Statistics and optimization
Calculation of linear estimators

Sylvain Sardy

Université de Genève

1.1 Least squares

The least squares estimator $\hat{\boldsymbol{\alpha}}^{\mathrm{LS}}$ of a linear model solves

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \|\mathbf{y} - X\boldsymbol{\alpha}\|_2^2,$$

where X is an $N \times P$ matrix.

The cost function is convex (strictly if $\ker X = \{\mathbf{0}\}$) and continuously differentiable over $\mathbb{R}^P$ convex, so a necessary and sufficient condition for $\boldsymbol{\alpha}$ to be solution is

$$X^{\mathrm{T}} X \boldsymbol{\alpha} = X^{\mathrm{T}} \boldsymbol{y}.$$

Simple case: X has P orthogonal columns.

Suppose $\mathrm{rank}(X) = P$:

- Cholesky factorisation of

$$X^{\mathrm{T}} X = G G^{\mathrm{T}},$$

  where G is $P \times P$ is a unique lower triangular matrix with positive diagonal entries.

  Requires $(N + P/3)P^2$ flops (FLoating-point Operations Per Second).

- QR factorisation of

$$X = QR = [Q_1 \, Q_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1 ,$$

  where:
  - Q is $N \times N$ and orthonormal
  - Q_1 is $N \times P$ and $\mathrm{range}(X) = \mathrm{range}(Q_1)$
  - Q_2 is $N \times (N - P)$ and $\mathrm{range}(X)^{\perp} = \mathrm{range}(Q_2)$
  - R_1 is upper triangular with positive diagonal entries. Moreover $R_1 = G^{\mathrm{T}}$.

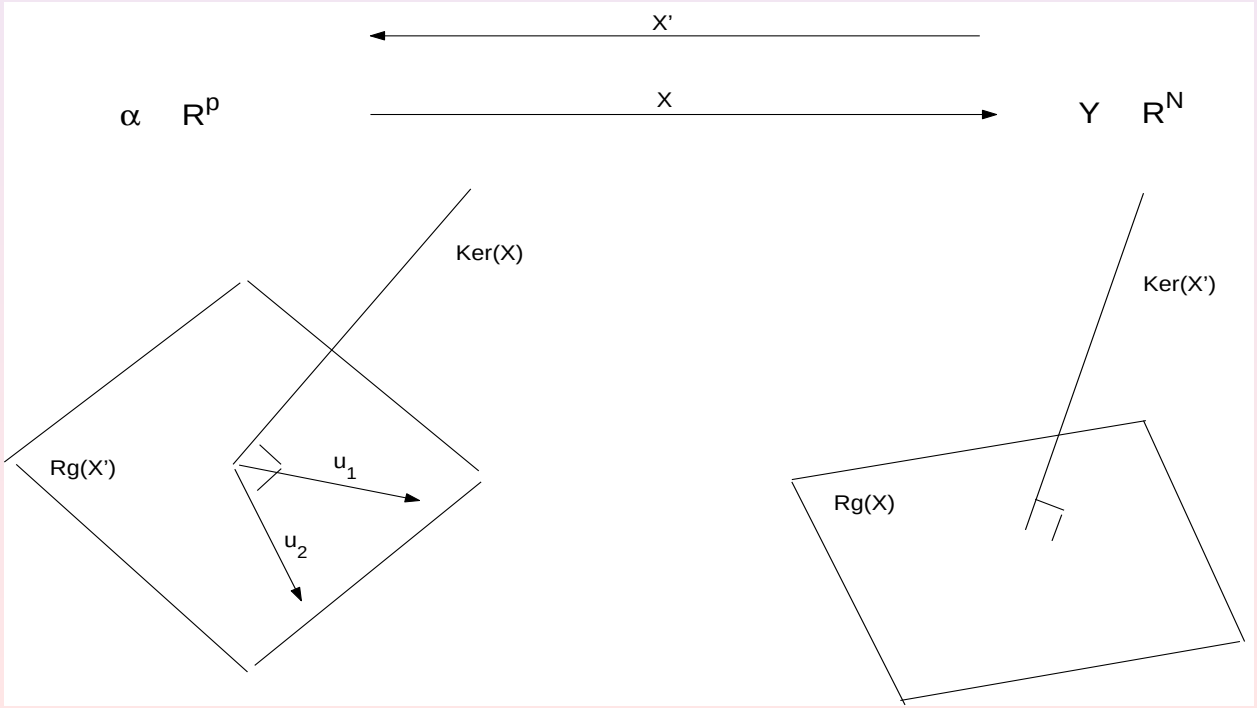
  Requires $2(N - P/3)P^2$ flops.
- Block coordinate relaxation converges to the solution.

