Chapter 1

Note on the notation: Throughout, Professor Bahadur used the symbols $\varphi(s)$, $\varphi_1(s)$, $\varphi_2(s)$, ... to denote functions of the sample that are generally of little importance in the discussion of the likelihood. These functions often arise in his derivations without prior definition.

Lecture 1

Review of L^2 geometry

Let $(S, \mathcal{A}, P)$ be a probability space. We call two functions f_1 and f_2 on S EQUIVALENT if and only if $P(f_1 = f_2) = 1$, and set

$$V = L^2(S, \mathcal{A}, P) := \{f : f \text{ is measurable and } E(f^2) = \int_S f(s)^2 dP(s) < \infty\},$$

where we have identified equivalent functions. We may abbreviate $L^2(S, \mathcal{A}, P)$ to $L^2(P)$ or, if the probability space is understood, to just L^2. For $f, g \in V$, we define $||f|| = +\sqrt{E(f^2)}$ and $(f, g) = E(f \cdot g)$, so that $||f||^2 = (f, f)$. Throughout this list f and g denote arbitrary (collections of equivalent) functions in V.

1. V is a real vector space.

2. $(\cdot, \cdot)$ is an inner product on V – i.e., a bilinear, symmetric and positive definite function.

3. CAUCHY-SCHWARZ INEQUALITY:

$$|(f, g)| \leq ||f|| \cdot ||g||,$$

with equality if and only if f and g are linearly dependent.

Proof. Let x and y be real; then, by expanding $||\cdot||$ in terms of $(\cdot, \cdot)$, we find that

$$0 \leq ||xf + yg||^2 = x^2||f||^2 + 2xy(f, g) + y^2||g||^2,$$

from which the result follows immediately on letting $x = ||g||$ and $y = ||f||$. □

4. TRIANGLE INEQUALITY:

$$||f+g|| \leq ||f|| + ||g||.$$

Proof.

$$||f+g||^2 = \left| ||f||^2 + 2(f,g) + ||g||^2 \right| \leq ||f||^2 + 2||f||\,||g|| + ||g||^2,$$

again by expanding $||\cdot||$ in terms of $(\cdot,\cdot)$ and using the Cauchy-Schwarz inequality. □

5. PARALLELOGRAM LAW:

$$||f+g||^2 + ||f-g||^2 = 2(||f||^2 + ||g||^2).$$

Proof. Direct computation, as above. □

6. $||\cdot||$ is a continuous function on V, and $(\cdot,\cdot)$ is a continuous function on $V \times V$.

Proof. Suppose $f_n \xrightarrow{L^2} f$; then

$$(||f_n|| \leq ||f|| + ||f_n - f|| \to ||f||) \Rightarrow (\overline{\lim}\,||f_n|| \leq ||f||)$$

and

$$(||f|| \leq ||f_n|| + ||f_n - f||) \Rightarrow (\underline{\lim}\,||f_n|| \geq ||f||).$$

From these two statements it follows that $\lim ||f_n|| = ||f||$. □

7. V is a complete metric space under $||\cdot||$ – i.e., if $\{g_n\}$ is a sequence in V and $||g_n - g_m|| \to 0$ as $n, m \to \infty$, then $\exists \gamma \in V$ such that $||g_n - \gamma|| \to 0$.

Proof. The proof proceeds in four parts.

1. $\{g_n\}$ is a Cauchy sequence in probability:

$$P(|g_m - g_n| > \varepsilon) = P(|g_m - g_n|^2 > \varepsilon^2) \leq \frac{1}{\varepsilon^2} E(||g_m - g_n||^2) = \frac{1}{\varepsilon^2}||g_m - g_n||^2.$$

2. Hence there exists a subsequence $\{g_{n_k}\}$ converging a.e.(P) to, say, g.
3. $g \in V$.

Proof.

$$E(|g|^2) = \int \left(\lim_{k\to\infty} g_{n_k}^2\right) dP \leq \underline{\lim} \int g_{n_k}^2\, dP$$

by Fatou's lemma; but $\left\{\int g_{n_k}^2 dP = ||g_{n_k}||^2\right\}$ is a bounded sequence, since $\{||g_n||\}$ is Cauchy. □

4. $||g_n - g|| \to 0$.

