Chapter 4

Hellinger differentiability

Modern statistical theory makes clever use of the fact that square roots of probability density functions correspond to unit vectors in spaces of square integrable functions. The Hellinger distance between densities corresponds to the $\mathcal{L}^2$ norm of the difference between the unit vectors. This Chapter explains some of the statistical consequences of differentiability in norm of the square root of the density, a property known as Hellinger differentiability.

Section 1 relates Hellinger differentiability to the classical regularity conditions for maximum likelihhod theory.

Section 2 derives some subtle consequences of norm differentiability for unit vectors.

Section 3 discusses connections between Hellinger differentiability and pointwise differentiability of densities, leading to a sufficient condition for Hellinger differentiability.

Section 4 derives the information inequality, as an illustration of the elegance brought into statistical theory by Hellinger differentiability.

Section 5 shows that Hellinger differentiability of marginal densities implies existence of a local quadratic approximation to the likelihood ratios for product measures. Apart from a few minor differences, the approximation implies the property known as local asymptotic normality.

Throughout the Chapter, $\mathcal{F} = \{ f_\theta(x) : \theta \in \Theta\}$ denotes a family of probability densities with respect to a measure λ, defining probability measures $\{P_\theta\}$ or $\{\mathbb{P}_\theta\}$, on a fixed space $\mathcal{X}$. Sometimes the parameter is written as a subscript, $f_\theta(x)$, and sometimes θ is replaced by another letter. The index set is usually a subset of Euclidean space $\mathbb{R}^k$, in most cases one-dimensional. The function $\xi_\theta(x)$, or $\xi(x,\theta)$, always denotes the positive square root of $f_\theta(x)$. The $\mathcal{L}^2(\lambda)$ norm is denoted by $\| \cdot \|_2$.

[§heuristics]

1. Heuristics

The traditional regularity conditions for maximum likelihood theory involve existence of two or three derivatives of density functions, together with domination assumptions to justify differentiation under integral signs. Le Cam (1970) noted that such conditions are unnecessarily stringent. He commented:

> Even if one is not interested in the maximum economy of assumptions one cannot escape practical statistical problems in which apparently "slight" violations of the assumptions occur. For instance the derivatives fail to exist at one point x which may depend on θ, or the distributions may not be mutually absolutely continuous or a variety of other difficulties may occur. The existing literature is rather unclear about what may happen in these circumstances. Note also that since the

conditions are imposed upon probability densities they may be satisfied for one choice of such densities but not for certain other choices.

Probably Le Cam had in mind examples such as the double exponential density, $\frac{1}{2}\exp(-|x-\theta|)$, for which differentiability fails at the point $\theta = x$. He showed that the traditional conditions can, for some purposes, be replaced by a simpler assumption of ***Hellinger differentiability***: differentiability in norm of the square root of the density as an element of an $\mathcal{L}^2$ space.

As you will see in Part III, much asymptotic theory can be made to work with classical regularity assumptions relaxed to assumptions of Hellinger differentiability. The derivation of the information inequality in Section 5 illustrates the point.

norm.diff <1> **Definition.** *A map τ from a subset Θ of a Euclidean space $\mathbb{R}^k$ into a normed vector space $\mathcal{V}$ is said to be differentiable (in norm) at a point θ_0 with derivative Δ if $\tau(\theta) = \tau(\theta_0) + (\theta-\theta_0)'\Delta + r(\theta)$, where $\|r(\theta)\| = o(|\theta-\theta_0|)$ as $\theta \to \theta_0$. The derivative Δ is a k-vector $(\Delta_1, \ldots, \Delta_k)$ of elements from $\mathcal{V}$, and $t'\Delta = \sum_i t_i \Delta_i$.*

Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a family of Hellinger differentiability of a family of probability densities $\mathcal{F} = \{f_\theta(x) : \theta \in \Theta\}$ with respect to a measure λ means just norm differentiability of $\sqrt{f_\theta(x)}$ as an element of $\mathcal{L}^2(\lambda)$. That is, the family is Hellinger differentiable at θ_0 if there exists a vector $\dot{\xi}(x)$ of functions in $\mathcal{L}^2(\lambda)$ such that

hell.diff <2>
$$\sqrt{f_\theta(x)} = \sqrt{f_{\theta_0}(x)} + (\theta-\theta_0)'\dot{\xi} + r(x,\theta),$$

with $\lambda r(x,\theta)^2 = o(|\theta-\theta_0|^2)$ as $\theta \to \theta_0$. Some authors (for example, Bickel, Klaassen, Ritov & Wellner (1993, page 202)) adopt a slightly different definition,

hell.diff2 <3>
$$\sqrt{f_\theta(x)} = \sqrt{f_{\theta_0}(x)} + \tfrac{1}{2}(\theta-\theta_0)'\Delta(x)\sqrt{f_{\theta_0}(x)} + r(x,\theta),$$

replacing the Hellinger derivative $\dot{\xi}$ by $\frac{1}{2}\Delta(x)\sqrt{f_{\theta_0}(x)}$. As explained inSection 2, the modification very cleverly adds an extra regularity assumption to the definition. The two definitions are not completely equivalent.

