Lecture notes for DDSM—work in progress
Federico A. Ramponi
Revision 0.0.1 - November 6, 2018.

1 Least squares

1.1 Fitting data with a function

[Rephrase. We do not yet assume an explicit model for y_i.]

Suppose that we have collected a sample of measures $(u_1, y_1), \dots, (u_N, y_N)$, where $u_i \in \mathcal{U}$ and $y_i \in \mathbb{R}$ for all $i = 1, \dots, N$. We want to "explain" the measures $y_1, \dots, y_N$ with a function $y_i \simeq f_\theta(u_i)$ that further depends on a parameter θ. The parameter $\theta \in \mathbb{R}^p$ is supposed to identify a function within a family of functions $\hat{\mathcal{M}} = \{f_\theta : \mathcal{U} \to \mathbb{R} \ : \ \theta \in \mathbb{R}^p\}$ of our choice ($\hat{\mathcal{M}}$ is our "theory" or "*a priori* knowledge"). This goal admits the following interpretation: the observed quantities y_i really behave according to a "true" function $f^* \in \hat{\mathcal{M}}$, and such function is properly identified by a "true" parameter θ°; however, the quantities y_i are in fact *measures* corrupted by a noise term $\boldsymbol{\varepsilon}_i$:

$$y_i = f^*(u_i) + \boldsymbol{\varepsilon}_i \quad \text{for all } i = 1, \dots N. \tag{1}$$

If the family $\hat{\mathcal{M}}$ is sufficiently regular (typical example: polynomials whose coefficients are the components of θ), a function f_θ is a good approximation of f^* when θ is close to θ°; hence our goal will be to find a good *estimate* $\hat{\theta}$ of θ°. For the moment, we don't make any assumption on $\boldsymbol{\varepsilon}_i$; later, it will be natural to model this source of *uncertainty* as a zero-mean random variable.

The least squares method prescribes to find an "optimal" estimate $\hat{\theta}$ by minimizing the sum of squares:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^N (y_i - f_\theta(u_i))^2 = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^N r_i(\theta)^2. \tag{2}$$

The expression $r_i(\theta) = y_i - f_\theta(u_i)$ is called the i-th *residual* with respect to the choice of θ. More generally, a *weighted* sum of squares can be minimized:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^N w_i \left(y_i - f_\theta(u_i)\right)^2, \tag{3}$$

where $w_i > 0$ for all i. A small w_i means that the corresponding measure y_i is supposed to be more "imprecise", i.e. more *noisy* ($\boldsymbol{\varepsilon}_i$ is likely to affect it significantly), and that therefore the corresponding $r_i(\theta)$ has to be considered "less important". When the weights w_i are all equal (without loss of generality, suppose $w_i = 1$ for all i), we fall back in the case (2).

1.2 Linear models

Let us introduce a fundamental simplifying hypothesis: $f_\theta(u)$ is linear in the parameter θ, i.e. it takes the form $f_\theta(u) = \varphi(u)^\top \theta$, where $\varphi : \mathcal{U} \to \mathbb{R}^p$, called *regressor function*, can still be nonlinear. For notational convenience, we let $\varphi_i = \varphi(u_i)$. The measurement model (1) becomes

$$y_i = \varphi_i^\top \theta + \boldsymbol{\varepsilon}_i \quad \text{for all } i = 1, \dots n, \tag{4}$$

and is known in literature as a *linear model*. The least squares problem (2) reads

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{N} \left(y_i - \varphi_i^\top \theta \right)^2, \tag{5}$$

and the weighted problem (3) reads

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{N} w_i \left(y_i - \varphi_i^\top \theta \right)^2. \tag{6}$$

Note that (5) and (6), besides yielding different results, are completely equivalent. On one hand, (6) is a particular case of (5) where $w_i = 1$ for all i; on the other hand, letting $\bar{y}_i = \sqrt{w_i}\ y_i$ and $\bar{\varphi}_i = \sqrt{w_i}\ \varphi_i$ the sum in (6) takes the same form of (5).

1.3 Normal equations

To find a solution of (6) we note that the sum of squares

$$\hat{J}(\theta) = \sum_{i=1}^{N} w_i \left(y_i - \varphi_i^\top \theta \right)^2 \tag{7}$$

is a *convex* and *differentiable* function of θ.

✎ **Tools from analysis: a characterization of convex differentiable functions**

Theorem 1.1 *Suppose that $\hat{J} : \mathbb{R}^p \to \mathbb{R}$ is convex and differentiable over $\mathbb{R}^p$ (its gradient $\nabla \hat{J}(\theta)$ exists at each point $\theta \in \mathbb{R}^p$). Then*

$$\hat{J}(\theta) \geq \hat{J}(\hat{\theta}) + \nabla \hat{J}(\hat{\theta})^\top (\theta - \theta^\circ)$$

for all $\theta, \hat{\theta} \in \mathbb{R}^p$.

Proof. See [1, p. 70]. □

An important consequence of the theorem is that if we find a point $\hat{\theta}$ such that $\nabla \hat{J}(\hat{\theta}) = 0$, then $\hat{J}(\theta) \geq \hat{J}(\hat{\theta})$ for all $\theta \in \mathbb{R}^p$, so that $\hat{\theta}$ is a minimum point for $\hat{J}$.

We differentiate (7) and set the result equal to zero:

$$\begin{aligned} \frac{\partial \hat{J}(\theta)}{\partial \theta} &= \sum_{i=1}^{N} w_i\ 2 \left(y_i - \varphi_i^\top \theta \right) \left(-\varphi_i^\top \right) \\ &= -2 \sum_{i=1}^{N} w_i \left(\varphi_i^\top y_i - \theta^\top \varphi_i \varphi_i^\top \right) = 0. \end{aligned}$$

After some algebraic manipulation, we come to the following equation, called *normal equations* (usually in the plural):

$$\left(\sum_{i=1}^{N} w_i\ \varphi_i \varphi_i^\top \right) \theta = \sum_{i=1}^{N} w_i\ \varphi_i y_i \quad \text{(weighted problem);} \tag{8}$$

$$\left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top \right) \theta = \sum_{i=1}^{N} \varphi_i y_i \quad \text{(non-weighted problem, i.e. } w_i \equiv 1\text{).} \tag{9}$$

We will show in the following sections that at least one solution to (8) (resp. (9)) always exists. If, moreover, the $p \times p$ matrix $\sum_{i=1}^{N} w_i \ \varphi_i \varphi_i^\top$ (resp. $\sum_{i=1}^{N} \varphi_i \varphi_i^\top$) is invertible, then the solution is unique and reads

$$\hat{\theta} = \left(\sum_{i=1}^{N} w_i \ \varphi_i \varphi_i^\top \right)^{-1} \sum_{i=1}^{N} w_i \ \varphi_i y_i. \tag{10}$$

1.4 Compact notation

To shorten notation, we let

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \Phi = \begin{bmatrix} \varphi_1^\top \\ \varphi_2^\top \\ \vdots \\ \varphi_N^\top \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}. \tag{11}$$

