldsymbol{A}^{-1}$. The matrix $\boldsymbol{A}^+$ as defined above satisfies these properties, and so is equivalent to the Moore-Penrose generalized inverse. If $\boldsymbol{A}$ is full rank and $m \geq n$ then we have $\boldsymbol{A}^+ = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T$. Finally, $\boldsymbol{A}\boldsymbol{A}^+$ and $\boldsymbol{A}^+\boldsymbol{A}$ are the orthogonal projections of $\boldsymbol{A}$ onto range($\boldsymbol{A}$) and range($\boldsymbol{A}^T$), respectively.

Rank deficient, $m \geq n$

The rank deficient case follows the use of the SVD in the full rank case: with the generalized inverse defined in for even rank-deficient $\boldsymbol{A}$, the minimum norm solution to the LSQ problem is $\mathbf{x} = \boldsymbol{A}^{+}\mathbf{b}$.

Numerical Rank Since we solve these problems on computers, the exact rank of a matrix is not always the same as the rank of the floating-point representation of a matrix. For example, A matrix $\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^{T}$ with σ_j very small might have $\sigma_j = 0$ in floating-point arithmetic. To make this notion precise, we define the **numerical rank** of a matrix $\boldsymbol{A}$ by $r(\varepsilon) = \#$ of singular values of $\boldsymbol{A}$ that are $\geq \varepsilon$.