Proof. For any $\varepsilon > 0$, choose $k = k(\varepsilon)$ so that $||g_m - g_n|| < \varepsilon$ whenever $m, n \geq k(\varepsilon)$. Then

$$\int |g_n - g|^2 dP = \int \big(\lim_{k\to\infty} |g_n - g_{n_k}|^2\big) dP$$

$$\overset{\text{Fatou}}{\leq} \lim_{k\to\infty} \int |g_n - g_{n_k}|^2 dP = \lim_{k\to\infty} ||g_n - g_{n_k}||^2 < \varepsilon,$$

provided that $n > k(\varepsilon)$. □

Let W be a subset of V. If W is closed under addition and scalar multiplication, then it is called a LINEAR MANIFOLD in V. If, furthermore, W is topologically closed, then it is called a SUBSPACE of V. Note that a finite-dimensional linear manifold must be topologically closed (hence a subspace).

If C is any collection of vectors in V, then let C_1 be the collection of all finite linear combinations of vectors in C and C_2 be the closure of C_1. Then C_2 is the smallest subspace of V containing C, and is called the subspace SPANNED by C. C_1 is called the linear manifold spanned by C.

Let W be a fixed subspace of V, and f a fixed vector in V. We say that the vector $g \in W$ is an ORTHOGONAL PROJECTION of f to W if and only if

$$||f - g|| = \inf_{h \in W} ||f - h||.$$

8. There exists a unique orthogonal projection g of f to W.

Proof. Let $\ell = \inf_{h\in W} ||f - h||$, and let $\{g_n\}$ be a sequence in W such that $||f - g_n|| \to \ell$; then we have

$$\left|\left|\frac{g_m - g_n}{2}\right|\right|^2 + \underbrace{\left|\left|\frac{g_m + g_n}{2} - f\right|\right|^2}_{\geq \ell} = \frac{1}{2}\underbrace{||g_m - f||^2}_{\text{converges to } \ell} + \frac{1}{2}\underbrace{||g_n - f||^2}_{\text{converges to } \ell},$$

from which we see that $||g_m - g_n||^2 \to 0$ as $m, n \to \infty$. Thus $\{g_n\}$ is a Cauchy sequence; but this means that there is some g such that $g_n \to g$. Since W is a subspace of V, it is closed; so, since each $g_n \in W$, so too is $g \in W$. □

Lecture 2

Definition. For two vectors $f_1, f_2 \in V$, we say that f_1 is ORTHOGONAL to f_2, and write $f_1 \perp f_2$, if and only if $(f_1, f_2) = 0$.

Throughout, we fix a subspace W of V and vectors $f, f_1, f_2 \in V$.

9. PYTHAGOREAN THEOREM (and its converse):

$$f_1 \perp f_2 \Leftrightarrow ||f_1 + f_2||^2 = ||f_1||^2 + ||f_2||^2.$$

10. a. Given the above definition of orthogonality, there are two natural notions of orthogonal projection:

(*) $\gamma \in W$ is an orthogonal projection of f on W if and only if

$$||f - \gamma|| = \inf_{g \in W} ||f - g||.$$

(**) $\gamma \in W$ is an orthogonal projection of f on W if and only if

$$(f - \gamma) \perp g \ \forall g \in W.$$

These two definitions are equivalent (i.e., γ satisfies (*) if and only if it satisfies (**)).

b. Exactly one vector $\gamma \in W$ satisfies (**) – i.e., a solution of the minimisation problem exists and is unique.

c. $||f||^2 = ||\gamma||^2 + ||f - \gamma||^2$.

Proof of (10).

a. ($\Rightarrow$) Choose $h \in W$. For all real x, $\gamma + xh \in W$ also. Therefore, if (*) holds, then (setting $\delta = f - \gamma$)

$$\begin{aligned}(||f - (\gamma + xh)||^2 \geq ||f - \gamma||^2 &\Rightarrow ||\delta||^2 - 2x(\delta, h) + x^2||h||^2 \geq ||\delta||^2 \\ &\Rightarrow x^2||h||^2 - 2x(\delta, h) \geq 0) \ \forall x \in \mathbb{R}.\end{aligned}$$

This is possible only if $(\delta, h) = 0$. Thus (**) holds.