With the compact notation (11), the measurement model (4) becomes

$$Y = \Phi \theta^\circ + \boldsymbol{\varepsilon} \quad \text{(linear model)}, \tag{12}$$

and the least squares problem (5) (without weights) reads

$$\begin{aligned} \hat{\theta} &= \arg\min_{\theta \in \mathbb{R}^p} \left(Y - \Phi^\top \theta \right)^\top \left(Y - \Phi^\top \theta \right) \\ &= \arg\min_{\theta \in \mathbb{R}^p} \left\| Y - \Phi^\top \theta \right\|^2 = \arg\min_{\theta \in \mathbb{R}^p} \left\| Y - \Phi^\top \theta \right\| \end{aligned} \tag{13}$$

(the last equality holds because taking a square root may change the *minimum* attained, but not the *minimum point*). If we further define

$$W = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_N \end{bmatrix},$$

then the *weighted* least squares problem (6) becomes

$$\begin{aligned} \hat{\theta} &= \arg\min_{\theta \in \mathbb{R}^p} \left(Y - \Phi^\top \theta \right)^\top W \left(Y - \Phi^\top \theta \right) \\ &= \arg\min_{\theta \in \mathbb{R}^p} \left\| Y - \Phi^\top \theta \right\|_W^2 = \arg\min_{\theta \in \mathbb{R}^p} \left\| Y - \Phi^\top \theta \right\|_W \end{aligned} \tag{14}$$

where the norm $\| \cdot \|_W$ is similar to the canonical Euclidean norm, but descends from the scalar product $\langle x, y \rangle_W = x^\top W y$ instead of the usual one $\langle x, y \rangle = x^\top y$. When $W = I$ (all $w_i = 1$), (13) and (14) are identical.

Finally, the normal equations (9) become

$$\left(\Phi^\top W \Phi \right) \theta = \Phi^\top W Y \quad \text{(weighted)}, \tag{15}$$

$$\left(\Phi^\top \Phi \right) \theta = \Phi^\top Y \quad \text{(non-weighted)}. \tag{16}$$

If the $p \times p$ matrices $\Phi^\top W \Phi$, resp. $\Phi^\top \Phi$ are invertible, then (15), (16) admit a unique solution (compare with (10)):

$$\hat{\theta} = \left(\Phi^\top W \Phi\right)^{-1} \Phi^\top W Y \quad \text{(weighted)}, \tag{17}$$

$$\hat{\theta} = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top Y \quad \text{(non-weighted)}. \tag{18}$$

1.5 The normal equations have at least a solution

We provide here a first, algebraic proof that the normal equations (16) do have at least one solution (we stick to the non-weighted version, but this is just for simplicity).

✎ **Tools from linear algebra: null space and range of the transpose**

For any matrix $M \in \mathbb{R}^{m \times n}$,

$$\begin{aligned} x \in \text{null } M \;&\Rightarrow\; Mx = 0 \;\Rightarrow\; M^\top M x = 0 \;\Rightarrow\; x \in \text{null } M^\top M; \\ x \in \text{null } M^\top M \;&\Rightarrow\; M^\top M x = 0 \;\Rightarrow\; x^\top M^\top M x = 0 \;\Rightarrow\; \|Mx\|^2 = 0 \\ &\Rightarrow\; Mx = 0 \;\Rightarrow\; x \in \text{null } M. \end{aligned}$$

Hence, null M = null $M^\top M$. Moreover,

$$\begin{aligned} x \in \text{null } M \;&\Leftrightarrow\; Mx = 0 \\ &\Leftrightarrow\; v^\top M x = 0 \quad \text{for all } v \in \mathbb{R}^m \\ &\Leftrightarrow\; \left(M^\top v\right)^\top x = 0 \quad \text{for all } v \in \mathbb{R}^m \\ &\Leftrightarrow\; M^\top v \perp x = 0 \quad \text{for all } v \in \mathbb{R}^m \\ &\Leftrightarrow\; x \perp \text{range } M^\top, \quad \text{that is,} \\ &\Leftrightarrow\; x \in \left(\text{range } M^\top\right)^\perp. \end{aligned}$$

It follows, by taking another complement and simplifying (this is OK because $\mathbb{R}^n$ is finite-dimensional), that

$$\begin{aligned} \text{range } M^\top &= (\text{null } M)^\perp; \quad \text{in the same way,} \\ \text{range } M^\top M &= \left(\text{null } M^\top M\right)^\perp; \end{aligned}$$

and finally, since null M = null $M^\top M$,

$$\text{range } M^\top = \text{range } M^\top M.$$

Let $M = \Phi$. The above result says that

$$\text{range } \Phi^\top = \text{range } \Phi^\top \Phi. \tag{19}$$

But since $\Phi^\top Y$ belongs to the left-hand subspace, it also belongs to range $\Phi^\top \Phi$, that is, to the set $\{\Phi^\top \Phi\ \theta \text{ for some } \theta \in \mathbb{R}^p\}$; it follows that there must exist at least $\hat{\theta}$ that solves the normal equations

$$\left(\Phi^\top \Phi\right) \hat{\theta} = \Phi^\top Y.$$

Note: the dimensions of the ranges in (19) are the respective ranks:

$$\text{rank } \Phi^\top \Phi = \text{rank } \Phi^\top = \text{rank } \Phi.$$

Therefore to check the invertibility of $\Phi^\top \Phi$ one does not need to actually compute it; it is sufficient to check the rank of Φ (a full rank $= p$ implies the invertibility of $\Phi^\top \Phi$).

1.6 Interpretation in terms of projections

The method of least squares is closely related with orthogonal projections. We start from a classical result, characterizing the point of a subspace which is closest to another point in the sense of the Euclidean distance:

Theorem 1.2 *Let $\mathcal{C}$ be a closed convex set of a Hilbert space*[1] *$\mathcal{H}$, and $Y \in \mathcal{H}$. Then there exists a unique vector $\hat{Y} \in \mathcal{C}$ such that $\|Y - \hat{Y}\| \leq \|Y - x\|$ for all $x \in \mathcal{C}$. A necessary and sufficient condition for $\hat{Y}$ to be the unique minimizing vector is that $\left\langle Y - \hat{Y}, x - \hat{Y} \right\rangle \leq 0$ for all $x \in \mathcal{C}$.*

[Insert nice figure.]

All finite-dimensional spaces like $\mathbb{R}^N$, endowed with the "standard" scalar product defined by $\langle x, y \rangle = x^\top y$, are Hilbert spaces, so Theorem 1.2 applies naturally to the closed convex sets of $\mathbb{R}^N$. Moreover, all subspaces of $\mathbb{R}^N$ are closed convex sets, and it is possible to show that when $\mathcal{C}$ is a subspace, a more stringent necessary and sufficient condition holds: $\left\langle Y - \hat{Y}, x - \hat{Y} \right\rangle = 0$. These particular cases are resumed in the following result.

Theorem 1.3 *Let $\mathcal{W}$ be a subspace of $\mathbb{R}^N$, and $Y \in \mathbb{R}^N$. Then there exists a unique vector $\hat{Y} \in \mathcal{W}$ such that $\|Y - \hat{Y}\| \leq \|Y - w\|$ for all $w \in \mathcal{W}$. A necessary and sufficient condition for $\hat{Y}$ to be the unique minimizer is that $Y - \hat{Y} \perp w$ for all $w \in \mathcal{W}$.*

The minimizer $\hat{Y}$ is called the *orthogonal projection* of Y on the subspace $\mathcal{W}$.

