6. Linear least-squares

- linear regression
- examples in engineering
- solving linear least-squares
- analysis of least-squares estimate
- computational aspects

Linear regression

- 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 $y \in \mathbf{R}^m$, $A \in \mathbf{R}^{m \times n}$, $x \in \mathbf{R}^n$

- 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

assume the model is the form of a polynomial of degree n

$$y(t) = a_0 + a_1 t + \cdots + a_r t^n$$

with unknown coefficients $a_0, \ldots, a_n$

this can be written in the form of linear regression as

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

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

Example 2: truncated weighting function

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

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

- an 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

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

Solving linear regressions

- the problem is to find an estimate of x from the measurements y and A
- if we choose the number of measurements, m 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

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

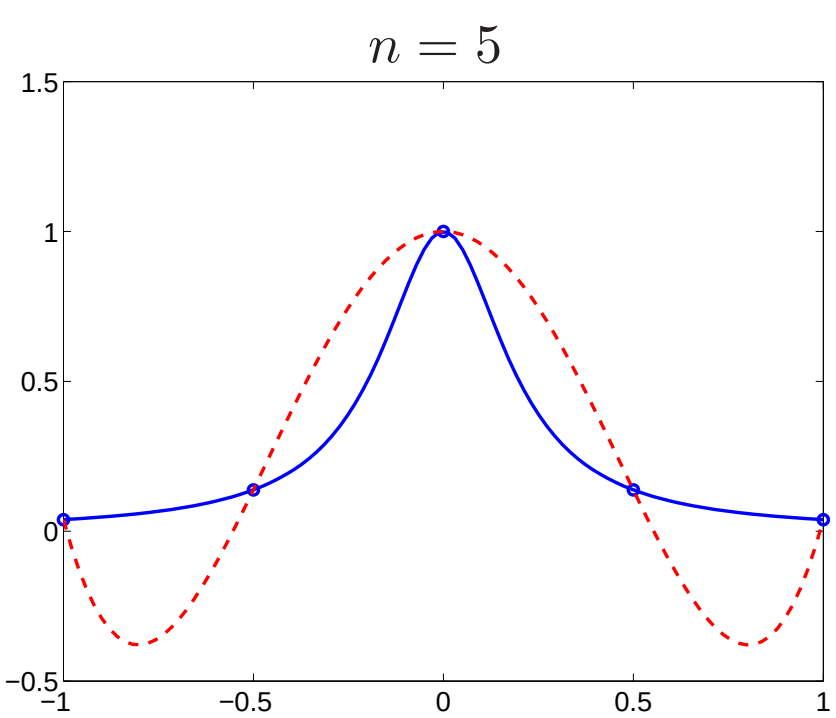
to data (t_1, y_1), (t_2, y_2), . . . , (t_m, y_m), i.e., choose the coefficients x_k so that

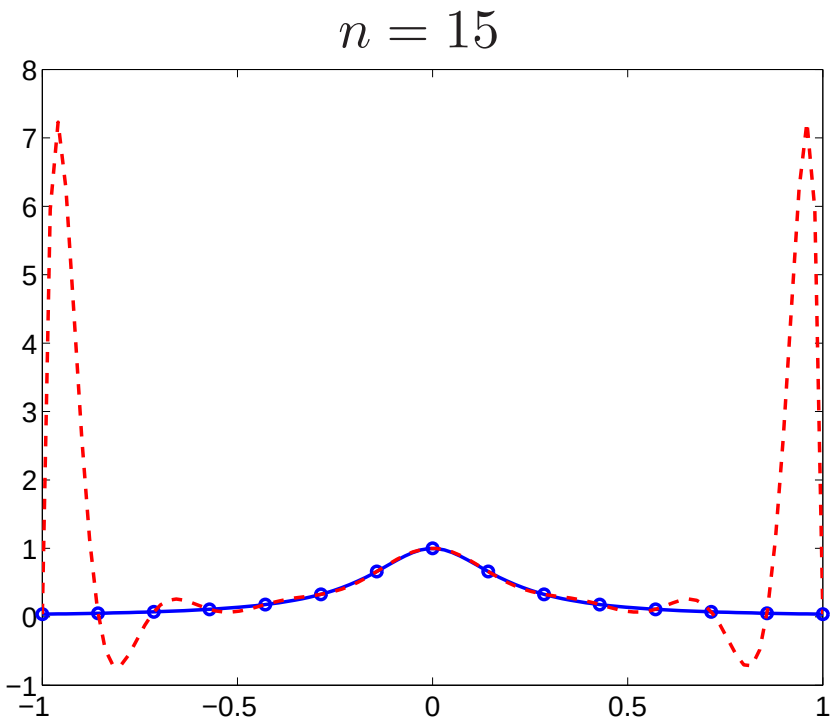
$$g(t_1) \approx y_1, \quad g(t_2) \approx y_2, \quad , g(t_m) \approx y_m$$