In classical statistical theory, the variance $\mathbb{I}(\theta)$ of the ***score vector*** $\partial \log f_\theta/\partial\theta$ is called the ***Fisher information matrix*** for the model. The classical regularity conditions justify differentiation under the integral sign to get

zero.deriv <4>
$$\mathbb{P}_\theta \frac{\partial}{\partial\theta}\log f_\theta = \int \frac{\partial}{\partial\theta} f_\theta(x)\,dx = \frac{\partial}{\partial\theta}\int f_\theta(x)\,dx = 0,$$

whence

$$\mathbb{I}(\theta) = \operatorname{var}_\theta\left(\frac{\partial}{\partial\theta}\log f_\theta\right) = \mathbb{P}_\theta\left(\frac{\partial}{\partial\theta}\log f_\theta\right)\left(\frac{\partial}{\partial\theta}\log f_\theta\right)'.$$

Under assumptions of Hellinger differentiability, the derivative Δ from <3> takes over the role of the score vector: Ignoring problems related to division by zero and distinctions between pointwise and $\mathcal{L}^2(\lambda)$ differentiability, we would have

$$\begin{aligned}\Delta = \frac{2\dot{\xi}(x)}{\sqrt{f_{\theta_0}(x)}} &= \frac{2}{\sqrt{f_{\theta_0}(x)}}\frac{\partial}{\partial\theta}\sqrt{f_\theta(x)}\Big|_{\theta=\theta_0} \\ &= \frac{1}{f_{\theta_0}(x)}\frac{\partial f_\theta(x)}{\partial\theta}\Big|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}\log f_\theta(x)\Big|_{\theta=\theta_0}.\end{aligned}$$

The equality <4> corresponds to

$$\mathbb{P}_{\theta_0}\Delta = 2\lambda\sqrt{f_{\theta_0}(x)}\dot{\xi} = 0,$$

an equality that Section 3 shows to be a consequence of Hellinger differentiability and the identity $\lambda f_\theta \equiv 1$.

The Fisher information at θ_0 corresponds to the matrix

$$\mathbb{P}_{\theta_0}\left(\Delta\Delta'\right) \stackrel{?}{=} 4\lambda\left(\dot\xi\dot\xi'\right).$$

Here I flag the equality as slightly suspect until we dispose (in Section 2) of hidden assumptions regarding a possible $0/0$ cancellation. Perhaps it would be better to write the last term as $\lambda(\dot\xi\dot\xi'\{f_{\theta_0} > 0\})$ until the problem is settled.

The classical assumptions also justify an integration by parts to derive the alternative representation

$$\mathbb{I}(\theta) = -\mathbb{P}_\theta \frac{\partial^2}{\partial\theta^2}\log f_\theta$$

for the information matrix. It might seem obvious that there can be no analog of this representation under an assumption of Hellinger differentiability. Indeed, how could an assumption of one-times differentiability, in norm, imply anything about a second derivative? Surprisingly, there is a way. If we think of second derivatives as coefficients of quadratic terms in local approximations then it turns out that Hellinger differentiability does have something to say about the classical dual representation for the information matrix. As shown in Chapter 6, the fact that $\sqrt{f_\theta}$ has constant $\mathcal{L}^2(\lambda)$ norm does lead to a quadratic approximation—a sort of Taylor expansion to quadratic terms without the usual assumption of twice differentiability. Remarkable.

[§intrinsic]

2. An intrinsic characterization of Hellinger differentiability

The choice of dominating measure λ for the family of probability measures $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is somewhat arbitrary.

hell.equiv <5>

Theorem. *Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, with $0 \in \Theta \subseteq \mathbb{R}^k$, is a dominated family of probability measures on a space $\mathcal{X}$, having densities $f_\theta(x)$ with respect to a sigma-finite measure λ. Define $\mathcal{U}$ as the set of unit vectors*

$$\mathcal{U} = \{u : \text{there exists a sequence } \{\theta_i\} \text{ in } \Theta \text{ such that } \theta_i/|\theta_i| \to u \text{ as } i \to \infty\}$$

Write $\tilde P_\theta$ for the part of P_θ that is absolutely continuous with respect to P_0 and $P_\theta^\perp = P_\theta - \tilde P_\theta$ for the part that is singular with respect to P_0. Write $\tilde p_\theta$ for the density $d\tilde P_\theta / dP_0$.

The following are equivalent.

(i) For some vector $\dot\xi$ of functions in $\mathcal{L}^2(\lambda)$,

$$\sqrt{f_\theta} = \sqrt{f_0} + \theta'\dot\xi + r_\theta \qquad \text{where } \lambda\left(r_\theta^2\right) = o(|\theta|^2) \text{ near } \theta = 0,$$

and $u'\dot\xi = 0$ a.e. $[\lambda]$ on $\{f_0 = 0\}$, for each u in $\mathcal{U}$.

(ii) For some vector Δ of functions in $\mathcal{L}^2(P_0)$,

$$\sqrt{f_\theta} = \sqrt{f_0} + 2\theta'\Delta\sqrt{f_0} + R_\theta \qquad \text{where } \lambda\left(R_\theta^2\right) = o(|\theta|^2) \text{ near } \theta = 0,$$

and $P_\theta^\perp \mathcal{X} = o(|\theta|^2)$.

(iii) For some vector $\tilde\Delta$ of functions in $\mathcal{L}^2(P_0)$,

$$\sqrt{\tilde p_\theta} = 1 + 2\theta'\tilde\Delta + \tilde r_\theta \qquad \text{where } P_0\left(\tilde r_\theta^2\right) = o(|\theta|^2) \text{ near } \theta = 0,$$

and $P_\theta^\perp \mathcal{X} = o(|\theta|^2)$.