[Insert nice figure.]

Let now $\mathcal{W} = \text{range } \Phi$. The vectors in $\mathcal{W}$ are precisely those with the form $w = \Phi\theta$ for some $\theta \in \mathbb{R}^p$. The least squares problem asks to minimize $\|Y - \Phi\theta\|^2$, but this is the same as to minimize $\|Y - \Phi\theta\|$, which in turn is equivalent to minimize $\|Y - w\|$ with respect to $w = \Phi\theta \in \text{range } \Phi$ that is, to find $\hat{Y} = \Phi\hat{\theta}$ such that $\|Y - \hat{Y}\|$ is minimal. Theorem 1.3 ensures that such a Y exists; hence a solution $\hat{\theta}$ of the least squares problem also exists. The theorem states that $\hat{Y}$ is unique; this does not imply that $\hat{\theta}$ is also unique! Indeed, $\hat{\theta}$ is unique if and only if Φ has full rank p.

Now let us apply the second part of Theorem 1.3: $\hat{Y} = \Phi\hat{\theta}$ is a minimizing vector, and $\hat{\theta}$ is the least squares solution, if and only if $Y - \Phi\hat{\theta} \perp w$ for all $w \in \text{range } \Phi$. Let $c_1, \ldots, c_p$ be the *columns* of Φ (whereas the regressors are its *rows*). Since range $\Phi = \text{span } \{c_1, \ldots, c_p\}$, to check the orthogonality condition it is sufficient to check that $Y - \Phi\hat{\theta} \perp c_i$ for all the columns c_i. Explicitly,

$$c_i^\top (Y - \Phi\hat{\theta}) = 0 \quad \text{for all } i = 1, \ldots, p. \tag{20}$$

[1] A Hilbert space $\mathcal{H}$ is a *complete Euclidean space*, i.e. a vector space endowed with a scalar product $\langle \cdot, \cdot \rangle$, with the norm defined by $\|x\| = \sqrt{\langle x, x \rangle}$, and with the distance defined by $d(x_1, x_2) = \|x_1 - x_2\|$, and such that every Cauchy sequence $(x_1, x_2, x_3, \ldots)$ has a limit $\bar{x} \in \mathcal{H}$.

Stacking the rows $c_i^\top$ on each other we get $\Phi^\top$, hence stacking the equations (20) on each other we obtain:

$$\Phi^\top (Y - \Phi\hat{\theta}) = 0,$$

which finally yields, again, the normal equations:

$$\left(\Phi^\top \Phi\right) \hat{\theta} = \Phi^\top Y.$$

Conclusion: $\hat{Y} = \Phi\hat{\theta}$ is the unique orthogonal projection of Y if and only if $\hat{\theta}$ solves the normal equations.

If $\left(\Phi^\top \Phi\right)$ is invertible, then $\hat{\theta} = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top Y$ and $\hat{Y} = \left(\Phi \left(\Phi^\top \Phi\right)^{-1} \Phi^\top\right) Y = \Pi_\Phi Y$. The matrix $\Pi_\Phi := \Phi \left(\Phi^\top \Phi\right)^{-1} \Phi^\top$ is a so-called *orthogonal projection matrix.*

✎ **Tools from linear algebra: projection matrices**

A matrix $\Pi \in \mathbb{R}^{N \times N}$ is called *symmetric* if $\Pi = \Pi^\top$ and *idempotent* if $\Pi^2 = \Pi$. An idempotent symmetric matrix is called an *orthogonal projection* matrix. Any such matrix has the form $\Pi = A \left(A^\top A\right)^{-1} A^\top$ for some "tall", full rank matrix A. Note that the pre-multiplication by Π leaves A unchanged:

$$\Pi A = A \left(A^\top A\right)^{-1} A^\top A = A;$$

therefore Π leaves any column of A (and hence any linear combination of columns of A) unchanged. On the other hand, if x is any vector orthogonal to all the columns of A, then

$$\Pi x = A \left(A^\top A\right)^{-1} \left(A^\top v\right) = 0.$$

Therefore the job of Π is to find the orthogonal projection of a vector on range A.

Example. To find the orthogonal projection of

$$Y = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \text{on} \quad \text{span} \left\{ \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}, \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix} \right\}, \quad \text{let} \quad A = \begin{bmatrix} 4 & 7 \\ 5 & 8 \\ 6 & 9 \end{bmatrix},$$

then compute $\Pi = A \left(A^\top A\right)^{-1} A^\top$, and the projection is given by $\hat{Y} = \Pi Y$.

Note that $(I - \Pi)$ is also symmetric and idempotent, indeed

$$(I - \Pi)^2 = I - \Pi - \Pi + \Pi^2 = I - \Pi - \Pi + \Pi = (I - \Pi).$$

Therefore $(I - \Pi)$ is another orthogonal projection matrix; if Π projects on range A, then $(I - \Pi)$ projects on $(\text{range } A)^\perp$, and

$$x = \Pi x + (I - \Pi)x = x_A + x_\perp$$

is the unique decomposition of x as the sum of a vector in range A and a vector in its orthogonal complement.

1.7 Constrained least squares

Sometimes, in solving a regression problem with the method of least squares, it is convenient to restrict θ to a subspace of $\mathbb{R}^p$. This is equivalent to impose a constraint of the form $K\theta = 0$, where the matrix $K \in \mathbb{R}^{m\times p}$ has full rank $m < p$ (each of its m rows imposes a constraint and reduces by 1 the dimension of the space where θ lives, so that θ is confined to a $(p-m)$-dimensional subspace). Assume that Φ has full rank and consider the following problem:

$$\begin{aligned} \min_{\theta\in\mathbb{R}^p} \quad & \|Y - \Phi\theta\|^2 \\ \text{subject to} \quad & K\theta = 0. \end{aligned} \tag{21}$$

Its solution may be substantially different from that of the standard, unconstrained problem $\min_{\theta\in\mathbb{R}^p} \|Y - \Phi\theta\|^2$. Before solving (21) it is good to take a little review of the fundamental tool of constrained optimization.

✎ **Tools from analysis: Lagrange's lemma**

The following proposition, albeit simple, is foundational in constrained optimization:

Lemma 1.1 *Let $F : \mathbb{R}^n \to \mathbb{R}$ and $\Lambda : \mathbb{R}^n \to \mathbb{R}$ be functions, and $C \subset \mathbb{R}^n$ be any subset. Suppose that a point $\bar{x} \in C$ satisfies*

1. *$\Lambda(\bar{x}) \geq \Lambda(x)$ for all $x \in C$,*
   i.e. $\bar{x} = \arg\max_{x\in C} \Lambda(x)$,
   i.e. $\bar{x}$ maximizes Λ over C;

2. *$F(\bar{x}) + \Lambda(\bar{x}) \leq F(x) + \Lambda(x)$ for all $x \in \mathbb{R}^n$,*
   i.e. $\bar{x} = \arg\min_{x\in\mathbb{R}^n} F(x) + \Lambda(x)$,
   i.e. $\bar{x}$ miminizes $F + \Lambda$ over the whole space (unconstrained minimization).