- $g_i(t) : \mathbf{R} \to \mathbf{R}$ 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 $Ax = y$



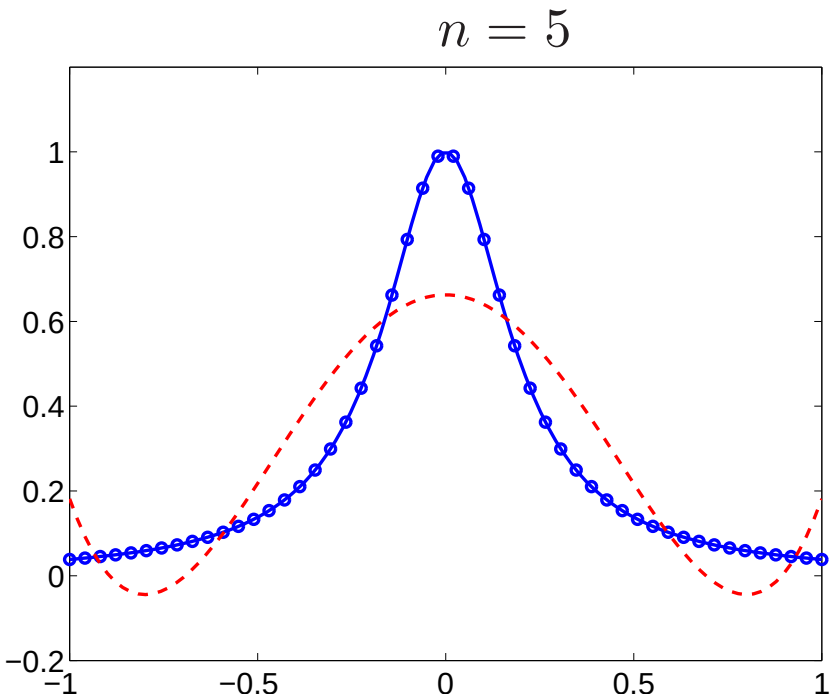


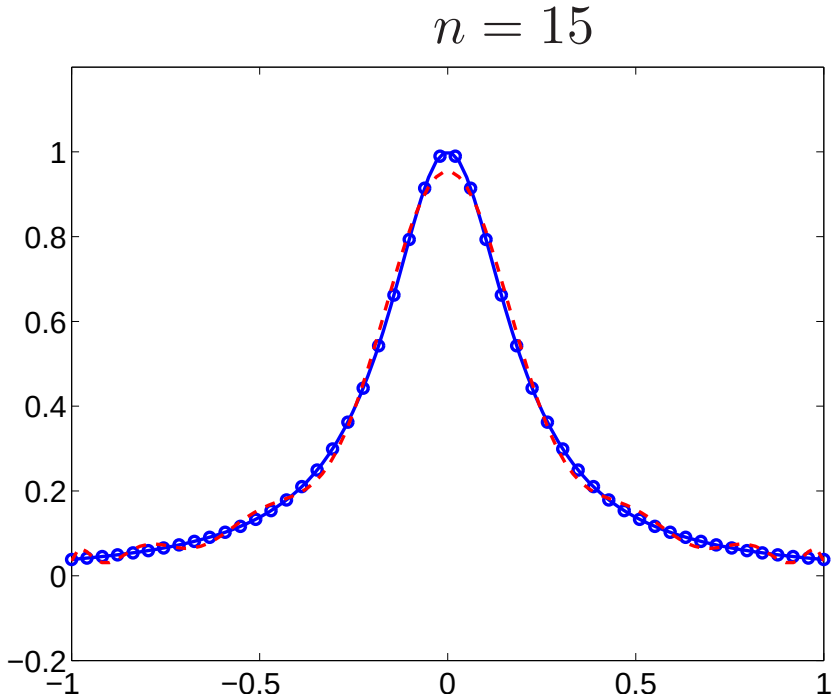
(blue solid line: f; red dashed line: polynomial g)