- Eigen-decomposition of $X^{\mathrm{T}}X$:

$$
\begin{aligned}
X^{\mathrm{T}}X &= UD^2U^{\mathrm{T}} \\
&= [U_R U_k] \begin{bmatrix} \mathrm{diag}(\mathbf{d}^2) & 0 \\ 0 & \mathrm{diag}(\mathbf{0}) \end{bmatrix} [U_R U_k]^{\mathrm{T}} \\
&= U_R D_R^2 U_R^{\mathrm{T}},
\end{aligned}
$$

where:

- U_R is an orthonormal basis for the range of X^{T},
- U_k is an orthonormal basis for the kernel of X.



1.2 Least squares with constaints

General formulation: consider the problem

$$\min_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) \quad \text{s.t.} \quad h(\boldsymbol{\alpha}) = \mathbf{0}, \tag{1}$$

where $f : \Theta \subset \mathbb{R}^P \to \mathbb{R}$ and $h : \Theta \subset \mathbb{R}^P \to \mathbb{R}^M$ are continuously differentiable and Θ is an open set containing a local minimum $\boldsymbol{\alpha}^*$.

Definition: the Lagrangian function $L : \Theta \subset \mathbb{R}^{P+M} \to \mathbb{R}$ is defined by

$$L(\boldsymbol{\alpha}, \boldsymbol{\lambda}) = f(\boldsymbol{\alpha}) + \boldsymbol{\lambda}^{\mathrm{T}} h(\boldsymbol{\alpha}).$$

Proposition (Lagrange multiplier theorem – Necessary condition): Suppose the M columns of $\nabla h(\boldsymbol{\alpha}^*)$ are linearly independent at a local minimum $\boldsymbol{\alpha}^*$ of (1). Then there exists a unique vector $\boldsymbol{\lambda}^*$

$$\nabla_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*) = \mathbf{0} \quad \text{and} \quad \nabla_{\boldsymbol{\lambda}} L(\boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*) = \mathbf{0}.$$

Consider the linear model

$$\mathbf{y} = X\boldsymbol{\alpha} + \boldsymbol{\epsilon}.$$

Application: Let $\mathbf{y} \in \mathbb{R}^N$ and $X \in \mathbb{R}^{N \times P}$ of rank N. Then

$$X^{\mathrm{T}}(XX^{\mathrm{T}})^{-1}\mathbf{y} = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^P:\ X\boldsymbol{\alpha} = \mathbf{y}} \|\boldsymbol{\alpha}\|_2$$

(it is unique!) whether $P = N$ or $P > N$.

Application: Consider testing $H_0 : A\boldsymbol{\alpha} = \mathbf{c}$ with A full row rank. Fisher F-test requires

$$\hat{\boldsymbol{\alpha}}_{H_0} = \arg \min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\alpha}\|_2^2 \quad \text{s.t.} \quad A\boldsymbol{\alpha} = \mathbf{c}.$$

Solution through Lagrange function
$L(\boldsymbol{\alpha}, \boldsymbol{\lambda}) = \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\alpha}\|_2^2 + \boldsymbol{\lambda}^{\mathrm{T}}(A\boldsymbol{\alpha} - \mathbf{c})$.