Then $F(\bar{x}) \leq F(x)$ for all $x \in C$, i.e. $\bar{x}$ minimizes F over C (constrained minimization), that is $\bar{x} = \arg\min_{x\in C} F(x)$.

Proof. For all $x \in C$,

$$\begin{aligned} F(\bar{x}) + \Lambda(\bar{x}) &\leq F(x) + \Lambda(x) \\ &\leq F(x) + \Lambda(\bar{x}), \quad \text{and hence} \\ F(\bar{x}) &\leq F(x). \end{aligned}$$

Curiously enough, the proof is shorter than the claim, and easier to understand. □

✎ **Tools from analysis: Lagrange multipliers**

A particular but very useful case of Lagrange's lemma is when Λ is constant over C: if this is the case, any point $\bar{x} \in C$ that satisfies just the *second* condition of the lemma miminizes F over C. A typical setup is the following:

- $x = (\theta, \lambda)$, where $\theta \in \mathbb{R}^p$, $\lambda \in \mathbb{R}^m$;
- $J : \mathbb{R}^p \to \mathbb{R}$ is a function to be minimized subject to constraints;
- $k : \mathbb{R}^p \to \mathbb{R}^m$ is a function designed to impose m constraints;
- $C := \{(\theta, \lambda) : k(\theta) = 0\}$;
- $F(x) := J(\theta)$ and $\Lambda(x) := \langle \lambda, k(\theta) \rangle$.

Clearly, $\Lambda(x) \equiv 0$ over C; thus, if a $\bar{\theta}$ satisfies $k(\bar{\theta}) = 0$ and attains the minimum in the *unconstrained* minimization problem

$$\min_{\theta \in \mathbb{R}^p} \quad J(\theta) + \langle \lambda, k(\theta) \rangle \tag{22}$$

then the same $\bar{\theta}$ solves the *constrained* minimization problem

$$\begin{aligned} \min_{\theta \in \mathbb{R}^p} \quad & J(\theta) \\ \text{subject to} \quad & k(\theta) = 0. \end{aligned} \tag{23}$$

Note that (22) is actually a *family* of minimization problems depending on λ, so that its solution $\bar{\theta} = \bar{\theta}(\lambda)$ is indeed a function of λ. Searching, among the values of this function, one $\bar{\theta}$ such that $k(\bar{\theta}) = 0$ amounts to search for a particular $\bar{\lambda}$: so, in a sense, the target of θ is to attain minimality, while the "dual" target of λ is to satisfy the constraint.

The function $J(\theta) + \langle \lambda, k(\theta) \rangle = J(\theta) + \lambda^\top k(\theta)$ is called Lagrangian, and λ is called a vector of Lagrangian multipliers. If J is convex and differentiable, then the search for a solution $\bar{\theta}$ of Problem (23) can proceed by equating to zero the derivatives of the Lagrangian both with respect to θ and with respect to λ.

To solve (21), we form the Lagrangian $\|Y - \Phi\theta\|^2 + \lambda^\top \theta$, compute gradients with respect to θ and λ, and set them equal to zero:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \left(\|Y - \Phi\theta\|^2 + \lambda^\top K\theta \right) \\ &= \frac{\partial}{\partial \theta} \left((Y - \Phi\theta)^\top (Y - \Phi\theta) + \lambda^\top K\theta \right) \\ &= \frac{\partial}{\partial \theta} \left(Y^\top Y - Y^\top \Phi\theta - \theta^\top \Phi^\top Y + \theta^\top \Phi^\top \Phi\theta + \lambda^\top K\theta \right) \\ &= 2\theta^\top \Phi^\top \Phi - 2Y^\top \Phi + \lambda^\top K, \quad \text{and transposing} \\ 0 &= \Phi^\top \Phi\theta - \Phi^\top Y + K^\top \frac{\lambda}{2}. \end{aligned} \tag{24}$$

Of course setting to zero the gradient with respect to λ we obtain the constraint $K\theta = 0$.

Multiplying (24) on the left by $(\Phi^\top\Phi)^{-1}$, we obtain

$$\begin{aligned}
\theta &= (\Phi^\top\Phi)^{-1}\Phi^\top Y - (\Phi^\top\Phi)^{-1}K^\top\frac{\lambda}{2},\\
K\theta &= K(\Phi^\top\Phi)^{-1}\Phi^\top Y - K(\Phi^\top\Phi)^{-1}K^\top\frac{\lambda}{2} = 0,\\
\frac{\lambda}{2} &= \left(K(\Phi^\top\Phi)^{-1}K^\top\right)^{-1} K(\Phi^\top\Phi)^{-1}\Phi^\top Y
\end{aligned}$$

(note that $K(\Phi^\top\Phi)^{-1}K^\top \in \mathbb{R}^{m\times m}$ must have full rank), and finally

$$\hat{\theta} = (\Phi^\top\Phi)^{-1}\Phi^\top Y - (\Phi^\top\Phi)^{-1}K^\top\left(K(\Phi^\top\Phi)^{-1}K^\top\right)^{-1} K(\Phi^\top\Phi)^{-1}\Phi^\top Y. \tag{25}$$

2 Statistical properties of the LS method

2.1 Estimators and desirable properties

2.2 Explicit model structure

2.3 Unbiasedness of the LS estimate

2.4 Consistency of the LS estimate

2.5 The Gauss-Markov theorem

Let Y, Φ, and $\boldsymbol{\varepsilon}$ be defined as in equation (11). In this section we will assume that the noise terms $\boldsymbol{\varepsilon}_i$ (and consequently the measures y_i) are random variables, while the regressors are deterministic. Random quantities will be written in boldface (for example $\mathbf{Y}$, $\boldsymbol{\varepsilon}$); the linear model (12) reads

$$\mathbf{Y} = \Phi\theta^\circ + \boldsymbol{\varepsilon}. \tag{26}$$

Suppose that $\boldsymbol{\varepsilon}_i$, $i = 1,\dots,N$ are uncorrelated, each with mean 0 and variance σ^2, so that $\mathsf{E}[\boldsymbol{\varepsilon}] = 0$ and let $\Sigma = \mathsf{Var}[\boldsymbol{\varepsilon}] = \sigma^2 I$. We search for the *best linear unbiased estimator* (BLUE for short) of θ° given $\mathbf{Y}$, that is an unbiased estimator $\hat{\boldsymbol{\theta}}$ of θ° having the form $\hat{\boldsymbol{\theta}} = L\mathbf{Y}$ for some $p\times N$ matrix L and whose variance is the minimum possible. Since $\mathsf{E}[L\mathbf{Y}] = \mathsf{E}[L\Phi\theta^\circ + L\boldsymbol{\varepsilon}] = L\Phi\theta^\circ$, asking for $\hat{\boldsymbol{\theta}}$ to be unbiased amounts to ask that L must satisfy the constraint $L\Phi = I$. The matrix $\left(\Phi^\top\Phi\right)^{-1}\Phi^\top$ does satisfy such constraint, hence L must have the form