Proof. Suppose (i) holds. Note that

sing.part <6> $$P_\theta^\perp \mathcal{X} = \lambda f_\theta\{f_0 = 0\} = \lambda \left(\theta'\dot{\xi} + r_\theta\right)^2 \{f_0 = 0\}$$

If this quantity is not of order $o(|\theta|^2)$ then there exists some $\epsilon > 0$ and a sequence $\theta_i \in \Theta$, with $\theta_i \to 0$, such that $P_{\theta_i}^\perp \mathcal{X}/|\theta_i|^2 \geq \epsilon$ for all i. With no loss of generality (or by subsequencing) we may assume that $\theta_i/|\theta_i| \to u \in \mathcal{U}$. We could then deduce that

$$\epsilon \leq \frac{P_{\theta_i}^\perp \mathcal{X}}{|\theta_i|^2} \to \lambda \left(u'\dot{\xi}\right)^2 \{f_0 = 0\},$$

which contradicts the fact that $u'\dot{\xi} = 0$ a.e. $[\lambda]$ on $\{f_0 = 0\}$. Define

$$\Delta = \frac{\dot{\xi}\{f_0 > 0\}}{2\sqrt{f_0}} \qquad \text{and} \qquad R_\theta = r_\theta + \theta'\dot{\xi}\{f_0 = 0\}$$

Note that $P_0|\Delta|^2 = \lambda|\dot{\xi}|^2\{f_0 > 0\} < \infty$. Argue as before along sequences $\{\theta_i\}$ for which $\theta_i/|\theta_i| \to u \in \mathcal{U}$ to deduce via <6> that $\lambda(\theta'\dot{\xi}\{f_0 = 0\})^2 = o(|\theta|^2)$. Thus (i) implies (ii).

Conversely, if (ii) holds, then each $\lambda \left(u'\dot{\xi}\right)^2 \{f_0 = 0\}$ can for $u \in \mathcal{U}$, can be recovered as the limit of a sequence $P_{\theta_i}^\perp \mathcal{X}/|\theta_i|^2 = o(1)$. If we take $\dot{\xi} = 2\Delta\sqrt{f_0}$ and $r_\theta = R_\theta$ we recover (i).

Similarly, if (ii) holds we can take $\tilde{\Delta} = \Delta$ and $\tilde{p}_\theta = f_\theta\{f_0 > 0\}/f_0$ and $\tilde{r}_\theta = R_\theta\{f_0 > 0\}\sqrt{f_0} - \{f_0 = 0\}$ to deduce (iii). (Remember that $\{f_0 = 0\} = 1$ a.e. $[P_0]$.)

Finally, if (iii) holds, and if $\mathcal{P}$ is dominated by λ with densities f_θ, we have only to note that

$$\lambda\left|\sqrt{f_\theta} - \sqrt{f_0\tilde{p}_\theta}\right|^2 = \lambda f_\theta\{f_0 = 0\} = o(|\theta|^2)$$

□ to recover (ii).

Le Cam (1986, page 575) calls the cone generated by the set of unit vectors $\mathcal{U}$ the ***contingent of*** Θ ***at*** 0. If the set of directions $\mathcal{U}$ is large enough to ensure that only the zero vector can have $u'x = 0$ for all $u \in \mathcal{U}$ then the property of the Hellinger derivative in (i) is equivalent to the assertion that $\dot{\xi} = 0$ a.e. $[\lambda]$ on$\{f = 0\}$. If 0 is an interior point of Θ then $\mathcal{U}$ consists of all unit vectors in $\mathbb{R}^k$, and the asserted equivalence holds, but at boundary points of the parameter space the situation becomes more delicate.

CounterEx <7> **Example.** Let λ be Lebesgue measure on the real line. Define

$$f_0(x) = x\{0 \leq x \leq 1\} + (2 - x)\{1 < x \leq 2\}.$$

For $0 \leq \theta \leq 1$ define densities

$$f(x, \theta) = (1 - \theta^2) f_0(x) + \theta^2 f_0(x - 2).$$

Notice that

RightDeriv <8> $$\lambda \left|\sqrt{f(x,\theta)} - \sqrt{f(x,0)} - \theta\sqrt{f(x,1)}\right|^2 = (\sqrt{1-\theta^2} - 1)^2 = O(\theta^4).$$

The family of densities is Hellinger differentiable at $\theta = 0$ with derivative

□ $\dot{\xi}(x) = \sqrt{f(x,1)}$. For this family, $\lambda\dot{\xi}^2\{f_0 = 0\} = 1$.

Failure of the requirement $P_\theta^\perp \mathcal{X} = o(|\theta|^2)$ near $\theta = 0$ would have several unfortunate consequences (see, for example, Section 5.1 and Section 6.3). To eliminate the problem once and for all, some authors, such as Le Cam & Yang (1990, page 101), include it as part of the definition of ***differentiability in quadratic mean (DQM)***. Some authors (for example, Bickel et al. 1993,

page 457) use the term DQM as a synonym for differentiability in L^2 norm. Following Le Cam & Yang, I will distinguish the term DQM from mere Hellinger differentiabilty by imposing the extra condition on the singular parts. Notice that the definition makes sense without any assumption that the family of probability measures is dominated.

dqm.def <9> **Definition.** *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, with $\theta_0 \in \Theta \subseteq \mathbb{R}^k$, be a (not necessarily dominated) family of probability measures on a space $\mathcal{X}$. Write $\tilde{P}_\theta$ for the part of P_θ that is absolutely continuous with respect to P_0 and $P_\theta^\perp = P_\theta - \tilde{P}_\theta$ for the part that is singular with respect to P_0. Say that $\mathcal{P}$ is differentiability in quadratic mean (DQM) at θ_0 if*

(i) For some vector Δ of functions in $\mathcal{L}^2(P_0)$,

$$\sqrt{\frac{d\tilde{P}_\theta}{dP_0}} = 1 + 2(\theta - \theta_0)'\Delta + r_\theta \qquad \text{where } P_0\left(r_\theta^2\right) = o(|\theta - \theta_0|^2) \text{ near } \theta_0,$$

(ii) $P_\theta^\perp \mathcal{X} = o(|\theta|^2)$.

