7. Linear least-squares

- Linear regression
- Linear least-squares problems
- Examples
- Analysis of least-squares estimate
- Computational aspects

Linear regression

- The linear regression is the simplest type of *parametric* model
- It explains a relationship between variables y and x using a linear function:

$$y = Ax$$

  where $y \in \mathbf{R}^N$, $A \in \mathbf{R}^{N \times n}$, $x \in \mathbf{R}^n$

- y contains the measurement variables and is called the *regressed variable* or *regressand*
- Each row vector a_k^T in matrix A is called *regressor*
- The matrix A is sometimes called *the design matrix*
- x is the *parameter vector*. Its element x_k is often called *regression coefficients*

Example 1: A Polynomial trend

Suppose the model is of the form

$$y(t) = a_0 + a_1 t + \ldots + a_r t^r$$

with unknown coefficients $a_0, \ldots, a_r$

This can be written in the form of linear regression as

$$\begin{bmatrix} y(t_1) \\ y(t_2) \\ \vdots \\ y(t_N) \end{bmatrix} = \begin{bmatrix} 1 & t_1 & \ldots & t_1^r \\ 1 & t_2 & \ldots & t_2^r \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_N & \ldots & t_N^r \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_r \end{bmatrix}$$

Given the measurements $y(t_i)$ for $t_1, t_2, \ldots, t_N$, we want to estimate the coefficents a_k

Example 2: Truncated weighting function

A truncated weighting function model (or FIR model) is given by

$$y(k) = \sum_{k=0}^{M-1} h(k)u(t-k)$$

The input u is known and applied to the system to measure the output y

The relationship between y and u can be fit into a linear regression as

$$\begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(k) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} u(0) & u(-1) & \ldots & u(-M+1) \\ u(1) & u(0) & \ldots & u(-M+2) \\ \vdots & \vdots & \vdots & \vdots \\ u(k) & u(k-1) & \ldots & u(k-M+1) \\ \vdots & \vdots & \vdots & \vdots \\ u(N) & u(N-1) & \ldots & u(N-M+1) \end{bmatrix} \begin{bmatrix} h(0) \\ h(1) \\ \vdots \\ h(M-1) \end{bmatrix}$$

Solving linear regressions

- The problem is to find an estimate $\hat{x}$ from the measurements y and A
- If we choose the number of measurements, N to be equal to n, then x can be solved by
$$x = A^{-1}y,$$
provided that A is invertible
- In practice, in the presence of noise and disturbance, more data should be collected in order to get a better estimate
- This leads to overdetermined linear equations where an exact solution does not usually exist
- However, it can be solved by linear least-squares formulation

Definition of Linear least-squares

Overdetermined linear equations

$$Ax = y \quad A \text{ is } m \times n \text{ with } m > n$$

for most y cannot solve for x

Linear least-squares formulation

$$\text{minimize } \|Ax - y\|_2 = \left(\sum_{i=1}^{m} (\sum_{j=1}^{n} a_{ij}x_j - y_i)^2 \right)^{1/2}$$

- $r = Ax = y$ is called *the residual error*
- x with smallest residual norm $\|r\|$ is called *the least-squares solution*
- equivalent to minimizing $\|Ax - y\|^2$