$$L = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top + K, \tag{27}$$

where $K\Phi = 0$. Now

$$\begin{aligned}
\mathsf{Var}\left[\hat{\boldsymbol{\theta}}\right] &= \mathsf{Var}\left[L\left(\Phi\theta^\circ + \boldsymbol{\varepsilon}\right)\right] = \mathsf{Var}\left[\theta^\circ + L\boldsymbol{\varepsilon}\right] = \mathsf{E}\left[(L\boldsymbol{\varepsilon})(L\boldsymbol{\varepsilon})^\top\right] = L\left(\sigma^2 I\right)L^\top\\
&= \sigma^2\left(\left(\Phi^\top\Phi\right)^{-1}\Phi^\top + K\right)\left(\left(\Phi^\top\Phi\right)^{-1}\Phi^\top + K\right)^\top\\
&= \sigma^2\left(\Phi^\top\Phi\right)^{-1}\Phi^\top\Phi\left(\Phi^\top\Phi\right)^{-1}\\
&\quad + \sigma^2\left(\Phi^\top\Phi\right)^{-1}\Phi^\top K^\top + \sigma^2 K\Phi\left(\Phi^\top\Phi\right)^{-1} + \sigma^2 KK^\top\\
&= \sigma^2\left(\Phi^\top\Phi\right)^{-1} + \sigma^2\, KK^\top.
\end{aligned} \tag{28}$$

Since the term $\left(\Phi^\top\Phi\right)^{-1}$ does not depend on K, the variance is minimum, *in the matricial sense*, when $K = 0$. It is minimum also in the *scalar* sense, because

$$\mathsf{var}\left[\hat{\boldsymbol{\theta}}\right] = \mathrm{tr}\ \mathsf{Var}\left[\hat{\boldsymbol{\theta}}\right]$$

also attains its minimum for $K = 0$. Hence, the BLUE is the least squares estimator:

$$\hat{\boldsymbol{\theta}} = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\mathbf{Y}, \tag{29}$$

and attains the variance

$$\mathsf{Var}\left[\hat{\boldsymbol{\theta}}\right] = \sigma^2\left(\Phi^\top\Phi\right)^{-1}.$$

Suppose now that the noise terms $\boldsymbol{\varepsilon}_i$, $i = 1, \ldots, n$ of $\boldsymbol{\varepsilon}$ are *not* assumed to be either uncorrelated or with the same variance. Let however $\mathsf{E}\left[\boldsymbol{\varepsilon}\right] = 0$ and let $\Sigma = \mathsf{Var}\left[\boldsymbol{\varepsilon}\right] > 0$ denote the *known* covariance matrix of $\boldsymbol{\varepsilon}$; here $\Sigma_{ij} = \mathsf{Cov}\left[\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j\right]$, and in particular $\Sigma_{ii} = \mathsf{var}\left[\boldsymbol{\varepsilon}_i\right]$. Then we can substitute (27) with

$$L = \left(\Phi^\top\Sigma^{-1}\Phi\right)^{-1}\Phi^\top\Sigma^{-1} + K,$$

where $K\Phi = 0$, and repeat the computation (28) without substantial changes. We recover that the BLUE is

$$\hat{\boldsymbol{\theta}} = \left(\Phi^\top\Sigma^{-1}\Phi\right)^{-1}\Phi^\top\Sigma^{-1}\ \mathbf{Y}, \tag{30}$$

attaining the variance

$$\mathsf{Var}\left[\hat{\boldsymbol{\theta}}\right] = \left(\Phi^\top\Sigma^{-1}\Phi\right)^{-1}.$$

Particular case: if the noise terms $\boldsymbol{\varepsilon}_i$ have mean zero and are uncorrelated, but have *different* variances σ_i^2, then (30) is the solution (10) of a *weighted* least squares problem (8) where $w_i = \frac{1}{\sigma_i^2}$.

2.6 Gaussian case: the LS estimator attains maximum likelihood

[Not part of the DDSM course.]

Suppose that $\mathbf{Y} = \Phi\theta^\circ + \boldsymbol{\varepsilon}$, where Φ is deterministic and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$, i.e. $\boldsymbol{\varepsilon}$ is Gaussian with mean 0 and *known* covariance matrix Σ (assume $\Sigma > 0$); then $\mathbf{Y} \sim \mathcal{N}(\Phi\theta^\circ, \Sigma)$. The density function of a $\mathcal{N}(\Phi\theta, \Sigma)$ vector is

$$f_{\mathbf{Y}}(Y;\theta) = \frac{1}{\sqrt{(2\pi)^n \det\ \Sigma}}\exp\left(-\frac{1}{2}(Y - \Phi\theta)^\top\Sigma^{-1}(Y - \Phi\theta)\right);$$

therefore the likelihood and the log-likelihood of θ given the observation Y are respectively

$$L(\theta; Y) = \frac{1}{\sqrt{(2\pi)^n \det\ \Sigma}}\exp\left(-\frac{1}{2}(Y - \Phi\theta)^\top\Sigma^{-1}(Y - \Phi\theta)\right);$$
$$\ell(\theta; Y) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log\det\ \Sigma - \frac{1}{2}(Y - \Phi\theta)^\top\Sigma^{-1}(Y - \Phi\theta).$$

The maximum-likelihood estimator is obtained equating $\frac{\partial\ell(\theta;\mathbf{Y})}{\partial\theta}$ to zero, thus obtaining

$$\frac{\partial\ell(\theta;\mathbf{Y})}{\partial\theta} = \Phi^\top\Sigma^{-1}(\mathbf{Y} - \Phi\theta) = 0;$$
$$\hat{\boldsymbol{\theta}} = \left(\Phi^\top\Sigma^{-1}\Phi\right)^{-1}\Phi^\top\Sigma^{-1}\ \mathbf{Y}.$$

This is precisely the expression (29) of the least-squares estimator weighted with the inverse of the covariance matrix Σ. Hence, in particular, in the Gaussian case the maximum-likelihood estimator $\hat{\boldsymbol{\theta}}$ happens to be linear and unbiased; but now this is coincidental property, not a *requirement* like in Section 2.5.

2.7 Gaussian case: the LS estimator is efficient

[Not part of the DDSM course.]

