These notes correspond to Section 5.3 in the text.

The Full-rank Linear Least Squares Problem

Minimizing the Residual

Given an $m \times n$ matrix $A$, with $m \geq n$, and an $m$-vector $b$, we consider the overdetermined system of equations $Ax = b$, in the case where $A$ has full column rank. If $b$ is in the range of $A$, then there exists a unique solution $x^*$. For example, there exists a unique solution in the case of

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix},$$

but not if $b = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$. In such cases, when $b$ is not in the range of $A$, then we seek to minimize $\|b - Ax\|_p$ for some $p$. Recall that the vector $r = b - Ax$ is known as the residual vector.

Different norms give different solutions. If $p = 1$ or $p = \infty$, then the function we seek to minimize, $f(x) = \|b - Ax\|_p$ is not differentiable, so we cannot use standard minimization techniques. However, if $p = 2$, $f(x)$ is differentiable, and thus the problem is more tractable. We now consider two methods.

The first approach is to take advantage of the fact that the 2-norm is invariant under orthogonal transformations, and seek an orthogonal matrix $Q$ such that the transformed problem

$$\min \|b - Ax\|_2 = \min \|Q^T(b - Ax)\|_2$$

is “easy” to solve. Let

$$A = QR = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1,$$

where $Q_1$ is $m \times n$ and $R_1$ is $n \times n$. Then, because $Q$ is orthogonal, $Q^T A = R$ and

$$\min \|b - Ax\|_2 = \min \|Q^T(b - Ax)\|_2 = \min \|Q^T b - (Q^T A)x\|_2 = \min \|Q^T b - \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x\|_2$$
If we partition 

$$Q^T b = \begin{bmatrix} c \\ d \end{bmatrix},$$

where $c$ is an $n$-vector, then

$$\min \| b - Ax \|_2^2 = \min \left\| \begin{bmatrix} c \\ d \end{bmatrix} - \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x \right\|^2_2 = \min \| c - R_1 x \|_2^2 + \| d \|_2^2.$$ 

Therefore, the minimum is achieved by the vector $x$ such that $R_1 x = c$ and therefore

$$\min_x \| b - Ax \|_2 = \| d \|_2 = \rho_{LS}.$$ 

It makes sense to seek a factorization of the form $A = QR$ where $Q$ is orthogonal, and $R$ is upper-triangular, so that $R_1 x = c$ is easily solved. This is called the QR factorization of $A$.

The second method is to define $(x) = \frac{1}{2}\| b - Ax \|_2^2$, which is a differentiable function of $x$. We can minimize $\phi(x)$ by noting that $\nabla \phi(x) = A^T (b - Ax)$, which means that $\nabla \phi(x) = 0$ if and only if $A^T A x = A^T b$. This system of equations is called the normal equations, and were used by Gauss to solve the least squares problem. If $m \gg n$ then $A^T A$ is $n \times n$, which is a much smaller system to solve than $Ax = b$, and if $\kappa(A^T A)$ is not too large, we can use the Cholesky factorization to solve for $x$, as $A^T A$ is symmetric positive definite.

Which is the better method? This is not a simple question to answer. The normal equations produce an $x^*$ whose relative error depends on $\kappa(A)^2$, whereas the QR factorization produces an $x^*$ whose relative error depends on $u(\kappa_2(A) + \rho_{LS}k_2(A)^2)$. The normal equations involve much less arithmetic when $m \gg n$ and they require less storage, but the QR factorization is often applicable if the normal equations break down.