($\Leftarrow$) If (**) holds then we have

$$\begin{aligned}((f - \gamma) \perp (\gamma - h) &\overset{(9)}{\Rightarrow} ||f - h||^2 = ||f - \gamma||^2 + ||\gamma - h||^2 \\ &\Rightarrow ||f - h||^2 \geq ||f - \gamma||^2) \ \forall h \in W\end{aligned}$$

Thus (*) holds.

b. Suppose that both γ_1 and γ_2 are solutions to (**) in W. Since $\gamma_1 - \gamma_2 \in W$, $(f - \gamma_1) \perp (\gamma_1 - \gamma_2)$ and hence, by (9),

$$||f - \gamma_2||^2 = ||f - \gamma_1||^2 + ||\gamma_1 - \gamma_2||^2.$$

By (a), however, γ_1 and γ_2 both also satisfy (*), so

$$||f - \gamma_1||^2 = \min_{g \in W} ||f - g||^2 = ||f - \gamma_2||^2$$

and hence $||\gamma_1 - \gamma_2||^2 = 0 \Rightarrow \gamma_1 = \gamma_2$.

c. Since $\gamma \in W$,

$$(f-\gamma) \perp \gamma \overset{(9)}{\Rightarrow} ||f||^2 = ||f-\gamma||^2 + ||\gamma||^2$$

as desired. $\square$

Definition. We denote by $\pi_W f$ the orthogonal projection of f on W.

Note. $||\pi_W f|| \leq ||f||$, with equality iff $\pi_W f = f$ – i.e., iff $f \in W$. (For, by 10(c), $||f||^2 = ||\pi_W f||^2 + ||f - \pi_W f||^2$.)

It's easy to see that

$$W = \{f \in V : \pi_W f = f\} = \{\pi_W f : f \in V\}.$$

Definition. The ORTHOGONAL COMPLEMENT of W in V is defined to be

$$W^\perp := \{h \in V : h \perp g \; \forall g \in W\}.$$

Note that $W^\perp = \{h \in V : \pi_W h = 0\}$.

11. $W^\perp$ is a subspace of V.

12. $\pi_W : V \to V$ is linear, idempotent and self-adjoint.

    Proof. We abbreviate π_W to π. Let $a_1, a_2 \in \mathbb{R}$ and $f, f_1, f_2 \in V$ be arbitrary. Then we have by (10) that $f_1 - \pi f_1$ and $f_2 - \pi f_2$ are in $W^\perp$ and hence by (11) that

    $$(a_1 f_1 + a_2 f_2) - (a_1 \pi f_1 + a_2 \pi f_2) = a_1(f_1 - \pi f_1) + a_2(f_2 - \pi f_2) \in W^\perp \qquad (*)$$

    Since $\pi f_1, \pi f_2 \in W$ and W is a subspace, $a_1 \pi f_1 + a_2 \pi f_2 \in W$; therefore, by (10) and (*) above, $\pi(a_1 f_1 + a_2 f_1) = a_1 \pi f_1 + a_2 \pi f_2$. Thus π is linear. We also have by (10) that $\pi(\pi f) = \pi f$, since $\pi f \in W$; thus π is idempotent.

    Finally, since $\pi f_1, \pi f_2 \in W$, once more by (10) we have that $(f_1 - \pi f_1, \pi f_2) = 0$; thus

    $$\begin{aligned}(f_1, \pi f_2) = \big(f_1 + (\pi f_1 - \pi f_1), \pi f_2\big) &= \big((f_1 - \pi f_1) + \pi f_1, \pi f_2\big) \\ &= (f_1 - \pi f_1, \pi f_2) + (\pi f_1, \pi f_2) = (\pi f_1, \pi f_2).\end{aligned}$$

    Similarly, $(\pi f_1, f_2) = (\pi f_1, \pi f_2)$, so that $(f_1, \pi f_2) = (\pi f_1, f_2)$. Thus π is self-adjoint. $\square$

13. We have from the above description of π_W that $W^\perp = \{f - \pi_W f : f \in V\}$.

14. (This is a converse to (12).) If $U : V \to V$ is linear, idempotent and self-adjoint, then U is an orthogonal projection to some subspace (i.e., there is a subspace W' of V so that $U = \pi_{W'}$).

15. Given an arbitrary $f \in V$, we may write uniquely $f = g + h$, with $g \in W$ and $h \in W^\perp$. In fact, $g = \pi_W f$ and $h = \pi_{W^\perp} f$. From this we conclude that $\pi_{W^\perp} \circ \pi_W \equiv 0 \equiv \pi_W \circ \pi_{W^\perp}$ and $(W^\perp)^\perp = W$.