Assume that the hypotheses of Section 2.6 hold. Fisher's score function is then

$$u(\theta; Y) = -\frac{\partial \ell(\theta; Y)}{\partial \theta} = \Phi^\top \Sigma^{-1}(\Phi\theta - Y);$$

Fisher's information matrix is defined by $\mathcal{I}(\theta) := \mathsf{E}\left[u(\theta; Y)u(\theta; Y)^\top; \theta\right]$; in our case,

$$\begin{aligned}\mathcal{I}(\theta) &= \mathsf{E}\left[\Phi^\top \Sigma^{-1}(\Phi\theta - Y)(\Phi\theta - Y)^\top \Sigma^{-1}\Phi; \theta\right] \\ &= \Phi^\top \Sigma^{-1} \mathsf{E}\left[(\Phi\theta - Y)(\Phi\theta - Y); \theta\right]\Sigma^{-1}\Phi \\ &= \Phi^\top \Sigma^{-1}\Sigma\Sigma^{-1}\Phi = \Phi^\top \Sigma^{-1}\Phi\end{aligned}$$

(note that here $\mathcal{I}(\theta)$ does not depend on θ); a famous result says that, given any *unbiased* estimator $\hat{\boldsymbol{\theta}}$ of θ°, the following inequality (Cramér-Rao lower bound) holds:

$$\mathsf{Var}\left[\hat{\boldsymbol{\theta}}\right] \geq \mathcal{I}(\theta)^{-1}. \tag{31}$$

In words, the right-hand side of (31) is a lower bound (in the matricial sense) for the variance of any unbiased estimator. If an estimator $\hat{\boldsymbol{\theta}}$ attains exactly the minimum possible variance, it is said to be *efficient*; and in the Gaussian case the LS estimator *does* attain the minimum possible variance, precisely because it attains the Cramér-Rao lower bound:

$$\mathsf{Var}\left[\hat{\boldsymbol{\theta}}\right] = \left(\Phi^\top \Sigma^{-1}\Phi\right)^{-1} = \mathcal{I}(\theta)^{-1}.$$

Resuming, in a linear model with Gaussian noise the maximum-likelihood estimator of the parameter is linear and unbiased, it is the LS estimator, it reaches the Cramér-Rao lower bound, and it has the minimum possible variance (among *all the unbiased estimators*, not only among the linear ones).

2.8 Residual variance

Assume the model $\mathbf{y}_i = \varphi_i^\top \theta^\circ + \boldsymbol{\varepsilon}_i$, where the noise terms $\boldsymbol{\varepsilon}_i$ are independent and identically distributed with mean zero and variance σ^2. In compact form it reads $\mathbf{Y} = \Phi\theta^\circ + \boldsymbol{\varepsilon}$, where $\mathsf{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right] = \sigma^2 I$. suppose that the regressors are deterministic and that Φ has full rank p. The following quantity, where $\hat{\theta}$ is the least squares solution, is needed for future computations:

$$\mathsf{E}\left[\sum_{i=1}^{N}\left(\mathbf{y}_i - \varphi_i^\top \hat{\boldsymbol{\theta}}\right)^2\right] = \mathsf{E}\left[\left\|\mathbf{Y} - \Phi\hat{\boldsymbol{\theta}}\right\|^2\right] \tag{32}$$

Recall that

$$\hat{\mathbf{Y}} = \Phi\hat{\boldsymbol{\theta}} = \left(\Phi\left(\Phi^\top\Phi\right)^{-1}\Phi^\top\right)\mathbf{Y} := \Pi_\Phi \mathbf{Y}.$$

The matrix $\Pi_\Phi = \Phi\left(\Phi^\top\Phi\right)^{-1}\Phi^\top \in \mathbb{R}^{N\times N}$ is the orthogonal projection matrix that projects on range Φ, and $(I-\Pi_\Phi)$ is the orthogonal projection matrix on $(\text{range } \Phi)^\perp$. It holds

$$\begin{aligned}\Pi_\Phi \mathbf{Y} &= \Pi_\Phi \Phi\theta^\circ + \Pi_\Phi \boldsymbol{\varepsilon} = \Phi\theta^\circ + \Pi_\Phi\boldsymbol{\varepsilon}\\ \mathbf{Y}-\hat{\mathbf{Y}} &= (I-\Pi_\Phi)\mathbf{Y} = \Phi\theta^\circ + \boldsymbol{\varepsilon} - \Phi\theta^\circ - \Pi_\Phi\boldsymbol{\varepsilon} = (I-\Pi_\Phi)\boldsymbol{\varepsilon}\end{aligned}$$

Then (32) becomes

$$\begin{aligned}\mathsf{E}\left[\left\|\mathbf{Y}-\Phi\hat{\boldsymbol{\theta}}\right\|^2\right] &= \mathsf{E}\left[\left\|\mathbf{Y}-\hat{\mathbf{Y}}\right\|^2\right] = \mathsf{E}\left[\left(\mathbf{Y}-\hat{\mathbf{Y}}\right)^\top\left(\mathbf{Y}-\hat{\mathbf{Y}}\right)\right]\\ &= \mathsf{E}\left[\boldsymbol{\varepsilon}^\top(I-\Pi_\Phi)^\top(I-\Pi_\Phi)\boldsymbol{\varepsilon}\right] = \mathsf{E}\left[\boldsymbol{\varepsilon}^\top(I-\Pi_\Phi)\boldsymbol{\varepsilon}\right]\\ &= \mathsf{E}\left[\text{tr } \boldsymbol{\varepsilon}^\top(I-\Pi_\Phi)\boldsymbol{\varepsilon}\right] = \text{tr } (I-\Pi_\Phi)\mathsf{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right] = \text{tr } (I-\Pi_\Phi)\ \sigma^2 I\\ &= \sigma^2(N-\text{tr } \Pi_\Phi)\end{aligned}$$

It remains to notice that

$$\text{tr } \Pi_\Phi = \text{tr } \Phi\left(\Phi^\top\Phi\right)^{-1}\Phi^\top = \text{tr } \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\Phi = \text{tr } I_p = p,$$

and we find the conclusion:

$$\mathsf{E}\left[\sum_{i=1}^N\left(\mathbf{y}_i - \varphi_i^\top\hat{\boldsymbol{\theta}}\right)^2\right] = \sigma^2(N-p). \tag{33}$$

Besides being interesting per se, (33) has an important consequence: the statistic

$$\hat{\boldsymbol{\sigma}}^2 = \frac{1}{N-p}\sum_{i=1}^N\left(\mathbf{y}_i - \varphi_i^\top\hat{\boldsymbol{\theta}}\right)^2 \tag{34}$$

is an unbiased estimator of the noise variance σ^2.

Example. Suppose that $\mathbf{y}_1,\dots,\mathbf{y}_N$ are independent and identically distributed random variables with mean μ and variance σ^2. We estimate μ letting $\varphi_i \equiv 1$ (here $p=1$) and applying the method of least squares to the model $\mathbf{y}_i = \varphi_i\mu + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i$ has mean 0 and variance σ^2. Solving the normal equations $(\sum_{i=1}^N 1\cdot 1)\hat{\boldsymbol{\mu}} = \sum_{i=1}^N 1\cdot\mathbf{y}_i$ we find that the least squares estimate is the sample average

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^N \mathbf{y}_i.$$

We already know that $\hat{\boldsymbol{\mu}}$ is an unbiased estimator of μ; the result that we have just proven allows us to add that the *corrected sample variance*

$$\hat{\boldsymbol{\sigma}}^2 = \frac{1}{N-1}\sum_{i=1}^N(\mathbf{y}_i - \hat{\boldsymbol{\mu}})^2$$

is an unbiased estimator of σ^2.

2.9 Bias-variance tradeoff and FPE

Theorems 1.2-1.3 can be adapted to the case of subspaces of the space $\mathcal{H}$ of square-summable random variables.