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

Some terminology

from the model $y = Ax + e$

variables y and A are commonly known as

y	A
endogenous variable	exogenous variable
dependent variable	independent variable
explained variable	explanatory variable
response variable	predictor
observable variable	regressor
	covariates
	manipulated variable

Closed-form of least-squares estimate

the zero gradient condition of LS objective is

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

which is equivalent to the **normal equation**

$$A^T Ax = A^T y$$

if A is full rank:

- 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 $x = (A^T A)^{-1} A^T y$
- $(A^T A)^{-1} A^T$ is a left inverse of A

Properties of full rank matrices

suppose A is an $m \times n$ matrix; we always have

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

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

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

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

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

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

Geometric interpretation of a LS problem

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

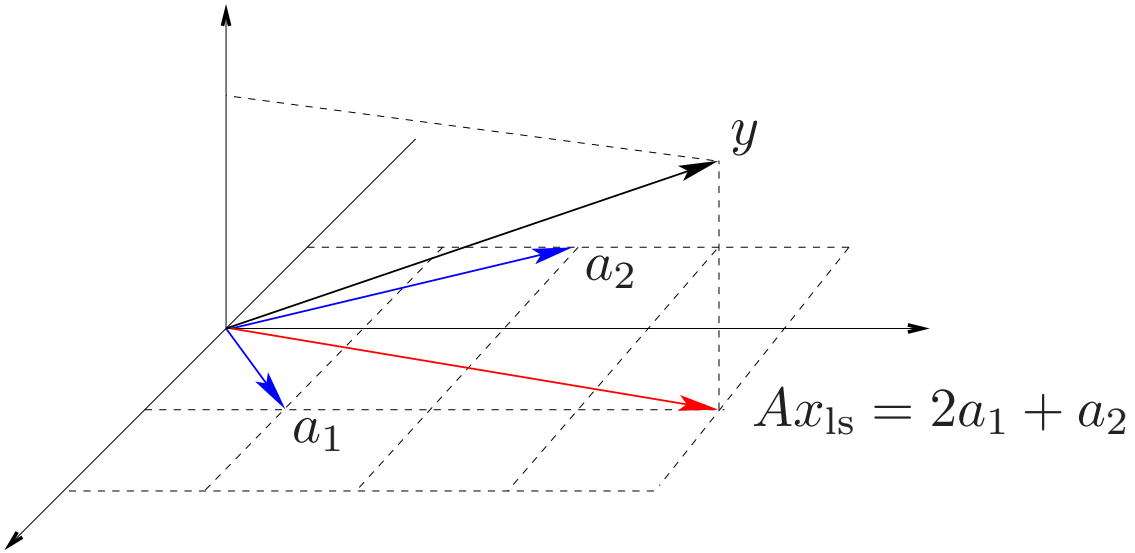
A is $m \times n$ with columns $a_1, a_2, \ldots a_m$