*Call the vector Δ the **score function** at θ_0.*

image.dqm <10> **Theorem.** *Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, with $0 \in \Theta \subseteq \mathbb{R}^k$, is DQM at $\theta = 0$ with score function Δ:*

$$\sqrt{\frac{d\tilde{P}_\theta}{dP_0}} = 1 + 2\theta'\Delta + r_\theta \qquad \text{with } P_0 r_\theta^2 = o(|\theta|^2),$$

Let T be a measurable map into a space $(\mathcal{Y}, \mathcal{B})$. Define $Q_\theta = TP_\theta$ and $\dot{\eta} = P_0(\Delta \mid T = t)$. Then $\mathcal{Q} = \{Q_\theta : \theta \in \Theta\}$ is DQM at $\theta = 0$ with score function $\dot{\eta}$:

$$\sqrt{\frac{d\tilde{Q}_\theta}{dQ_0}} = 1 + 2\theta'\dot{\eta} + \rho_\theta \qquad \text{with } Q_0\rho_\theta^2 = o(|\theta|^2),$$

Proof. To simplify notation, I will assume Θ is one-dimensional. No extra conceptual difficulties arise in higher dimensions.

Abbreviate $\sqrt{d\tilde{P}_\theta/dP_0}$ to ξ_θ. Write π_t for the conditional expectation operator $P_0(\cdot \mid T = t)$: for $g \in \mathcal{L}^(Q_0)$ and $f \in \mathcal{L}^2(P_0)$,

$$P_0\left(f(x)g(Tx)\right) = Q_0\left(g(t)\pi_t(f)\right)$$

First note that the image measure $T\tilde{P}_\theta$ is absolutely continuous with respect to Q_0 with density $\tilde{q}_\theta^2 = d(T\tilde{P}_\theta)/dQ_0 = \pi_t\xi_\theta^2$, because

$$(T\tilde{P}_\theta)(g) = P_0(\xi_\theta^2 g \circ T) = Q_0\left(g(t)\pi_t\xi_\theta^2\right)$$

For the sake of notational simplicity, I will act as if $T\tilde{P}_\theta = \tilde{Q}_\theta$, the part of Q_θ that is absolutely continuous with respect to Q_0, which is not quite true: the image measure $TP_\theta^\perp$ might also have a part absolutely continuous with respect to Q_0. As explained at the end of the proof, any extra contribution could be absorbed into a $o(\theta^2)$ term, thereby having no further effect on the main argument.

A rough heuristic, which ignores all terms of order $o(|\theta|)$ in various senses, shows why $\dot{\eta}$ might play the role of score function:

$$\sqrt{\pi_t\xi_\theta^2} = \sqrt{\pi_t\left(1 + 2\theta\Delta + r_\theta\right)^2} \approx \sqrt{\pi_t\left(1 + 4\theta\Delta\right)} \approx 1 + 2\theta\pi_t(\Delta)$$

Define

eta.def <11> $$\eta_\theta(t) = \pi_t\xi_\theta = 1 + 2\theta\dot{\eta} + \tilde{\rho}_\theta \qquad \text{where } \tilde{\rho}_\theta = \pi_t r_\theta$$

Note that $Q_0\left(\pi_t r_\theta\right)^2 \le Q_0\pi_t\left(r_\theta^2\right) = P_0 r_\theta^2 = o(\theta^2)$.

The conditional analog of the formula $\mathbb{P}X^2 = \operatorname{var}(X) + (\mathbb{P}X)^2$ gives

$$\begin{aligned}\tilde{q}_\theta &= \pi_t \xi_\theta^2 \\ &= \pi_t \left(\xi_\theta - \eta_\theta(t)\right)^2 + \eta_\theta(t)^2 \\ &= B_\theta(t) + \eta_\theta(t)^2 \qquad \text{where } B_\theta(t) = \pi_t \left(\theta(\Delta - \dot\eta) + r_\theta - \rho_\theta\right)^2 .\end{aligned} \tag{image.sum <12>}$$

Notice the cancellation of the leading constants in the definition of the nonnegative term B_θ. Write $\sqrt{\tilde{q}_\theta} = \epsilon_\theta(t) + \eta_\theta(t)$ with $\epsilon_\theta \geq 0$ a.e. $[Q_0]$. I claim that $Q_0\epsilon_\theta^2 = o(\theta^2)$, which would leave

$$\sqrt{\frac{dQ_\theta}{dQ_0}} = 1 + 2\theta\dot\eta + (\tilde\rho_\theta + \epsilon_\theta)$$

as the desired DQM expansion.

From the defining equality <12>,

$$2\epsilon_\theta\eta_\theta + \epsilon_\theta^2 = B_\theta(t) \leq 2\theta^2\pi_t(\Delta - \dot\eta)^2 + 2\pi_t(r_\theta - \tilde\rho_\theta)^2 \tag{eps.bnd <13>}$$

Write $G(t)$ for the Q_0-square integrable function $\sqrt{\pi_t(\Delta - \dot\eta)^2}$ and δ_θ for the square root of the other term. Note that $Q_0\delta_\theta^2 \leq 2P_0 r_\theta^2 = o(\theta^2)$.

Roughly speaking $\eta_\theta \approx 1$, so that <13> should give an upper bound for ϵ_θ. To formalize the argument we need to consider separately contributions from cases where quanties behave as suggested by their $\mathcal{L}^2(Q_0)$ norms or not. Let M_θ be a positive constant for which $M_\theta \to \infty$ and $1/4 \geq |\theta| M_\theta \to 0$ as $\theta \to 0$. Define a truncation region

$$\Gamma_\theta = \{t : G(t) \leq M_\theta,\ |\tilde\rho_\theta| \leq 1/4,\ |\delta_\theta| \leq 1\}$$

Note that $Q_0\Gamma_\theta^c \to 0$, which lets us deduce via Dominated Convergence that

$$Q_0\epsilon_\theta^2\Gamma_\theta^c \leq Q_0 B_\theta(t)\Gamma_\theta^c \leq \theta^2 Q_0 G(t)^2\Gamma_\theta^c + Q_0\delta_\theta^2 = o(\theta^2).$$

On the set Γ_θ we have

$$\eta_\theta(t) \geq 1 - |\theta\dot\eta| - |\tilde\rho_\theta| \geq 1 - |\theta| M_\theta - \tfrac{1}{4} \geq \tfrac{1}{2}$$

Thus $\epsilon_\theta\Gamma_\theta \leq B_\theta\Gamma_\theta \leq \theta^2 M_\theta G(t) + \delta_\theta(t)$, from which we get

$$Q_0\epsilon_\theta^2\Gamma_\theta \leq 2\theta^4 M_\theta^2 Q_0 G^2 + 2Q_0\delta_\theta^2 = o(\theta^2),$$

as asserted.