Example: Data fitting

fit a function

$$y = g(t) = x_1 g_1(t) + x_2 g_2(t) + \ldots + x_n g_n(t)$$

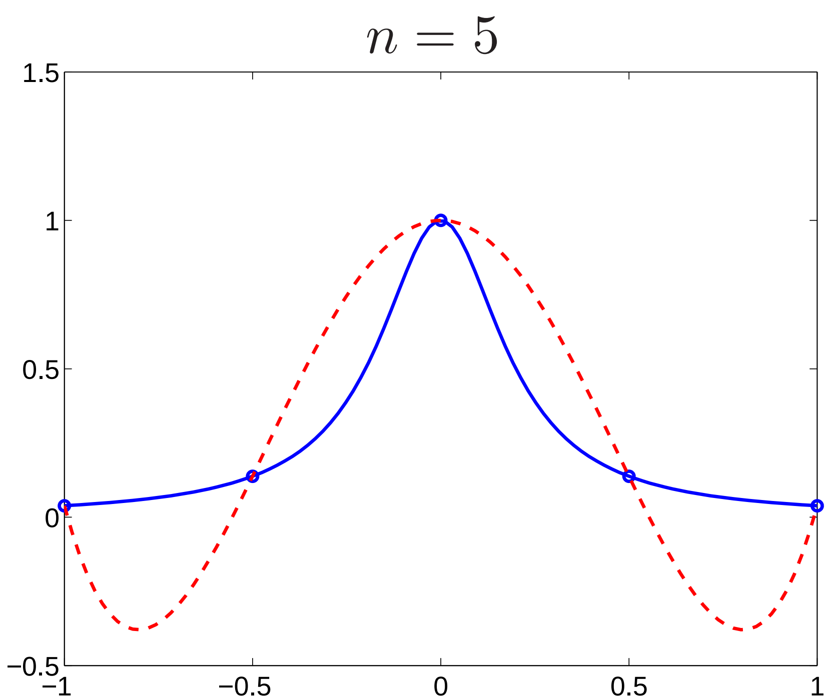
to data (t_1, y_1), (t_2, y_2), . . . , (t_m, y_m), i.e., choose the coefficients x_k so that

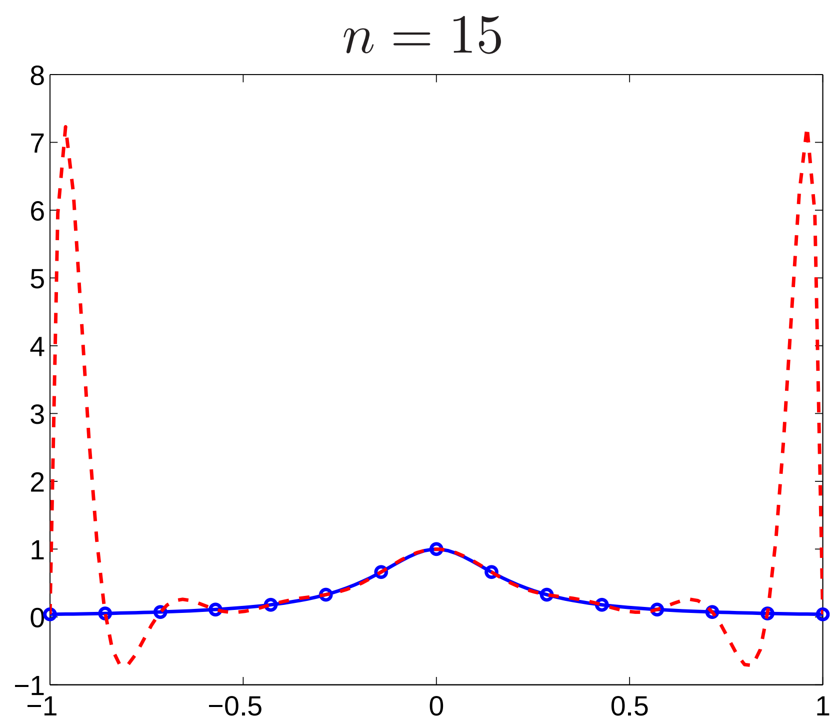
$$g(t_1) \approx y_1, \quad g(t_2) \approx y_2, \quad , g(t_m) \approx y_m$$

- $g_i(t) : \mathbf{R} \to \mathbf{R}$ are given functions (*basis functions*)
- problem variables: the coefficients $x_1, x_2, \ldots, x_n$
- usually $m \gg n$, hence no exact solution with $g(t_i) = y_i$ for all i
- applications: developing simple, approximate model of observed data

Example: fit a polynomial to $f(t) = 1/(1 + 25t^2)$ on $[-1, 1]$

- pick $m = n$ points t_i in $[-1, 1]$ and calculate $y_i = 1/(1 + 25t_i^2)$
- interpolate by solving $Ax = y$