16. Suppose that W_1 and W_2 are two subspaces of V such that $W_2 \subseteq W_1$. Then $\pi_{W_2} f = \pi_{W_2}(\pi_{W_1} f)$ and $||\pi_{W_2} f|| \leq ||\pi_{W_1} f||$, with equality iff $\pi_{W_1} f \in W_2$.

Lecture 3

Note. The above concepts and statements (regarding projections etc.) are valid in any Hilbert space, but we are particularly interested in the case $V = L^2(S, \mathcal{A}, P)$.

Note. If V is a Hilbert space and W is a subspace of V, then W is a Hilbert space when equipped with the same inner product as V.

Homework 1

1. If $V = L^2(S, \mathcal{A}, P)$, show that V is finite-dimensional if P is concentrated on a finite number of points in S. You may assume that the one-point sets $\{s\}$ are measurable.

2. Suppose that $S = [0,1]$, $\mathcal{A}$ is the Borel field (on $[0,1]$) and P is the uniform probability measure. Let $V = L^2$ and, for I, J fixed disjoint subintervals of S, define

$$W = W_{I,J} := \{f \in V : f = 0 \text{ a.e. on } I \text{ and } f \text{ is constant a.e. on } J\}.$$

   Show that W is a subspace and find $W^\perp$. Also compute $\pi_W f$ for $f \in V$ arbitrary.

3. Let $S = \mathbb{R}^1$, $\mathcal{A} = \mathcal{B}^1$ and P be arbitrary, and set $V = L^2$. Suppose that $s \in V$ is such that $E(e^{ts}) < \infty$ for all t sufficiently small (i.e., for all t in a neighbourhood of 0). Show that the subspace spanned by $\{1, s, s^2, \ldots\}$ is equal to V. (HINT: Check first that the hypothesis implies that $1, s, s^2, \ldots$ are indeed in V. Then check that, if $g \in V$ satisfies $g \perp s^2$ for $r = 0, 1, 2, \ldots$, then $g = 0$ a.e.(P). This may be done by using the uniqueness of the moment-generating function.)

Definition. Let $S = \{s\}$ and $V = L^2(S, \mathcal{A}, P)$. Let $(R, \mathcal{C})$ be a measurable space, and let $F : S \to R$ be a measurable function. If we let $Q = P \circ F^{-1}$ (so that $Q(T) = P(F^{-1}[T])$), then $F(s)$ is called a STATISTIC with corresponding probability space $(R, \mathcal{C}, Q)$. $W = L^2(R, \mathcal{C}, Q)$ is isomorphic to the subspace $\tilde{W} = L^2(S, F^{-1}[\mathcal{C}], P)$ of V.

Application to prediction

Let $S = \mathbb{R}^{k+1}$, $\mathcal{A} = \mathcal{B}^{k+1}$ be the Borel field in $\mathbb{R}^{k+1}$, P be arbitrary and $V = L^2$. Let $s = (X_1, \dots, X_k; Y)$.

A PREDICTOR of Y is a Borel function $G = G(\underline{X})$ of $\underline{X} = (X_1, \dots, X_k)$. We assume that $E(Y^2) < \infty$ and take the MSE of G, i.e., $E(|G(\underline{X}) - Y|^2)$, as a criterion. What should we mean by saying that G is the "best" predictor of Y?

i. No restriction on G: Consider the set W of all measurable $G = G(\underline{X})$ with $E(|G|^2) < \infty$. W is clearly (isomorphic to) a subspace of V and, for $G \in W$, $E(G - Y)^2 = ||Y - G||^2$.

   Then the *best* predictor of Y is just the orthogonal projection of Y on W, which is the same as the conditional expectation of Y given $\underline{X} = (X_1, \dots, X_k)$.

   Proof (informal). Let $G^*(\underline{X}) = E(Y \mid \underline{X})$. For an arbitrary $G = G(\underline{X}) \in W$,

   $$||Y - G||^2 = ||Y - G^*||^2 + ||G - G^*||^2 + 2(Y - G^*, G^* - G),$$

   but

   $$\begin{aligned}(Y - G^*, G^* - G) &= E\big((Y - G^*)(G^* - G)\big) \\ &= E\big[E((Y - G^*)(G^* - G) \mid \underline{X})\big] \\ &= E\big[(G^* - G)E(Y - G^* \mid \underline{X})\big] = 0,\end{aligned}$$

   so that $||Y - G||^2 = ||Y - G^*||^2 + ||G - G^*||^2$, whence G^* must be the unique projection.

ii. G an affine function: We require that G be an affine function of $\underline{X}$ – i.e., that there be constants $a_0, a_1, \dots, a_k$ such that $G(\underline{X}) = G(X_1, \dots, X_k) = a_0 + \sum_{i=1}^k a_i X_i$ for all $\underline{X}$. The class of such G is a subspace W' of the space W defined in the previous case. The best predictor of Y in this class is the orthogonal projection of Y on W', which is called the LINEAR REGRESSION of Y on $(X_1, \dots, X_k)$.