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

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

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

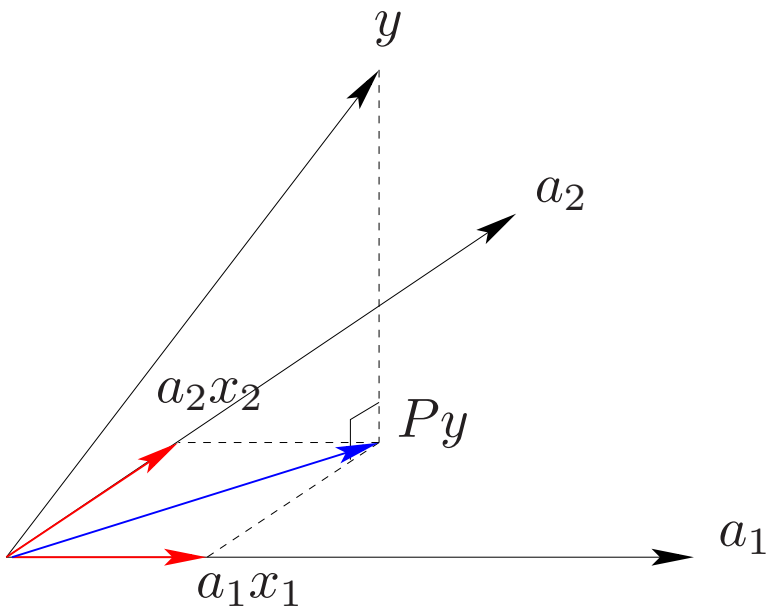
Example: $A = \begin{bmatrix} 1 & -1 \\ 1 & 2 \\ 0 & 0 \end{bmatrix}$, $y = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}$



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**

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

(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 $\{a_k\}$

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

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

- this also leads to the **normal equations**

$$A^T A x = A^T y$$

- $Ax_{\rm ls} = Py$ with

$$P = A(A^TA)^{-1}A^T$$

  if A has **full rank**

Definition: any orthogonal projection operator satisfies

- $P = P^T$
- $P^2 = P$ (Idempotent operator)

from its definition, any orthogonal projection operator obeys

- $\|Px\| \leq \|x\|$ for any x (contraction operator)
- $I - P \succeq 0$

Least-squares estimation

suppose y is generated under the dgp (data generating process)

$$y = Ax + e$$

- 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 an estimate $\hat{x}$ 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 $(e = 0)$

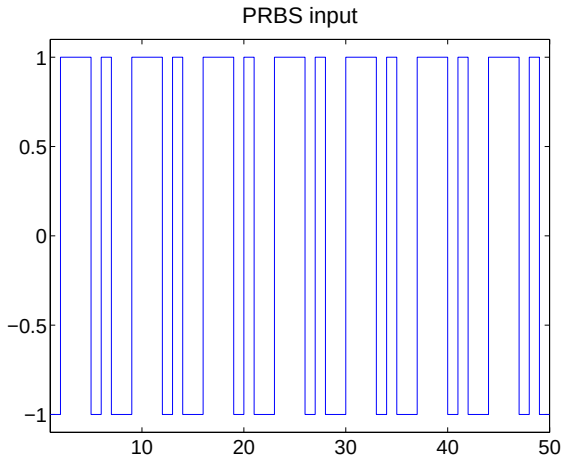
Example: first-order linear model

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

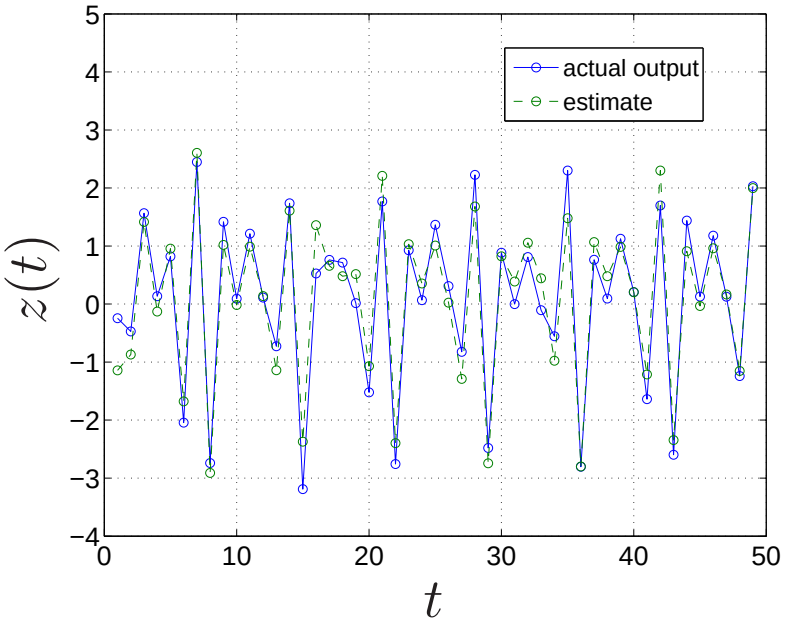
$$\sum_{t=1}^{N} \|z(t) - (\hat{a}z(t-1) + \hat{b}u(t-1))\|^2 = \|Ax - b\|^2$$