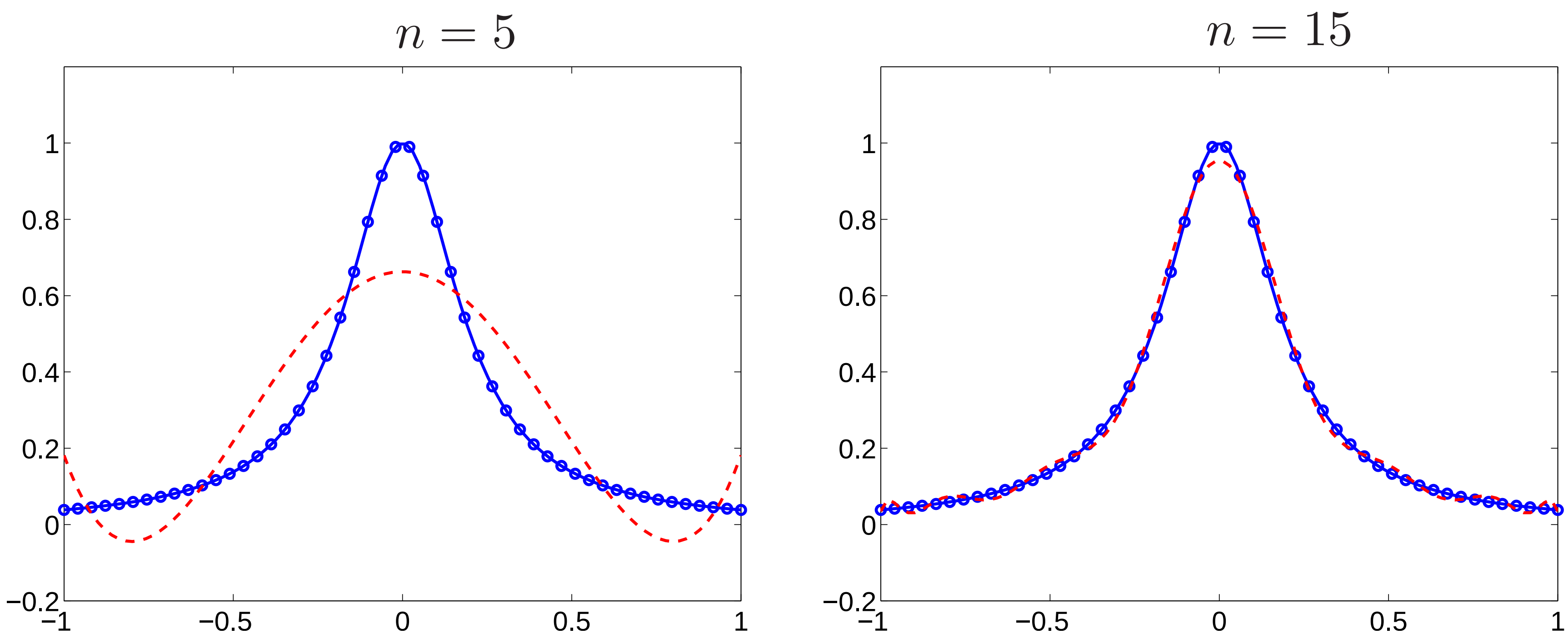


(blue solid line: f; red dashed line: polynomial g)

increase n does not improve the overall quality of the fit

Same example by approximation

- pick $m = 50$ points t_i in $[-1, 1]$
- fit polynomial by minimizing $\|Ax - y\|$



blue solid line: f; red dashed line: polynomial g)