2. Ridge regression (*Hoerl, 1962; Hoerl and Kennard, 1970*)

For a given regularization parameter $\lambda \geq 0$, RR solves

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{Y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^2.$$

Ways to calculate the RR estimate:

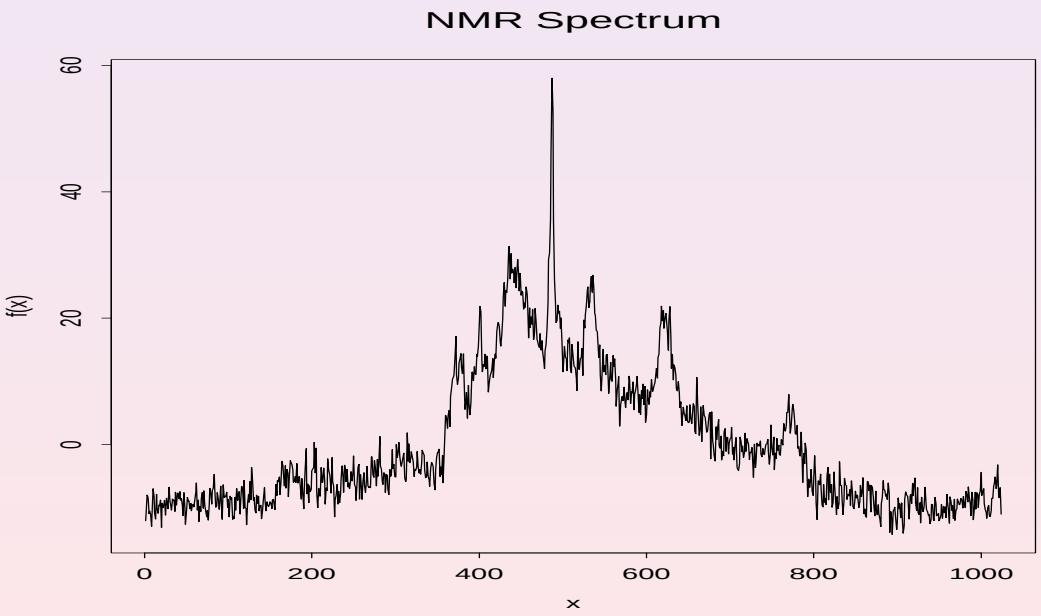
- Matrix inversion
- SD, CG
- SVD has an overhead cost, but efficient if several λ's have to be used.

3. Splines-based smoothing (*Wahba, 1980's; Duchon, 1976*)

Motivation: regression problem

$$y_n = \mu(x_n) + \epsilon_n, \quad \text{where} \quad \epsilon_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2).$$

Goal: estimate univariate function μ.



Existing estimator: running me(di)an, kernel, local polynomials

Thin plate splines come from approximation theory: Let $W_{2,2}$ be the Sobolev space of function that has squared integrable derivatives up to second derivative. Then

$$\min_{\mu \in W_{2,2}} \|\mathbf{y} - \mu(\mathbf{x})\|_2^2 + \lambda \int (\mu''(x))^2 dx,$$

where $\mu(\mathbf{x}) = (\mu(\mathbf{x_1}), \ldots, \mu(\mathbf{x_N}))$ has a solution of the form

$$\mu(x) = \sum_{n=1}^{N} \alpha_n \phi_n(x),$$

where ϕ_n are known thin plate splines.
In other words, an infinite dimensional problem becomes a finite dimensional problem.

Letting Φ be the $N \times N$ regression matric such that

$$\Phi_{ij} = \phi_j(x_i), \quad i, j = 1, \ldots, N,$$

the problem becomes

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^{\mathrm{N}}} \|\mathbf{y} - \Phi\boldsymbol{\alpha}\|_2^2 + \lambda \boldsymbol{\alpha}^{\mathrm{T}} Q \boldsymbol{\alpha},$$

where $\boldsymbol{\alpha}^{\mathrm{T}} Q \boldsymbol{\alpha} = \int (\mu''(x))^2 dx$. Because $\int (\mu''(x))^2 dx = 0$ for polynomials of degree 1, then Q has rank $N - 2$.