Lecture 4

We return to predicting Y using an affine function of $\underline{X}$. We define

$$W := \text{Span}\{1, X_1, \dots, X_k\}$$

and denote by $\hat{Y}$ the orthogonal projection of Y on W. $\hat{Y}$ is characterized by the two facts that

(*) $Y - \hat{Y} \perp 1$, and

(**) $Y - \hat{Y} \perp X_i^0$ for $i = 1, \ldots, k$

where $X_i^0 = X_i - EX_i$. Since $W = \mathrm{Span}\{1, X_1, \ldots, X_k\}$, we may suppose that $\hat{Y} = \beta_0 + \sum_{i=1}^k \beta_i X_i^0$. From (*), $\beta_0 = EY$; and, from (**), $\Sigma\beta = \mathfrak{c}$ (the 'normal equation'), where $\beta = (\beta_1, \ldots, \beta_k)^T$, $\mathfrak{c} = (c_1, \ldots, c_k)^T$, $\Sigma = (\sigma_{ij})$, $c_i = E(Y^0 X_i^0) = \mathrm{Cov}(X_i, Y)$, $\sigma_{ij} = E(X_i^0 X_j^0) = \mathrm{Cov}(X_i, X_j)$ and $Y^0 = Y - EY$. We have (by considering the minimization problem) that there exists a solution β to these two equations; and (by uniqueness of the orthogonal projection) that, if β is any such solution, then $\hat{Y} = \beta_0 + \sum_{i=1}^k \beta_i X_i^0$. Σ is positive semi-definite and symmetric.

Homework 1

4. Show that Σ is nonsingular iff, whenever $P(a_1 X_1^0 + \cdots + a_k X_k^0 = 0) = 1$, $a_1 = \cdots = a_k = 0$; and that this is true iff, whenever $P(b_0 + b_1 X_1 + \cdots + b_k X_k = 0) = 1$, $b_0 = b_1 = \cdots = b_k = 0$.

Let us assume that Σ is nonsingular; then $\beta = \Sigma^{-1}\mathfrak{c}$ and $\hat{Y} = EY + \sum_{i=1}^k \beta_i X_i^0$.

Note.

i. $\hat{Y}$ is called the LINEAR REGRESSION of Y on $(X_1, \ldots, X_k)$, or the AFFINE REGRESSION or the LINEAR REGRESSION of Y on $(1, X_1, \ldots, X_k)$.

ii. $\hat{Y}^0 = \sum_{i=1}^k \beta_i X_i^0$ is the projection of Y^0 on $\mathrm{Span}\{X_1^0, \ldots, X_k^0\}$. Thus

$$\mathrm{Var}\, Y = ||Y^0||^2 = ||Y^0 - \hat{Y}^0||^2 + ||\hat{Y}^0||^2 = \mathrm{Var}(Y - \hat{Y}) + \mathrm{Var}\, \hat{Y}$$

or, more suggestively, Var(predictand) = Var(residual) + Var(regression).

A related problem concerns

$$R := \sup_{a_1, \ldots, a_k} \mathrm{Corr}(Y, a_1 X_1 + \cdots + a_k X_k) = ?$$