much better fit overall

Geometric interpretation of a LS problem

$$\text{minimize} \ \ \|Ax - y\|^2$$

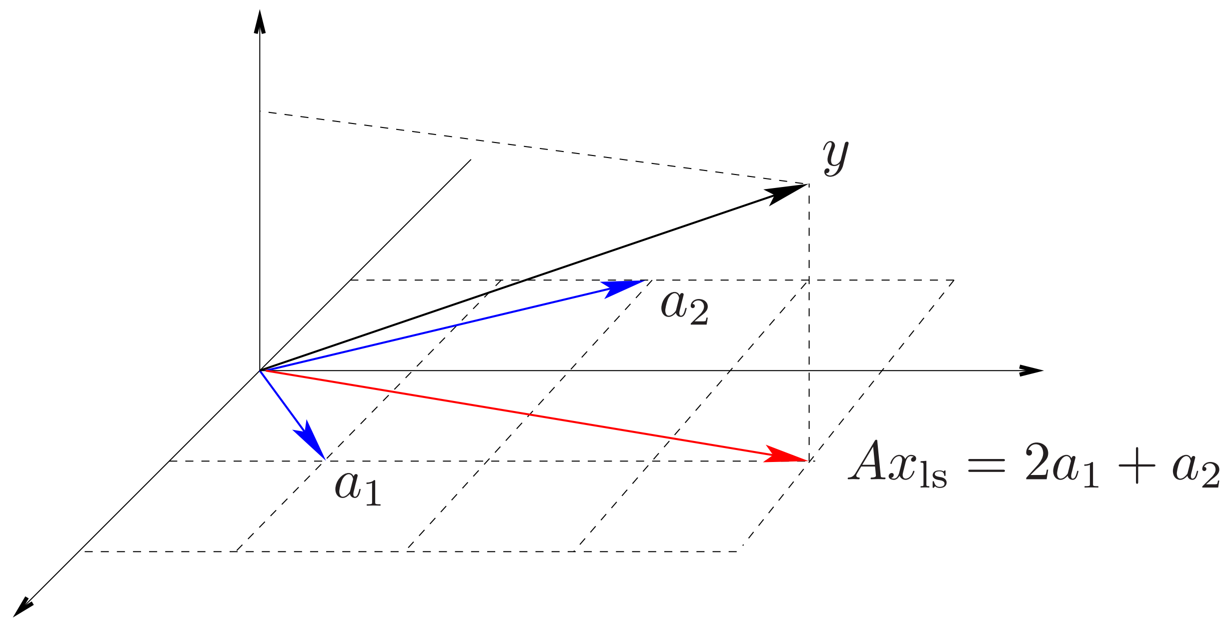
A is $m \times n$ with colums $a_1, a_2, \ldots a_m$

- $\|Ax - y\|$ is the distance of y to the vector

$$Ax = a_1x_1 + a_2x_2 + \ldots a_nx_n$$

- solution x_{ls} gives the linear combination of the columns of A closest to y
- Ax_{ls} is the **projection** of y to the range of A

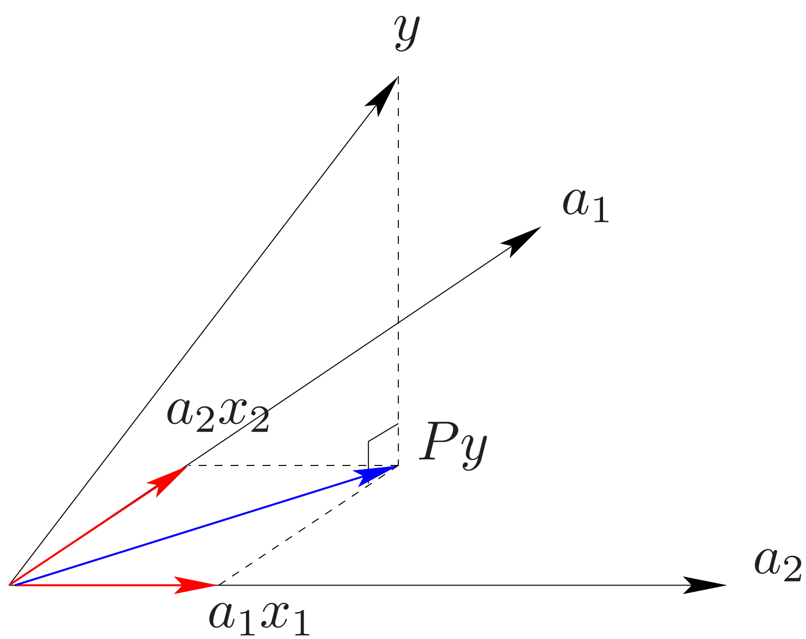
Example: $A = \begin{bmatrix} 1 & -1 \\ 1 & 2 \\ 0 & 0 \end{bmatrix}$, $y = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}$



Least-squares solution $x_{\rm ls}$

$$Ax_{\rm ls} = \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix}, \quad x_{\rm ls} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Orthogonal projection



- Py is the orthogonal projection of y onto $\mathcal{R}(A)$ spanned by $a_1, \ldots, a_n$
- The projection satisfies the **orthogonality condition**

$$\langle a_k, Py - y \rangle = 0, \quad \forall k$$

(The optimal residual must be orthogonal to any vector in $\mathcal{R}(A)$)