Using the Normal Equations

Instead of using the Cholesky factorization, we can solve the linear least squares problem using the normal equations

$$A^T A x = A^T b$$

as follows: first, we solve the above system to obtain an approximate solution $\hat{x}$, and compute the residual vector $r = b - A\hat{x}$. Now, because $A^T r = A^T b - A^T A \hat{x} = 0$, we obtain the system

$$r + A \hat{x} = b$$

$$A^T r = 0$$

or, in matrix form,

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ \hat{x} \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$ 

This is a large system, but it preserves the sparsity of $A$. It can be used in connection with iterative refinement, but unfortunately this procedure does not work well because it is very sensitive to the residual.
The Pseudo-Inverse

The singular value decomposition is very useful in studying the linear least squares problem. Suppose that we are given an \(m\)-vector \(b\) and an \(m \times n\) matrix \(A\), and we wish to find \(x\) such that

\[
\|b - Ax\|_2 = \text{minimum}.
\]

From the SVD of \(A\), we can simplify this minimization problem as follows:

\[
\begin{align*}
\|b - Ax\|_2^2 &= \|b - U\Sigma V^T x\|_2^2 \\
&= \|U^T b - \Sigma V^T x\|_2^2 \\
&= \|c - \Sigma y\|_2^2 \\
&= (c_1 - \sigma_1 y_1)^2 + \cdots + (c_r - \sigma_r y_r)^2 + \\
&\quad c_{r+1}^2 + \cdots + c_m^2
\end{align*}
\]

where \(c = U^T b\) and \(y = V^T x\). We see that in order to minimize \(\|b - Ax\|_2\), we must set \(y_i = c_i / \sigma_i\) for \(i = 1, \ldots, r\), but the unknowns \(y_i\), for \(i = r + 1, \ldots, n\), can have any value, since they do not influence \(\|c - \Sigma y\|_2\). Therefore, if \(A\) does not have full rank, there are infinitely many solutions to the least squares problem. However, we can easily obtain the unique solution of minimum 2-norm by setting \(y_{r+1} = \cdots = y_n = 0\).

In summary, the solution of minimum length to the linear least squares problem is

\[
x = V y \\
= V \Sigma^+ c \\
= V \Sigma^+ U^T b \\
= A^+ b
\]

where \(\Sigma^+\) is a diagonal matrix with entries

\[
\Sigma^+ = \begin{bmatrix}
\sigma_1^{-1} & & \\
& \ddots & \\
& & \sigma_r^{-1}
\end{bmatrix}
\]

and \(A^+ = V \Sigma^+ U^T\). The matrix \(A^+\) is called the pseudo-inverse of \(A\). In the case where \(A\) is square and has full rank, the pseudo-inverse is equal to \(A^{-1}\). Note that \(A^+\) is independent of \(b\). It also has the properties

1. \(AA^+A = A\)
2. $A^+ AA^+ = A^+$
3. $(A^+ A)^T = A^+ A$
4. $(AA^+)^T = AA^+$

The solution $\mathbf{x}$ of the least-squares problem minimizes $\|\mathbf{b} - A \mathbf{x}\|_2$, and therefore is the vector that solves the system $A \mathbf{x} = \mathbf{b}$ as closely as possible. However, we can use the SVD to show that $\mathbf{x}$ is the exact solution to a related system of equations. We write $\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2$, where $\mathbf{b}_1 = AA^+ \mathbf{b}$, $\mathbf{b}_2 = (I - AA^+) \mathbf{b}$.

The matrix $AA^+$ has the form

$$AA^+ = U \Sigma V^T V \Sigma + U^T = U \Sigma \Sigma + U^T = U \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} U^T,$$

where $I_r$ is the $r \times r$ identity matrix. It follows that $\mathbf{b}_1$ is a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_r$, the columns of $U$ that form an orthogonal basis for the range of $A$.

From $\mathbf{x} = A^+ \mathbf{b}$ we obtain

$$A \mathbf{x} = AA^+ \mathbf{b} = P \mathbf{b} = \mathbf{b}_1,$$

where $P = AA^+$. Therefore, the solution to the least squares problem, is also the exact solution to the system $A \mathbf{x} = P \mathbf{b}$. It can be shown that the matrix $P$ has the properties

1. $P = P^T$
2. $P^2 = P$

In other words, the matrix $P$ is a projection. In particular, it is a projection onto the space spanned by the columns of $A$, i.e. the range of $A$. That is, $P = U_r U_r^T$, where $U_r$ is the matrix consisting of the first $r$ columns of $U$.

