Least Squares Estimation

- Consider an system of m equations in n unknown, with $m > n$, of the form
$$y = Ax.$$
- Assume that the system is inconsistent: there are more equations than unknowns, and these equations are non linear combinations of one another.
- In these conditions, there is no x such that $y - Ax = 0$. However, one can write $e = y - Ax$, and find x that minimizes $\|e\|$.
- In particular, the problem
$$\min_x \|e\|_2 = \min_x \|y - Ax\|_2$$
is a least squares problem. The optimal x is the least squares estimate.

Computing the Least-Square Estimate

- The set $M := \{z \in \mathbb{R}^m : z = Ax,\ x \in \mathbb{R}^n\}$ is a subspace of $\mathbb{R}^m$, called the range of A, $\mathcal{R}(A)$, i.e., the set of all vectors that can be obtained by linear combinations of the columns of A.
- Recall the projection theorem. Now we are looking for the element of M that is "closest" to y, in terms of 2-norm. We know the solution is such that

$$e = (y - A\hat{x}) \perp \mathcal{R}(A).$$

- In particular, if a_i is the i-th column of A, it is also the case that

$$\begin{aligned}(y - A\hat{x}) \perp \mathcal{R}(A) \quad &\Leftrightarrow \quad a_i'(y - A\hat{x}) = 0, \qquad i = 1, \ldots, n \\ & \quad A'(y - A\hat{x}) = 0 \\ & \quad A'A\hat{x} = A'y\end{aligned}$$

- $A'A$ is a $n \times n$ matrix; is it invertible? It if were, then at this point it is easy to recover the least-square solution as

$$\hat{x} = (A'A)^{-1}A'y.$$

The Gram product

- Let us take a more abstract look at this problem, e.g., to address the case that the data vector y is infinite-dimensional.
- Given an array of n_A vectors $A = [a_1| \ldots |a_{n_A}]$, and an array of n_B vectors $B = [b_1| \ldots |b_{n_B}]$, both from an inner vector space V, define the Gram Product $\prec A, B \succ$ as a $n_A \times n_B$ matrix such that its (i, j) entry is $\langle a_i, b_j \rangle$.
- For the usual Euclidean inner product in an m-dimensional space,

$$\prec A, B \succ = A'B.$$

- Symmetry and linearity of the inner product imply symmetry and linearity of the Gram product.

The Least Squares Estimation Problem

- Consider again the problem of computing

$$\min_{x \in \mathbb{R}^n} \| \underbrace{y - Ax}_{e} \| = \min_{\hat{y} \in \mathcal{R}(A)} \|y - \hat{y}\|.$$

- y can be an infinite-dimensional vector—as long as n is finite.
- We assume that the columns of $A = [a_1, a_2, \ldots, a_n]$ are independent.

Lemma (Gram matrix)

The columns of a matrix A are independent $\Leftrightarrow \prec A, A \succ$ is invertible.

Proof— If the columns are dependent, then there is $\eta \neq 0$ such that $A\eta = \sum_j a_j \eta_j = 0$. But then $\sum_j \langle a_i, a_j \rangle \eta j = 0$ by the linearity of inner product.
That is, $\prec A, A \succ \eta = 0$, and hence $\prec A, A \succ$ is not invertible.
Conversely, if $\prec A, A \succ$ is not invertible, then $\prec A, A \succ \eta = 0$ for some $\eta \neq 0$. In other words $\eta' \prec A, A \succ \eta = 0$, and hence $A\eta = 0$.

- y has a unique decomposition $y = y_1 + y_2$, where $y_1 \in \mathcal{R}(A)$, and $y_2 \in \mathcal{R}^\perp(A)$.
- To find this decomposition, let $y_1 = A\alpha$, for some $\alpha \in \mathbb{R}^n$. Then, ensure that $y_2 = y - y_1 \in \mathcal{R}^\perp(A)$. For this to be true,

$$\langle a_i, y - A\alpha \rangle = 0, \quad i = 1, \ldots, n,$$

  i.e.,

$$\prec A, y - A\alpha \succ = 0.$$

- Rearranging, we get

$$\prec A, A \succ \alpha = \prec A, y \succ$$

- if the columns of A are independent,

$$\alpha = \prec A, A \succ^{-1} \prec A, y \succ$$

The Projection theorem and least squares estimation 2

- Decompose $e = e_1 + e_2$ similarly ($e_1 \in \mathcal{R}(A)$, and $e_2 \in \mathcal{R}^{\perp}(A)$).
- Note $\|e\|^2 = \|e_1\|^2 + \|e_2\|^2$.
- Rewrite $e = y - Ax$ as

$$e_1 + e_2 = y_1 + y_2 - Ax,$$

i.e.,

$$e_2 - y_2 = y_1 - e_1 - Ax.$$

- Each side must be 0, since they are on orthogonal subspaces!
- $e_2 = y_2$—can't do anything about it.
- $e_1 = y_1 - Ax = A(\alpha - x)$—minimize by choosing $x = \alpha$. In other words

$$\hat{x} = \prec A, A \succ^{-1} \prec A, y \succ .$$

Examples

- If $y, e \in \mathbb{R}^m$, and it is desired to minimize $\|e\|^2 = e'e = \sum_{i=1}^m |e_i|^2$, then

$$\hat{x} = (A'A)^{-1}A'y$$

  (If the columns of A are mutually orthogonal, $A'A$ is diagonal, and inversion is easy)

- if $y, e \in \mathbb{R}^m$, and it is desired to minimize $e'Se$, where S is a Hermitian, positive-definite matrix, then

$$\hat{x} = (A'SA)^{-1}A'Sy.$$

- Note that if S is diagonal, then $e'Se = \sum_{i=1}^m s_{ii}|e_i|^2$, i.e., we are minimizing a weighted least square criterion. A large s_{ii} penalizes the i-th component of the error more relative to the others.
- In a general stochastic setting, the weight matrix S should be related to the noise covariance, i.e.,

$$S = (E[ee'])^{-1}.$$