- Py gives the best approximation; for any $\hat{y} \in \mathcal{R}(A)$ and $\hat{y} \neq Py$

$$\|y - Py\| < \|y - \hat{y}\|$$

- From the orthogonality condition and Py is a linear combination of $\{a_k\}$

$$\langle a_k, y \rangle = \langle a_k, Py \rangle = \langle a_k, \sum_{j=1}^{n} a_j x_j \rangle \quad \forall k$$

$$\begin{bmatrix} \langle a_1, y \rangle \\ \langle a_2, y \rangle \\ \vdots \\ \langle a_n, y \rangle \end{bmatrix} = \begin{bmatrix} \langle a_1, a_1 \rangle & \langle a_1, a_2 \rangle & \dots & \langle a_1, a_n \rangle \\ \langle a_2, a_1 \rangle & \langle a_2, a_2 \rangle & \dots & \langle a_2, a_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle a_n, a_1 \rangle & \langle a_n, a_2 \rangle & \dots & \langle a_n, a_n \rangle \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- This leads to **the normal equations**

$$A^* A x = A^* y$$

- $Ax_{\rm ls} = Py$ with
$$P = A(A^*A)^{-1}A^*$$

Facts: Any orthogonal projection operator satisfies

- $P = P^*$
- $P^2 = P$ (Idempotent operator)
- $\|Px\| \leq \|x\|$ for any x (contraction operator)
- $I - P \succeq 0$

Properties of full rank matrices

Suppose A is an $m \times n$ matrix. Then we always have

$$\mathbf{rank}(A) \leq \min(m, n)$$

If A is **full rank with** $m \geq n$

- $\mathbf{rank}(A) = n$ and $\mathcal{N}(A) = \{0\}$ ($Ax = 0 \Leftrightarrow x = 0$)
- A^*A is positive definite: for any $x \neq 0$ then

$$\langle A^*Ax, x\rangle = \langle Ax, Ax\rangle = \|Ax\|^2 > 0$$

Similarly, if A is **full rank** with $m \leq n$

- $\mathbf{rank}(A) = m$ and $\mathcal{N}(A^*) = \{0\}$
- AA^* is positive definite

The normal equations

$$A^*Ax = A^*b$$

- equivalent to the zero gradient condition:

$$\frac{d}{dx}\|Ax - y\|_2^2 = A^*(Ax - y) = 0$$

if A has a zero nullspace:

- least-squares solution can be found by solving the normal equations
- n equations in n variables with a positive definite coefficient matrix
- the closed-form solution is $x = (A^*A)^{-1}A^*y$
- $(A^*A)^{-1}A^*$ is a left inverse of A

Least-squares estimation

$$y = Ax + e$$

- x is what we want to estimate or reconstruct
- y is our measurements
- e is an unknown *noise* or *measurement error*
- ith row of A characterizes $i$the sensor or ith measurement (and A is deterministic)

Least-squares estimation: Choose as estimate the vector $\hat{x}$ that minimizes

$$\|A\hat{x} - y\|$$

i.e., minimize the deviation between what we actually observed (y), and what we would observe if $x = \hat{x}$, and there were no noise $(w = 0)$

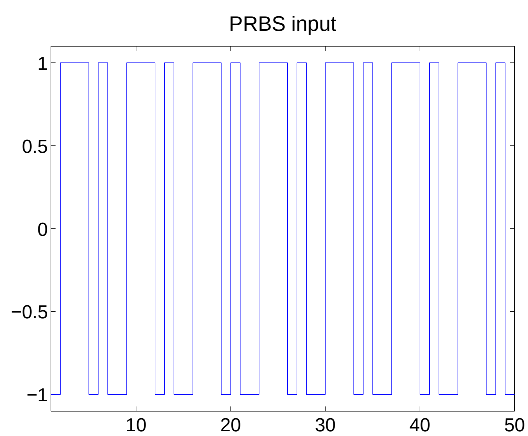
Example: first-order linear model

estimate the parameters a, b in a linear model

$$z(t) = az(t-1) + bu(t-1) + e(t)$$

from the measurement $z(t)$ and the input $u(t)$

- true parameters: $a = 0.8$, $b = 1$
- $u(t)$ is a PRBS sequence of magnitude -1,1 with period $M = 7$
- $e(t)$ is a zero mean white noise with variance 0.1