We have that

$$\mathrm{Corr}\Big(Y, \sum a_i X_i\Big) = \mathrm{Corr}\Big(Y^0, \sum a_i X_i^0\Big) = \frac{1}{||Y^0||\,||L||}\mathrm{Cov}(Y^0, L)$$
$$= \frac{1}{||Y^0||\,||L||}(Y^0, L) = \frac{1}{||Y^0||}\left(Y^0, \frac{L}{||L||}\right),$$

where $L = \sum a_i X_i^0$. Since $Y^0 = (Y^0 - \hat{Y}^0) + \hat{Y}^0$,

$$\left(Y^0, \frac{L}{||L||}\right) = \left(\hat{Y}^0, \frac{L}{||L||}\right) \le ||\hat{Y}^0||$$

with equality iff $\frac{L}{||L||} = d\hat{Y}^0$ for some $d > 0$ (we have used the Cauchy-Schwarz inequality). In particular, $c(\beta_1, \ldots, \beta_k)$ (with c a positive constant) are the maximizing choices of $(a_1, \ldots, a_k)$. Plugging in any one of these maximizing choices gives us that $R = \frac{||\hat{Y}^0||}{||Y^0||}$ and hence that $R^2 = \frac{\operatorname{Var}\hat{Y}}{\operatorname{Var}Y}$, from which we conclude that

$$(1 - R^2)\operatorname{Var}Y = \operatorname{Var}(Y - \hat{Y}).$$

From the above discussion we see that Hilbert spaces are related to regression, and hence to statistics.

Note. Suppose that $k = 1$, and that we have data

Serial #	
1	(x_1, y_1)
2	(x_2, y_2)
$\vdots$	$\vdots$
n	(x_n, y_n).

We may then let S be the set consisting of the points $(1; x_1, y_1), \ldots, (n; x_n, y_n)$, to each of which we assign probability $1/n$. If we define $X(i, x_i, y_i) = x_i$ and $Y(i, x_i, y_i) = y_i$ for $i = 1, 2, \ldots, n$, then $EX = \overline{x}$ and $EY = \overline{y}$. $\hat{Y}$ is the affine regression of y on x and R is the correlation between x and y, which is

$$\frac{1}{S_x S_y}\left[\left(\sum x_i y_i\right) - n\overline{x}\,\overline{y}\right].$$

This extends also to the case $k > 1$.

Lecture 5

Classical estimation problem for inference

In the following, S is a sample space, with sample point s; $\mathcal{A}$ is a σ-field on S; and $\mathcal{P}$ is a set of probability measures P on $\mathcal{A}$, indexed by a set $\Theta = \{\theta\}$. We call Θ the PARAMETER SPACE. (The distinction between probability and statistics is that, in probability, Θ has only one element, whereas, in statistics, Θ is richer.)

Suppose we are given a function $g : \Theta \to \Theta$ and a sample point $s \in S$. We are interested in estimating the actual value of g using s, and describing its quality.

Example 1. Estimate $g(\theta)$ from iid $X_i = \theta + e_i$, where the e_i are iid with distribution symmetric around 0. We let $S = \{X_1, \ldots, X_n\}$ and $\Theta = (-\infty, \infty)$, and define g by $g(\theta) = \theta$ for all $\theta \in \Theta$. We might have:

a. X_is iid $N(\theta, 1)$.

b. X_is iid double exponential with density $\frac{1}{2}e^{-|x-\theta|}$ (for $-\infty < x < \infty$), with respect to Lebesgue measure.

c. X_is iid Cauchy, with density $\frac{1}{\pi(1+(x-\theta)^2)}$.

Possible estimates are $t_1(s) = \overline{X}$, $t_2(s) = \text{median}\{X_1, \ldots, X_n\}$ and

$$t_3(s) = 10\% \text{ of the trimmed mean in } \{X_1, \ldots, X_n\};$$

there are many others.

In the general case, $(S, \mathcal{A}, P_\theta)$, $\theta \in \Theta$, an ESTIMATE (of $g(\theta)$) is a measurable function t on S such that

$$E_\theta(t^2) = \int_S t(s)^2 dP_\theta(s) < \infty \; \forall \theta \in \Theta.$$

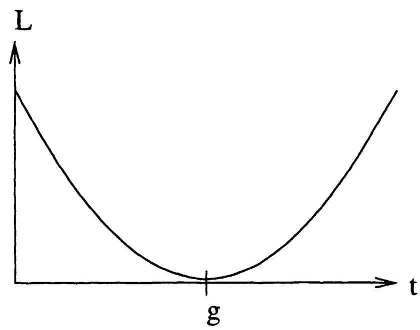
What is a "good" estimate?

Suppose that the loss involved in estimating $g(\theta)$ to be t when it is actually g is $L(t, g)$. (Some important choices of loss functions are $L(t, g) = |t - g|$ – the absolute error – and $L(t, g) = |t - g|^2$ – the square error.) Then the EXPECTED LOSS for a particular estimate t (and $\theta \in \Theta$) is

$$R_t(\theta) = E_\theta\big(L(t(s), g(\theta))\big).$$

R_t is called the RISK FUNCTION for t. For t to be a "good" estimate, we want R_t "small".