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

results:

from one realization of $e(t)$,

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



Analysis of the LS estimate (static case)

assumptions:

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

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

- the information 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

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

the expression of $\mathbf{cov}(\hat{x}) = (A^T A)^{-1} A^T \Sigma A (A^T A)^{-1}$ suggests that

- if A can be arbitrarily chosen, pick A that the covariance is small
- the covariance of the LS estimate depends on noise covariance

special case: noise covariance is diagonal

- $\Sigma = \mathbf{diag}(\sigma_1^2, \ldots, \sigma_N^2)$ (heteroskedasticity): e_i has different variances
- $\Sigma = \sigma^2 I$ (homoskedasticity): e_i has uniform variance

for homoskedasticity case, the covariance of the LS estimate reduces to

$$\mathbf{cov}(\hat{x}) = \sigma^2 (A^T A)^{-1}$$

BLUE property

under the dgp: $y = Ax + e$ and *homoskedasticity* of e, the LS estimator

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

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
- $\mathbf{cov}(\hat{z}) = BB^T$
- $\mathbf{cov}(\hat{x}) = BA(A^T A)^{-1} A^T B^T$ (apply $BA = I$)

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

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

suppose the covariance matrix of e is *not* I, says

$$\mathbf{E}ee^T = \Sigma$$

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

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

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

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

this is a special case of **weighted least-squares** problems

Weighted least-squares

given W a positive definite matrix and can be factorized as $W = L^T L$

a weighted least-squares problem is

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

- equivalent formulation: $\text{minimize}_x \quad \|L(Ax - y)\|_F^2$
- can be solved from the modified normal equations

$$A^T W A x = A^T W y$$

- 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

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

- the observations $y(1), y(2), \ldots, y(N)$ are available
- θ is the dynamical model parameter

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 $e(t)$ is white noise with covariance Σ

the LS estimate $\hat{\theta}_N$ (depending on N) given by

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

has the following properties (under some assumptions):

- $\hat{\theta}_N$ is consistent, $i.e.$, it converges to the true parameter in probability

$$\operatorname{plim} \hat{\theta}_N = \theta \quad \Longleftrightarrow \quad \lim_{N\to\infty} P(|\hat{\theta}_N - \theta| > \epsilon) = 0$$

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

$$P = \Sigma_x^{-1} \Sigma_{ux} \Sigma_x^{-1}$$

  Σ_x involves $\mathbf{E}[H(t)^T H(t)]$ and Σ_{ux} involes $\mathbf{E}[H(t)e(t)e(t)^T H(t)^T]$

the consistency results of LS estimate are based on *some assumptions*

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

$$= \left(\frac{1}{N}\sum_{t=1}^{N} H(t)^T H(t)\right)^{-1} \left(\frac{1}{N}\sum_{t=1}^{N} H(t)^T e(t)\right)$$

hence, $\hat{\theta}_N$ is consistent if

- $\mathbf{E}[H(t)^T H(t)]$ is nonsingular
  satisfied in most cases, except u is not persistently exciting of order n
- $\mathbf{E}[H(t)^T e(t)] = 0$
  not satisfied in most cases, except $e(t)$ is white noise

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 $(1/3)n^3$ flops

solve the least-squares problem from the normal equations

$$A^T Ax = A^T y$$

we have $A^T A \succ 0$ 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 $\begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \in \mathbf{R}^{m \times m}$ 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 $\mathcal{R}(A)^\perp$

Summary

- the linear least-squares method can be applied to models that are linear in the parameters
- a LS solution is unique if there is no colinearity (A is full rank)
- the method is mature, can be solve efficiently and is available in many softwares
- LS estimate has the BLUE property under the assumption that the noise in data generating process is homoskedastic
- LS estimate is consistent if the additive noise is uncorrelated with the regressors and the system is persistently excited

References

L. Ljung, *System Identification: Theory for the User*, Prentice Hall, Second edition, 1999

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

W.H. Greene, *Econometric Analysis*, Prentice Hall, 2008

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`