Estimation: choose $\hat{a}, \hat{b}$ that minimizes

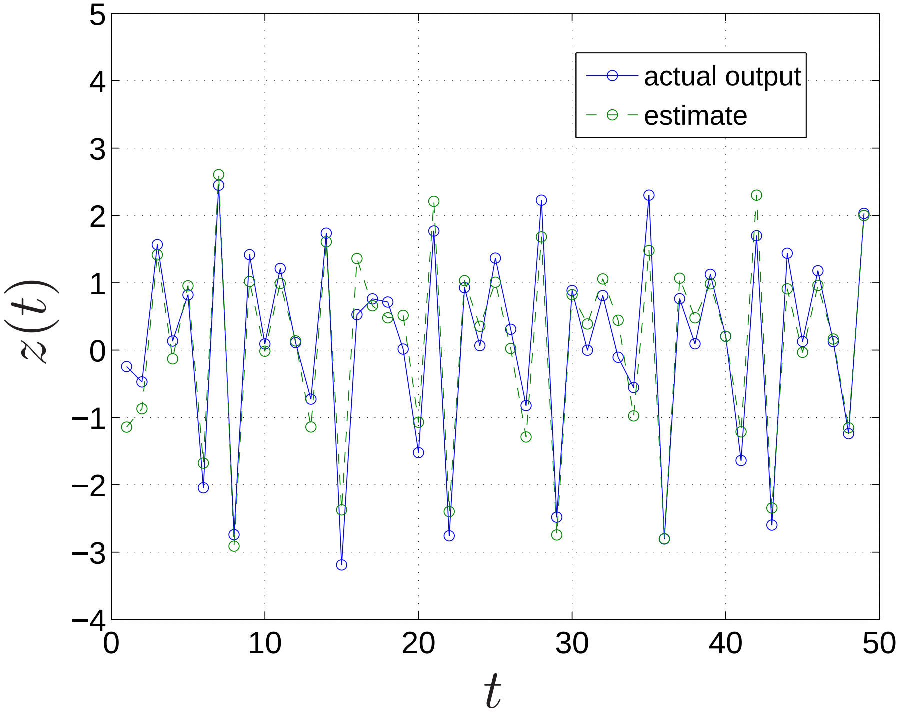
$$\sum_{t=1}^{N} \|z(t) - (\hat{a}z(t-1) + \hat{b}u(t-1))\|^2 = \|Ax - b\|^2$$

$$y = \begin{bmatrix} z(1) \\ \vdots \\ z(N) \end{bmatrix}, \quad A = \begin{bmatrix} z(0) & u(0) \\ \vdots & \vdots \\ z(N-1) & u(N-1) \end{bmatrix}, \quad x = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}$$

Results:

from one realization of $e(t)$,

$$\hat{a} = 0.7485, \quad \hat{b} = 1.0768$$



Analysis of the LS estimate (static case)

Assume that

- e is *white noise* with zero mean and covariance matrix I
- the least-square estimate is given by

$$\hat{x} = \operatorname{argmin} \|Ax - y\|$$

- The matrix A is *deterministic*

Then the following properties hold:

- $\hat{x}$ is an unbiased estimate of x ($\mathbf{E}\,\hat{x} = x$, or $\hat{x} = x$ when $e = 0$)
- The covariance matrix of $\hat{x}$ is given by

$$\mathbf{cov}(\hat{x}) = \mathbf{E}(\hat{x} - \mathbf{E}\,\hat{x})(\hat{x} - \mathbf{E}\,\hat{x})^* = (A^*A)^{-1}$$

BLUE property

The estimator defined by

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

is the *optimum unbiased linear least-mean-squares* estimator of x

Assume $\hat{z} = By$ is any other linear estimator of x

- require $BA = I$ in order for $\hat{z}$ to be unbiased
- $\mathbf{cov}(\hat{z}) = BB^*$
- $\mathbf{cov}(\hat{x}) = BA(A^*A)^{-1}A^*B^*$ (apply $BA = I$)

Using $I - P \succeq 0$, we conclude that

$$\mathbf{cov}(\hat{z}) - \mathbf{cov}(\hat{x}) = B(I - A(A^*A)^{-1}A^*)B^* \succeq 0$$