Theorem 2.1 *Let $\mathcal{W}$ be a closed subspace of the Hilbert space $\mathcal{H}$ of square-summable random variables, and let $\mathbf{Y} \in \mathcal{H}$. Then there exists a unique random variable $\hat{\mathbf{Y}} \in \mathcal{W}$ such that $\|\mathbf{Y} - \hat{\mathbf{Y}}\| \leq \|\mathbf{Y} - \mathbf{w}\|$ for all $\mathbf{w} \in \mathcal{W}$. A necessary and sufficient condition for $\hat{\mathbf{Y}}$ to be the unique minimizing random variable is that $\mathbf{Y} - \hat{\mathbf{Y}} \perp \mathbf{w}$ for all $\mathbf{w} \in \mathcal{W}$.*

Note: with the scalar product defined as $\langle \mathbf{w}, \mathbf{Y} \rangle = \mathsf{E}\left[\mathbf{w}\mathbf{Y}\right]$,

- the minimality claim translates to $\mathsf{E}\left[(\mathbf{Y} - \hat{\mathbf{Y}})^2\right] \leq \mathsf{E}\left[(\mathbf{Y} - \mathbf{w})^2\right]$ for all $\mathbf{w} \in \mathcal{W}$;
- the orthogonality condition reads $\mathsf{E}\left[(\mathbf{Y} - \hat{\mathbf{Y}})\mathbf{w}\right] = 0$ for all $\mathbf{w} \in \mathcal{W}$.

Assume the predictive model class

$$\hat{\mathcal{M}} = \{\hat{\mathbf{y}}_i(\theta) = \varphi(\mathbf{u}_i)^\top \theta \ : \ \theta \in \mathbb{R}^p\} \tag{35}$$

The *optimal predictor in* $\hat{\mathcal{M}}$ is the predictor $\hat{\mathbf{y}}_i(\theta^\circ) = \varphi(\mathbf{u}_i)^\top \theta^\circ$ corresponding to the parameter θ° that solves

$$\theta^\circ = \arg\min_{\theta \in \mathbb{R}^p} \bar{J}(\theta),$$

where $\bar{J}$ is the cost function defined as follows:

$$\bar{J}(\theta) = \mathsf{E}\left[(\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta))^2\right].$$

Under mild assumption on $\varphi(\cdot)$ ($\varphi(\mathbf{u}_i)$ must have finite second order moment), we let $\mathcal{W} = \hat{\mathcal{M}}$; in fact this is a closed subspace of the space of square-summable variables. Then Theorem 2.1 asserts that $\hat{\mathbf{y}}(\theta^\circ)$ exists and is unique, hence an optimal parameter θ° exists and is unique. The orthogonality condition says that

$$\mathbf{y}_i = \varphi(\mathbf{u}_i)^\top \theta^\circ + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{\varepsilon}_i$ is orthogonal (i.e. uncorrelated) to $\varphi(\mathbf{u}_i)$. Consequences:

1. $\mathsf{E}\left[(\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta^\circ))^2\right] = \mathsf{E}\left[\boldsymbol{\varepsilon}_i^2\right]$. For the sake of brevity, denote $\sigma_{\boldsymbol{\varepsilon}}^2 = \mathsf{E}\left[\boldsymbol{\varepsilon}_i^2\right]$, whether or not $\boldsymbol{\varepsilon}_i$ has mean 0.
2. If $\boldsymbol{\varepsilon}_i$ has mean 0 (this is the case, for example, if the regressor function $\varphi(\cdot)$ contains the component 1), then the least squares estimate $\hat{\boldsymbol{\theta}}$ is unbiased:

$$\mathsf{E}\left[\hat{\boldsymbol{\theta}}\right] = \mathsf{E}\left[\theta^\circ + (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{\varepsilon}\right] = \theta^\circ + \mathsf{E}\left[(\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top\right] \mathsf{E}\left[\boldsymbol{\varepsilon}\right] = \theta^\circ,$$

and its variance is

$$\begin{aligned}\mathsf{Var}\left[\hat{\boldsymbol{\theta}}\right] &= \sigma_{\boldsymbol{\varepsilon}}^2 \, \mathsf{E}\left[(\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}\right] = \sigma_{\boldsymbol{\varepsilon}}^2 \, \mathsf{E}\left[\left(\sum_{i=1}^N \varphi(\mathbf{u}_i)\varphi(\mathbf{u}_i)^\top\right)^{-1}\right] \\ &= \frac{\sigma_{\boldsymbol{\varepsilon}}^2}{N} \, \mathsf{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \varphi(\mathbf{u}_i)\varphi(\mathbf{u}_i)^\top\right)^{-1}\right],\end{aligned}$$

where $\sigma_{\boldsymbol{\varepsilon}}^2$ is the variance of $\boldsymbol{\varepsilon}_i$.

For a fixed θ,

$$
\begin{aligned}
\bar{J}(\theta) &= \mathsf{E}\left[(\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta))^2\right] \\
&= \mathsf{E}\left[(\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta^\circ) + \hat{\mathbf{y}}_i(\theta^\circ) - \hat{\mathbf{y}}_i(\theta))^2\right] \\
&= \mathsf{E}\left[(\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta^\circ))^2\right] + \mathsf{E}\left[(\hat{\mathbf{y}}_i(\theta^\circ) - \hat{\mathbf{y}}_i(\theta))^2\right] - 2\mathsf{E}\left[(\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta^\circ))(\hat{\mathbf{y}}_i(\theta^\circ) - \hat{\mathbf{y}}_i(\theta))\right] \\
&= \sigma_{\boldsymbol{\varepsilon}}^2 + \mathsf{E}\left[(\varphi(\mathbf{u}_i)^\top(\theta^\circ - \theta))^2\right] - 2\mathsf{E}\left[(\mathbf{y}_i - \varphi(\mathbf{u}_i)^\top\theta^\circ)\ \underbrace{\varphi(\mathbf{u}_i)^\top(\theta^\circ - \theta)}_{\mathbf{w}\in\mathcal{W}}\right] \\
&\quad \text{(the last term vanishes due to Theorem (2.1))} \\
&= \sigma_{\boldsymbol{\varepsilon}}^2 + \mathsf{E}\left[(\varphi(\mathbf{u}_i)^\top(\theta^\circ - \theta))^2\right] \\
&= \sigma_{\boldsymbol{\varepsilon}}^2 + (\theta^\circ - \theta)^\top \mathsf{E}\left[\varphi(\mathbf{u}_i)\varphi(\mathbf{u}_i)^\top\right](\theta^\circ - \theta) \\
&\quad \text{(denote } \mathsf{E}\left[\varphi(\mathbf{u}_i)\varphi(\mathbf{u}_i)^\top\right] = \Sigma) \\
&= \sigma_{\boldsymbol{\varepsilon}}^2 + (\theta^\circ - \theta)^\top \Sigma(\theta^\circ - \theta).
\end{aligned}
$$