If the image measure $TP_\theta^\perp$ has a part absolutely continuous with respect to Q_0 we could add its density σ_θ to $\tilde{q}_\theta$ to get dQ_θ/dQ_0. We could repeat the argument, but with δ_θ^2 replaced by $\delta_\theta^2 + \sigma_\theta$, whose Q_0-integral is still of order $o(\theta^2)$, because $Q_0\sigma_\theta \leq (TP_\theta^\perp)\mathcal{Y} = P_\theta^\perp\mathcal{X} = o(\theta^2)$. □

REMARK. It is common practice to write $d\nu/d\mu 0$ to denote the density with respect to μ of the part of ν that is absolutely continuous with respect to μ.

[§unit.vector] 3. Differentiability of unit vectors

For a differentiable map τ into some inner product space $\mathcal{H}$ (such as $\mathcal{L}^2(\lambda)$), the Cauchy-Schwarz inequality implies that the inner product $\langle \tau(\theta_0), r(\theta)\rangle$ converges to zero at a $o(|\theta - \theta_0|)$ rate. It would usually be a blunder to assume naively that the bound must therefore be of order $O(|\theta - \theta_0|^2)$; typically, higher-order differentiability assumptions are needed to derive approximations with smaller errors. However, if $\|\tau(\theta)\|$ is constant—that is, if the function is constrained to take values lying on the surface of a sphere—then the naive assumption turns out to be no blunder. Indeed, in that case, at least if θ_0 is an

interior point of the parameter space, $\langle \tau(\theta_0), r(\theta)\rangle$ can be written as a quadratic in $\theta-\theta_0$ plus an error of order $o(|\theta-\theta_0|^2)$. The sequential form of the assertion will be more convenient for the calculations in Section SECT.LAN.

UNITvector <14> **Lemma.** *Let $\{\delta_n\}$ be a sequence of constants tending to zero. Let $\tau_0, \tau_1, \ldots$ be elements of norm one for which $\tau_n = \tau_0 + \delta_n W + r_n$, with W a fixed element of $\mathcal{H}$ and $\|r_n\| = o(\delta_n)$. Then $\langle \tau_0, W\rangle = 0$ and $2\langle \tau_0, r_n\rangle = -\delta_n^2\|W\|^2 + o(\delta_n^2)$.*

Proof. Because both τ_n and τ_0 have unit length,

$$\begin{aligned} 0 = \|\tau_n\|^2 - \|\tau_0\|^2 = 2\delta_n\langle \tau_0, W\rangle & \quad \text{order } O(\delta_n) \\ + 2\langle \tau_0, r_n\rangle & \quad \text{order } o(\delta_n) \\ + \delta_n^2\|W\|^2 & \quad \text{order } O(\delta_n^2) \\ + 2\delta_n\langle W, r_n\rangle + \|r_n\|^2 & \quad \text{order } o(\delta_n^2). \end{aligned}$$

The Cauchy-Schwarz inequality delivers the $o(\delta_n)$ and $o(\delta_n^2)$ rates of convergence. The exact zero on the left-hand side of the equality exposes the leading $2\delta_n\langle \tau_0, W\rangle$ as the only $O(\delta_n)$ term. It must be of smaller order, which can happen only if $\langle \tau_0, W\rangle = 0$, leaving

$$0 = 2\langle \tau_0, r_n\rangle + \delta_n^2\|W\|^2 + o(\delta_n^2),$$

□ as asserted.

Without the fixed length property, the difference $\|\tau_n\|^2 - \|\tau_0\|^2$ might contain terms of order δ_n. The inner product $\langle \tau_0, r_n\rangle$, which inherits $o(\delta_n)$ behaviour from $\|r_n\|$, might then not decrease at the $O(\delta_n^2)$ rate.

unit2 <15> **Corollary.** *Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a famility of probability densities indexed by a subset Θ of $\mathbb{R}^k$. If $\mathcal{F}$ has a Hellinger derivative Δ at an interior point θ_0 of Θ, then $\lambda(\sqrt{f_{\theta_0}}\Delta) = 0$ and*

$$2\lambda\left(\sqrt{f_{\theta_0}}r_\theta\right) = -(\theta-\theta_0)'\lambda(\Delta\Delta')(\theta-\theta_0) + o(|\theta-\theta_0|^2) \qquad \text{near } \theta_0$$

Proof. Abbreviate $\sqrt{f_{\theta_0}}$ to ξ_0. Choose $\theta = \delta_n u$ for a fixed unit vector. Invoke the Lemma with $W = u'\Delta$ to deduce that $\lambda(\xi_0 u'\Delta) = 0$ for *every* unit vector u in $\mathbb{R}^k$, because θ_0 is an interior point of Θ. Hence $\lambda(\xi_0\Delta) = 0$.