Suppose the covariance matrix of e is *not* I, say

$$\mathbf{E}\, ee^* = \Sigma$$

Scale the equation $y = Ax + e$ by $\Sigma^{-1/2}$

$$\Sigma^{-1/2} y = \Sigma^{-1/2} A x + \Sigma^{-1/2} e$$

The optimal unbiased linear least-mean-squares estimator of x is

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

The solution is a special case of a *weighted least-squares* problem

Weighted least-squares

$$\underset{x}{\text{minimize}} \quad \mathbf{tr}(Ax - y)^* W (Ax - y)$$

- W is a given positive definite matrix
- can be solved from the modified normal equations

$$A^* W A x = A^* W y$$

- Ax_{wls} is the *orthogonal projection* on $\mathcal{R}(A)$ w.r.t the new inner product

$$\langle x, y \rangle_W = \langle Wx, y \rangle$$

Analysis of the LS estimate (dynamic case)

Suppose we apply the LS method to a dynamical system

$$y(t) = H(t)\theta + \nu(t)$$

where the observations $y(1), y(2), \ldots, y(N)$ are available

Typically, $H(t)$ contains the past outputs and inputs

$$y(1), \ldots, y(t-1), u(1), \ldots u(t-1)$$

(hence $H(t)$ is no longer deterministic)

and $\nu(t)$ is white noise with covariance Λ

We obtain the following results

- The LS estimate is given by

$$\hat{\theta} = \left[\frac{1}{N}\sum_{t=1}^{N} H(t)^* H(t)\right]^{-1} \left[\frac{1}{N}\sum_{t=1}^{N} H(t)^* y(t)\right]$$

- $\hat{\theta}$ is consistent, $i.e.$,

$$\lim_{N\to\infty} \hat{\theta} = \theta$$

- $\sqrt{N}(\hat{\theta} - \theta)$ is asymptotically Gaussian distributed $\mathcal{N}(0, P)$ where

$$P = \Lambda[\mathbf{E}\, H(t)^* H(t)]^{-1}$$

Solving LS via Cholesky factorization

Every positive definite $B \in \mathbf{S}^n$ can be factored as

$$B = LL^T$$

where L is lower triangular with positive diagonal elements

Fact: For $B \succ 0$, a linear equation

$$Bx = b$$

can be solved in $(1/3)n^2$ flops

Solve the least-squares problem from the normal equations

$$A^* A x = A^* y$$

we have $A^* A \succ 0$ when A is full rank

Solving LS via QR factorization

- full QR factorization:

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

  with $[Q_1 \;\; Q_2] \in \mathbf{R}^{m\times m}$ orthogonal, $R_1 \in \mathbf{R}^{n\times n}$ upper triangular, invertible

- multiplication by orthogonal matrix doesn't change the norm, so

$$\begin{aligned} \|Ax-y\|^2 &= \left\| \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - y \right\|^2 \\ &= \left\| \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}^T \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} x - \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}^T y \right\|^2 \end{aligned}$$

$$= \left\| \begin{bmatrix} R_1 x - Q_1^T y \\ -Q_2^T y \end{bmatrix} \right\|^2$$

$$= \|R_1 x - Q_1^T y\|^2 + \|Q_2^T y\|^2$$

- this can be minimized by the choice $x_{\rm ls} = -R_1^{-1} Q_1^T y$ (which makes the first term zero)
- residual with optimal x is

$$Ax_{\rm ls} - y = -Q_2 Q_2^T y$$

- $Q_1 Q_1^T$ gives projection on $\mathcal{R}(A)$
- $Q_2 Q_2^T$ gives projection on $\mathcal{R}(A)^\perp$

References

Chapter 4 in
T. Söderström and P. Stoica, *System Identification*, Prentice Hall, 1989

Chapter 2-3 in
T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*, Prentice Hall, 2000

Lectures on
Linear least-squares and *The solution of a least-squares problem*, EE103, Lieven Vandenberghe, UCLA,
`http://www.ee.ucla.edu/~vandenbe/ee103.html`

Lectures on
Least-squares and *Least-squares applications*, EE263, Stephen Boyd, Stanford, `http://www.stanford.edu/class/ee263/letures.html`