We consider now a heuristic for the square error function:



Assume that $L \geq 0$ and that, for each g, $L(g, g) = 0$ and $L(\cdot, g)$ is a smooth function of t. Then

$$L(t, g) = 0 + (t - g)\frac{\partial}{\partial t}L(t, g)\bigg|_g + \frac{1}{2}a(g)(t - g)^2 + \cdots = \frac{1}{2}a(g)(t - g)^2 + \cdots$$

where $a(g) \geq 0$. Let us assume that in fact $a(g) > 0$; then we define

$$R_t(\theta) := \frac{1}{2}a(g)E_\theta\big(t(s) - g(\theta)\big)^2,$$

so that R_t is locally proportional to $E_\theta(t - g)^2$, the MSE in t at θ.

Assume henceforth that $R_t(\theta) = E_\theta(t - g)^2$ and denote by $b_t(\theta) = E_\theta(t) - g(\theta)$ the 'bias' of t at θ.

1. $R_t(\theta) = \mathrm{Var}_\theta(t) + \left[b_t(\theta)\right]^2$.

Note. It is possible to regard $P_\theta(|t(s) - g(\theta)| > \varepsilon)$ (for $\varepsilon > 0$ small) – i.e., the distribution of t – as a criterion for how "good" the estimate t is. Now, for $Z \geq 0$, $EZ = \int_0^\infty P(Z \geq z)dz$; hence

$$R_t(\theta) = \int_0^\infty P_\theta\left(|t(s) - g(\theta)| > \sqrt{z}\right)dz.$$

There are several approaches to making R_t small. Three of them are:

ADMISSIBILITY: The estimate t is INADMISSIBLE if there is some estimate t' such that $R_{t'}(\theta) \leq R_t(\theta)$ for all $\theta \in \Theta$, and the inequality is strict for at least one θ. t_0 is admissible if it is not inadmissible. (This may be called the "sure-thing principle".)

MINIMAXITY: The estimate t_0 is MINIMAX if

$$\sup_{\theta\in\Theta} R_{t_0}(\theta) \leq \sup_{\theta\in\Theta} R_t(\theta)$$

for all estimates t.

BAYES ESTIMATION: Let λ be a probability on Θ and let $\overline{R}_t = \int_\Theta R_t(\theta)d\lambda$ be the average risk with respect to λ. The estimate t^* is then BAYES (with respect to λ) if $\overline{R}_{t^*} = \inf_t \overline{R}_t$.

2. If t^* has constant risk, i.e., $R_{t^*}(\theta) = c$ for all $\theta \in \Theta$, and t^* is Bayes with respect to some probability λ on Θ, then t^* is minimax.

   Proof. Let t be arbitrary; then

   $$c = \sup_\theta R_{t^*}(\theta) = \overline{R}_{t^*} \leq \overline{R}_t \leq \sup_\theta R_t(\theta).$$

   □

3. If t^* is the essentially unique Bayes estimate with respect to a probability λ on Θ, then t^* is admissible.

   Proof. Suppose that t is such that $R_t(\theta) \leq R_{t^*}(\theta)$ for all $\theta \in \Theta$; then $\overline{R}_t \leq \overline{R}_{t^*}$. Hence, by the definition of essential uniqueness,

   $$P_\theta(t^* = t) = 1 \ \forall\theta \in \Theta;$$

   it follows that $R_{t^*}(\theta) = R_t(\theta)$ for all $\theta \in \Theta$. □

Another approach to making R_t small is:

UNBIASEDNESS: We require all estimates t to be unbiased – i.e., $E_\theta(t) = g(\theta) \Leftrightarrow b_t(\theta) = 0$ for all $\theta \in \Theta$.

Several questions arise:

i. Are there any unbiased estimates at all?

ii. If so, which t, if any, has minimum variance at a given θ? (We call such a t a LOCALLY MINIMUM-VARIANCE UNBIASED ESTIMATE.)

iii. If there is a locally minimum variance unbiased estimate, is it independent of θ? (If so, then it is the uniformly minimum-variance unbiased estimate. If this estimate exists, what is it?)

There are two approaches: (I) general; and (II) sufficiency (i.e., via complete sufficient statistics).