If we plug in the least squares estimate $\hat{\boldsymbol{\theta}}$ we get

$$
\begin{aligned}
\bar{J}(\hat{\boldsymbol{\theta}}) &= \mathsf{E}\left[(\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta^\circ))^2\right] + (\hat{\boldsymbol{\theta}} - \theta^\circ)^\top \Sigma(\hat{\boldsymbol{\theta}} - \theta^\circ) \\
&= \sigma_{\boldsymbol{\varepsilon}}^2 + (\hat{\boldsymbol{\theta}} - \theta^\circ)^\top \Sigma(\hat{\boldsymbol{\theta}} - \theta^\circ),
\end{aligned}
$$

that is a random variable (depending on the observations $(\boldsymbol{\varphi}_i, \mathbf{y}_i)$, $i = 1, \ldots, N$). The *expected cost* is

$$
\begin{aligned}
\mathsf{E}\left[\bar{J}(\hat{\boldsymbol{\theta}})\right] &= \mathsf{E}\left[(\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta^\circ))^2\right] + \mathsf{E}\left[(\hat{\boldsymbol{\theta}} - \theta^\circ)^\top \Sigma(\hat{\boldsymbol{\theta}} - \theta^\circ)\right] \\
&= \sigma_{\boldsymbol{\varepsilon}}^2 + \mathsf{E}\left[(\hat{\boldsymbol{\theta}} - \theta^\circ)^\top \Sigma(\hat{\boldsymbol{\theta}} - \theta^\circ)\right] \\
&= \sigma_{\boldsymbol{\varepsilon}}^2 + \mathsf{E}\left[\mathrm{tr}\ (\hat{\boldsymbol{\theta}} - \theta^\circ)^\top \Sigma(\hat{\boldsymbol{\theta}} - \theta^\circ)\right] \\
&= \sigma_{\boldsymbol{\varepsilon}}^2 + \mathrm{tr}\ \mathsf{E}\left[(\hat{\boldsymbol{\theta}} - \theta^\circ)(\hat{\boldsymbol{\theta}} - \theta^\circ)^\top\right]\Sigma.
\end{aligned}
\tag{36}
$$

We can provide an approximation of this quantity. On one hand, for big N,

$$
\begin{aligned}
\mathsf{E}\left[\bar{J}(\hat{\boldsymbol{\theta}})\right] &= \sigma_{\boldsymbol{\varepsilon}}^2 + \frac{\sigma_{\boldsymbol{\varepsilon}}^2}{N}\ \mathrm{tr}\ \mathsf{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \varphi(\mathbf{u}_i)\varphi(\mathbf{u}_i)^\top\right)^{-1}\right]\Sigma \\
&\simeq \sigma_{\boldsymbol{\varepsilon}}^2 + \frac{\sigma_{\boldsymbol{\varepsilon}}^2}{N}\ \mathrm{tr}\ \Sigma^{-1}\Sigma = \sigma_{\boldsymbol{\varepsilon}}^2 + \frac{\sigma_{\boldsymbol{\varepsilon}}^2}{N}\ \mathrm{tr}\ I_p \\
&= \frac{N+p}{N}\ \sigma_{\boldsymbol{\varepsilon}}^2,
\end{aligned}
\tag{37}
$$

because we know that $\frac{1}{N}\sum_{i=1}^N \varphi(\mathbf{u}_i)\varphi(\mathbf{u}_i)^\top \to \Sigma$ almost surely for the strong law of large numbers; on the other hand, we know an unbiased estimate of $\sigma_{\boldsymbol{\varepsilon}}^2$:

$$
\hat{\sigma}_{\boldsymbol{\varepsilon}}^2 = \frac{1}{N-p}\sum_{i=1}^N \left(\mathbf{y}_i - \varphi(\mathbf{u}_i)^\top\hat{\boldsymbol{\theta}}\right)^2,
$$

and hence we consider

$$
\begin{aligned}
\mathsf{E}\left[\bar{J}(\hat{\boldsymbol{\theta}})\right] &\simeq \frac{N+p}{N-p}\ \left(\frac{1}{N}\sum_{i=1}^N \left(\mathbf{y}_i - \varphi(\mathbf{u}_i)^\top\hat{\boldsymbol{\theta}}\right)^2\right) \\
&= \frac{N+p}{N-p}\ \hat{J}(\hat{\boldsymbol{\theta}}),
\end{aligned}
\tag{38}
$$

which is a readily computable from data. The estimator (38) of the expected cost is called the *Final Prediction Error* (FPE).

Now suppose that the measures $\mathbf{y}_i$, $i = 1, \ldots, N$, are generated by the following model:

$$\mathbf{y}_i = f^*(\mathbf{u}_i) + \boldsymbol{\eta}_i,$$

where $\boldsymbol{\eta}_i$ are independent random variables with mean 0 and variance $\sigma^2_{\boldsymbol{\eta}}$, and $\boldsymbol{\eta}_i$ is independent of $\boldsymbol{\varphi}_i$. We do *not* assume that $f^*(\cdot)$, the "true" function, belongs to $\hat{\mathcal{M}}$ defined in (35), i.e. we do not assume that $f^*(\cdot)$ has the form $f^*(\cdot) = \varphi(\cdot)^\top \theta$ for any $\theta \in \mathbb{R}^p$; therefore the optimal predictor $\hat{\mathbf{y}}_i(\theta^\circ) = \varphi(\mathbf{u}_i)^\top \theta^\circ$ is not necessarily the best possible one, which is instead $\hat{\mathbf{y}}^*_i = f^*(\mathbf{u}_i)$. In this case the first term of (36) can be decomposed:

$$\begin{aligned}
\mathsf{E}\left[\bar{J}(\hat{\boldsymbol{\theta}})\right] &= \mathsf{E}\left[(\mathbf{y}_i - \hat{\mathbf{y}}_i(\theta^\circ))^2\right] + \operatorname{tr} \mathsf{Var}\left[\hat{\boldsymbol{\theta}}\right]\Sigma \\
&= \mathsf{E}\left[(\boldsymbol{\eta}_i + f^*(\mathbf{u}_i) - \hat{\mathbf{y}}_i(\theta^\circ))^2\right] + \operatorname{tr} \mathsf{Var}\left[\hat{\boldsymbol{\theta}}\right]\Sigma \\
&= \underbrace{\sigma^2_{\boldsymbol{\eta}}}_{\text{“noise”}} + \underbrace{\mathsf{E}\left[(f^*(\mathbf{u}_i) - \hat{\mathbf{y}}_i(\theta^\circ))^2\right]}_{\text{“bias}^2\text{”}} + \underbrace{\operatorname{tr} \mathsf{Var}\left[\hat{\boldsymbol{\theta}}\right]\Sigma}_{\text{“variance”}}
\end{aligned}$$

This is the so-called "bias-variance" decomposition. If $f^* \in \hat{\mathcal{M}}$, then the second term vanishes and the noise term $\sigma^2_{\boldsymbol{\eta}}$ coincides with $\sigma^2_{\boldsymbol{\varepsilon}}$. In any case, we can provide a similar approximation as in (37):

$$\mathsf{E}\left[\bar{J}(\hat{\boldsymbol{\theta}})\right] \simeq \frac{N+p}{N}\,\sigma^2_{\boldsymbol{\varepsilon}} = \frac{N+p}{N}\,\left(\sigma^2_{\boldsymbol{\eta}} + \mathsf{E}\left[(f^*(\mathbf{u}_i) - \hat{\mathbf{y}}_i(\theta^\circ))^2\right]\right).$$

2.10 Choice of model order

2.11 Regularization

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, New York, NY, USA, 2004.