For the second assertion consider sequences $\theta_n = \theta_0 + \delta_n u_n$, with $\{u_n\}$ a sequence of unit vectors and $\delta_n \to 0$. Abbreviate ξ_{θ_n} to ξ_n and r_{θ_n} to r_n. We need to show that

$$2\lambda(\xi_0 r_n) = -\delta_n^2\lambda(u_n'\Delta)^2 + o(\delta_n^2)$$

With no loss of generality we may assume that u_n converges to some unit vector u. (Equivalently, argue along sub-subsequences.) Then

$$\xi_n = \xi_0 + \delta_n u_n'\Delta + r_n = \xi_0 + \delta_n u'\Delta + r_n + \delta_n(u_n - u)'\Delta.$$

Note that $\|r_n + \delta_n(u_n-u)'\Delta\|_2 = o(|\delta_n|)$. Invoke the Lemma with $W = u'\Delta$ and $r_n + \delta_n(u_n-u)'\Delta$ playing the role of r_n to deduce that

$$2\lambda(\xi_0 r_n) = -\delta_n^2\lambda(u'\Delta)^2 + o(\delta_n^2) = -\delta_n^2\lambda(u_n'\Delta)^2 + o(\delta_n^2),$$

□ as required.

REMARK. If θ_0 were not an interior point of the parameter space, there might not be enough directions u along which to approach θ_0, and it might not follow that $\lambda(\sqrt{f_{\theta_0}}\Delta) = 0$. Roughly speaking, the set of such directions is called the ***contingent*** of Θ at θ_0. If the contingent is 'rich enough', we do not need to assume that θ_0 is an interior point. See Le Cam & Yang (1988, Section 6.2) and Le Cam (1986, page 575) for further details.

[§pwise]

4. A sufficient condition for Hellinger differentiability

How does Hellinger differentiability relate to the classical assumption of pointwise differentiability?

Consider the case of a one-parameter family $\mathcal{F} = \{f_t : -\delta < t < \delta\}$ of probability densities with respect to a measure λ. Once again write ξ_t for $\sqrt{f_t}$ and $\|\cdot\|_2$ for the $\mathcal{L}^2(\lambda)$ norm.

First suppose $\mathcal{F}$ has a Hellinger derivative $\Delta(x)$ at $t = 0$. That is, suppose the remainder term

$$r(x,t) = \xi_t(x) - \xi_0(x) - t\Delta(x),$$

has $\mathcal{L}^2(\lambda)$ norm of order $o(|t|)$ as $t \to 0$. If a sequence $\{t_n\}$ tends to zero fast enough then

$$\| \sum_n \frac{|r(\cdot,t_n)|}{|t_n|} \|_2 \le \sum_n \frac{\|r(\cdot,t_n)\|_2}{|t_n|} < \infty,$$

from which it follows that $|r(x,t_n)| = o(|t_n|)$ for λ-almost all x. Unfortunately the aberrant negligible set of x might depend on $\{t_n\}$, so we cannot invoke the usual subsequencing argument to deduce that $|r(x,t)| = o(|t|) \mod[\lambda]$. That is, it does not follow immediately that $t \mapsto \xi_t(x)$ is differentiable at $t = 0$ for λ-almost all x. However, if by some means we can show that the pointwise derivative $\xi_t'(x)$ does exist then we must have $\Delta(x) = \xi_0'(x) \mod[\lambda]$. For example, if $t \mapsto f_t(x)$ has derivative $f_t'(x)$ at $t = 0$, and if $f_t(x) > 0$, then $2\xi_0'(x) = f_0'(x)/\xi_0(x)$. At points x where $f_0(x) = 0$, the derivative $f_0'(x)$ must also be zero, for otherwise $f_t(x)$ would be strictly negative for some small t, either positive or negative. Perhaps there is no harm in writing $f_0'(x)/\xi_0(x)$, with the interpretation that $0/0 = 0$, but I think it is safer to append an explicit indicator function $\{\xi_0 > 0\}$ to avoid inadvertent disastrous cancellation of zeros. In short, if the pointwise derivative $f_0'(x)$ exists then

$$\xi_0'(x) = \frac{f_0'(x)\{\xi_0(x) > 0\}}{2\xi_0(x)}$$

is, up to a λ-equivalence, the only candidate for a Hellinger derivative at $t = 0$.

Now consider the situation where we have pointwise differentiability, and we wish to deduce Hellinger differentiability. What more is needed? The answer requires careful attention to the problem of when functions of a real variable can be recovered as integrals of their derivatives.

abs.cty.def <16> **Definition.** *A real valued function H defined on an interval $[a,b]$ of the real line is said to be* ***absolutely continuous*** *if to each $\epsilon > 0$ there exists a $\delta > 0$ such that $\sum_i |H(b_i) - H(a_i)| < \epsilon$ for all finite collections of disjoint subintervals (a_i, b_i) of $[a,b]$ for which $\sum_i (b_i - a_i) < \delta$.*

Absolute continuity of a function defined on the whole real line is taken to mean absolute continuity on each finite subinterval.

The connection between absolute continuity and integration of derivatives is one of the most celebrated results of classical analysis (see, for example, Chapter 5 of Royden 1974):

fundamental <17> **Theorem.** *A real valued function H defined on an interval $[a,b]$ is absolutely continuous if and only if the following three conditions hold*

(i) the derivative $H'(x)$ exists at Lebesgue almost all points of $[a,b]$

(ii) the derivative H' is integrable

(iii) $H(x)-H(a)=\int_a^x H'(t)\,dt$ for each x in $[a,b]$

Put another way, a function H is absolutely continuous on an interval $[a,b]$ if and only if there exists an integrable function h for which

ac.integral <18> $$H(x)=\int_a^x h(t)\,dt \qquad \text{for all } x \text{ in } [a,b]$$

The function H must then have derivative $h(x)$ at almost all x. As a systematic convention we could take h equal to the measurable function

$$\dot H(x)=\begin{cases} H'(x) & \text{at points } x \text{ where the derivative exists} \\ 0 & \text{elsewhere}\end{cases}$$

I will refer to $\dot H$ as the ***density***. Of course it is actually immaterial how $\dot H$ is defined on the Lebesgue negligible set of points at which the derivative does not exist, but the convention helps to avoid ambiguity.