The residual vector $\mathbf{r} = \mathbf{b} - A \mathbf{x}$ can be expressed conveniently using this projection. We have

$$\mathbf{r} = \mathbf{b} - A \mathbf{x} = \mathbf{b} - AA^+ \mathbf{b} = \mathbf{b} - P \mathbf{b} = (I - P) \mathbf{b} = P^\perp \mathbf{b}.$$

That is, the residual is the projection of $\mathbf{b}$ onto the orthogonal complement of the range of $A$, which is the null space of $A^T$. Furthermore, from the SVD, the 2-norm of the residual satisfies

$$\rho_{LS}^2 \equiv \|\mathbf{r}\|_2^2 = c_{r+1}^2 + \cdots + c_m^2,$$

where, as before, $\mathbf{c} = U^T \mathbf{b}$. 
Perturbation Theory

Suppose that we perturb the data, so that we are solving \((A + \epsilon E)x(\epsilon) = b\). Then what is \(\|x - x(\epsilon)\|_2\) or \(\|r - r(\epsilon)\|_2\)? Using the fact that \(PA = AA^+A = A\), we differentiate with respect to \(\epsilon\) and obtain

\[
P \frac{dA}{d\epsilon} + \frac{dP}{d\epsilon} A = \frac{dA}{d\epsilon}.
\]

It follows that

\[
\frac{dP}{d\epsilon} A = (I - P) \frac{dA}{d\epsilon} = P_\perp \frac{dA}{d\epsilon}.
\]

Multiplying through by \(A^+\), we obtain

\[
\frac{dP}{d\epsilon} = P_\perp \frac{dA}{d\epsilon} A^+.
\]

Because \(P\) is a projection,

\[
\frac{d(P^2)}{d\epsilon} = P \frac{dP}{d\epsilon} + \frac{dP}{d\epsilon} P = \frac{dP}{d\epsilon},
\]

so, using the symmetry of \(P\),

\[
\frac{dP}{d\epsilon} = P_\perp \frac{dA}{d\epsilon} A^+ + (A^+)^T \frac{dA^T}{d\epsilon} P_\perp.
\]

Now, using a Taylor expansion around \(\epsilon = 0\), as well as the relations \(\hat{x} = A^+ b\) and \(r = P_\perp b\), we obtain

\[
\begin{align*}
\mathbf{r}(\epsilon) &= \mathbf{r}(0) + \epsilon \frac{dP_\perp}{d\epsilon} b + O(\epsilon^2) \\
&= \mathbf{r}(0) + \epsilon \frac{d(I - P)}{d\epsilon} b + O(\epsilon^2) \\
&= \mathbf{r}(0) - \epsilon \frac{dP}{d\epsilon} b + O(\epsilon^2) \\
&= \mathbf{r}(0) - \epsilon [P_\perp E \hat{x}(0) + (A^+)^T E^T r(0)] + O(\epsilon^2).
\end{align*}
\]

Taking norms, we obtain

\[
\frac{\|\mathbf{r}(\epsilon) - \mathbf{r}(0)\|_2}{\|\hat{x}\|_2} = |\epsilon| \|E\|_2 \left(1 + \|A^+\|_2 \frac{\|\mathbf{r}(0)\|_2}{\|\hat{x}(0)\|_2}\right) + O(\epsilon^2).
\]

Note that if \(A\) is scaled so that \(\|A\|_2 = 1\), then the second term above involves the condition number \(\kappa_2(A)\). We also have

\[
\frac{\|\mathbf{x}(\epsilon) - \mathbf{x}(0)\|_2}{\|\hat{x}\|_2} = |\epsilon| \|E\|_2 \left(2\kappa_2(A) \frac{\|\mathbf{r}(0)\|_2}{\|\hat{x}(0)\|_2}\right) + O(\epsilon^2).
\]

Note that a small perturbation in the residual does not imply a small perturbation in the solution.