Now consider a nonnegative function H that is differentiable at a point x. If $H(x)>0$ then the chain rule of elementary calculus implies that the function $2\sqrt{H}$ is also differentiable at x, with derivative $H'(x)/\sqrt{H(x)}$. At points where $H(x)=0$ the question of differentiability becomes more delicate, because the map $s\mapsto\sqrt{s}$ is not differentiable at the origin. If x is an internal point of the interval then we must have $H'(x)=0$, whence $H(y)=o(|y-x|)$ near x. If $\sqrt{H}$ had a derivative D at x then we would have $H(y)=(y-x)^2D+o(|y-x|^2)$ near x, a stronger requirement. Clearly we need to take some care with the question of differentiability at points where H equals zero.

Even more subtle is the fact that absolute continuity of the nonnegative function H need not imply absolute continuity of the function $\sqrt{H}$, without further assumptions—even if H is everywhere differentiable (Problem [1]).

sqrt.ac <19> **Lemma.** *Suppose the nonnegative function H is absolutely continuous on an interval $[a,b]$, with density $\dot H$. Let $\Delta(x)=\frac12\dot H(x)\{H(x)>0\}/\sqrt{H(x)}$. If $\int_a^b|\Delta(x)|\,dx<\infty$ then $\sqrt{H}$ is absolutely continuous, with density Δ, that is,*

$$\sqrt{H(x)}-\sqrt{H(a)}=\int_a^x\Delta(t)\,dt \qquad \text{for all } x \text{ in } [a,b]$$

Proof. Fix an $\eta>0$. The function $H_\eta=\eta+H$ is bounded away from zero, and hence $\sqrt{H_\eta}$ has derivative $H_\eta'=H'/(2\sqrt{H+\eta})$ at each point where the derivative H' exists. Moreover, absolute continuity of H_η follows directly from the Definition <16>, because

$$|\sqrt{H_\eta(b_i)}-\sqrt{H_\eta(a_i)}|=\frac{|H_\eta(b_i)-H_\eta(a_i)|}{\sqrt{H_\eta(b_i)}+\sqrt{H_\eta(a_i)}}\le\frac{|H(b_i)-H(a_i)|}{2\sqrt{\eta}}$$

for each interval (a_i,b_i).

From Theorem <17>, for each x in $[a,b]$,

$$\sqrt{H(x)+\eta}-\sqrt{H(a)+\eta}=\int_a^x\frac{\dot H(t)}{2\sqrt{H(t)+\eta}}\,dt$$

As η decreases to zero, the left-hand side converges to $\sqrt{H(x)}-\sqrt{H(a)}$. The integrand on the right-hand side converges to $\Delta(t)$ at points where $H(t)>0$. For almost all t in $\{H=0\}$ the derivative exists and equals zero; the integrand also converges to $\Delta(t)$ at those points. By Dominated Convergence, the right-hand side converges to $\int_a^x\Delta(t)\,dt$. □

The integral representation for the square root of an absolutely continuous function is often the key to proofs of Hellinger differentiability.

Hdiff.suff <20> **Theorem.** *Let $\mathcal{F} = \{f_\theta(x) : |\theta| < \delta\}$ be a family of probability densities with respect to a measure λ, with $(x,\theta) \mapsto f_\theta(x)$ measurable. Suppose*

(i) for λ almost all x, the function $\theta \mapsto f_\theta(x)$ is absolutely continuous on $(-\delta,\delta)$, with density $\dot f_\theta(x)$;

(ii) for λ almost all x, the function $\theta \mapsto f_\theta(x)$ is differentiable at $t = 0$

(iii) for each t the function

$$\Delta_\theta(x) = \frac{\dot f_\theta(x)\{f_\theta(x) > 0\}}{2\sqrt{f_\theta(x)}}$$

is square integrable with respect to λ and $\lambda\Delta_\theta^2 \to \lambda\Delta_0^2$ as $\theta \to 0$.

Then $\mathcal{F}$ has Hellinger derivative $\Delta_0(x)$ at $\theta = 0$.

REMARK. Assumption (ii) might appear redundant, because (i) implies differentiability of $\theta \mapsto f_\theta(x)$ at Lebesgue almost all θ, for λ-almost all x. A mathematical optimist (or Bayesian) might be prepared to gamble that 0 does not belong to the bad negligible set; a mathematical pessimist might prefer Assumption (ii).

Proof. As before write $\xi_\theta(x)$ for $\sqrt{f_\theta(x)}$, and define

$$r_\theta(x) = \xi_\theta(x) - \xi_0(x) - \theta\Delta_0(x).$$

We need to prove that $\lambda r_\theta^2 = o(|\theta|^2)$ as $\theta \to 0$.

For simplicity of notation, consider only positive θ. The arguments for negative θ are analogous.

With no loss of generality (or by a suitable decrease in δ) we may assume that $\lambda\Delta_\theta^2$ is bounded, so that, by Fubini,

$$\infty > \lambda^x \int_{-\delta}^{\delta} \Delta_t(x)^2\,dt = \int_{-\delta}^{\delta} \lambda^x \Delta_t(x)^2\,dt.$$

For λ-almost all x, the function $t \mapsto \Delta_t(x)$ is Lebesgue (square) integrable. From Lemma <19>,

$$\frac{\xi_\theta(x) - \xi_0(x)}{\theta} = \frac{1}{\theta}\int_0^\theta \Delta_t(x)\,dt \qquad \text{mod}[\lambda].$$

By Jensen's inequality for the uniform distribution on $[0,\theta]$, and (iii),

limsup.diff <21> $$\lambda\left|\frac{\xi_\theta(x) - \xi_0(x)}{\theta}\right|^2 \le \frac{1}{\theta}\int_0^\theta \lambda\Delta_t(x)^2\,dt \to \lambda\Delta_0^2 \qquad \text{as } \theta \to 0.$$

Define nonnegative, measurable functions

$$g_\theta(x) = 2\left|\frac{\xi_\theta(x) - \xi_0(x)}{\theta}\right|^2 + 2\Delta_0(x)^2 - \left|\frac{r_\theta(x)}{\theta}\right|^2.$$

By (ii), for almost all x at which $\xi_0(x) > 0$ we have $r_\theta(x)/\theta \to 0$, and hence $g_\theta(x) \to 4\Delta_0(x)^2$. At points where $\xi_0(x) = 0$ we have $\Delta_0(x) = 0$. Thus $\liminf g_\theta(x) \ge 4\Delta_0(x)^2$ mod[λ]. By Fatou's Lemma (applied along subsequences), followed by an appeal to <21>,

$$4\lambda\Delta_0^2 \le \liminf_{\theta\to 0} \lambda g_\theta \le 4\lambda\Delta_0^2 - \limsup_{\theta\to 0} \lambda\left|\frac{r_\theta(x)}{\theta}\right|^2.$$

□ That is, $\lambda r_\theta^2 = o(\theta^2)$, as required for Hellinger differentiability.

shift.family <22> **Example.** Let f be a probability density with respect to Lebesgue measure λ on the real line. Suppose f is absolutely continuous, with density $\dot f$ for which

$\mathbb{I} := \lambda\left(\{f > 0\}\dot{f}^2/f\right) < \infty$. Define the corresponding shift family of densities, $\mathcal{F} = \{f_\theta : \theta \in \mathbb{R}\}$, by $f_\theta(x) = f(x-\theta)$.

Each f_θ is also absolutely continuous, with density $\dot{f}_\theta(x) = \dot{f}(x-\theta)$. For each fixed θ_0, differentiability of f at almost all x implies differentiability of $\theta \mapsto f_{\theta+\theta}(x)$ at $\theta = 0$ for almost all x. The integral $\lambda\left(\{f_{\theta_0} > 0\}\dot{f}_{\theta_0}^2/f_{\theta_0}\right)$ equals $\mathbb{I}$ for every θ_0. From the Lemma, applied to $f_{\theta_0+\theta}$, we get Hellinger differentiability of $\mathcal{F}$ at every θ_0. □

[§info]

5. Information inequality

Let $\mathcal{F} = \{f_\theta : \theta \in J\}$ be a family of probability densities on a space $\mathcal{X}$, with corresponding probability measures $\{\mathbb{P}_\theta\}$. For simplicity, suppose $\Theta \subseteq \mathbb{R}$. The information inequality bounds the variance of an estimator $T(x)$ from below by a function of the expected value, $\gamma(\theta) = \mathbb{P}_\theta T$, of the statistic and the Fisher information: under suitable regularity conditions,

$$\operatorname{var}_\theta(T) \geq \frac{\gamma'(\theta)^2}{\mathbb{I}(\theta)}$$

The classical proof of the inequality imposes assumptions that derivatives can be passed inside integral signs, typically justified by more primitive assumptions involving pointwise differentiability of densities and domination assumptions about their derivatives.

By contrast, the proof of the information inequality based on an assumption of Hellinger differentiability replaces the classical requirements by simple properties of $\mathcal{L}^2(\lambda)$ norms and inner products. The gain in elegance and economy of assumptions illustrates the typical benefits of working with Hellinger differentiability. The main technical ideas ar captured by the following Lemma.

Tdiff <23>

Lemma. *Suppose the familiy of probability densities $\mathcal{F}$ has Hellinger derivative Δ at θ_0. Suppose $\sup_{\theta\in J}\mathbb{P}_\theta T(x)^2 < \infty$, for some neighborhood J of θ_0. Then the expected value, $\gamma(\theta) = \mathbb{P}_\theta T = \lambda(f_\theta(x)T(x))$, has derivative $2\lambda(\sqrt{f_{\theta_0}}\Delta T)$ at θ_0.*

REMARK. Notice that $\mathbb{P}_\theta T$ is well defined throughout J, because of (i). Also $\sqrt{f_{\theta_0}}\Delta T$ is integrable, because $\left(\lambda|\sqrt{f_{\theta_0}}\Delta T|\right)^2 \leq \left(\lambda f_\theta T^2\right)\left(\lambda\Delta^2\right) < \infty$.

Proof. Without loss of generality suppose $\theta_0 = 0$. Use $\|\cdot\|_2$ to denote $\mathcal{L}^2(\lambda)$ norms. Hellinger differentiability means that

$$\xi_\theta = \xi_0 + \theta\Delta + r_\theta \qquad \text{with } \|r_\theta\|_2 = o(|\theta|) \text{ as } \theta \to 0.$$

Write C^2 for $\sup_{\theta\in J}\mathbb{P}_\theta T(x)^2$. Then $\|\xi_\theta T\|_2 \leq C$ for each θ in J.

The proof is easy if T is bounded by a finite constant K:

$$\begin{aligned}
\operatorname{rem}(\theta) &:= |\gamma(\theta) - \gamma(0) - 2\theta\lambda(\xi_0\Delta T)| \\
&= |\lambda\left(\xi_\theta^2 - \xi_0^2 - 2\theta\xi_0\Delta\right)T| \\
&= \lambda\left|\theta^2\Delta^2 + r_\theta^2 + 2\xi_0 r_\theta + 2\theta\Delta r_\theta\right||T| \qquad \text{remainder } \langle 24\rangle \\
&\leq K\theta^2\|\Delta\|_2^2 + K\|r_\theta\|_2^2 \\
&\qquad + 2\|\xi_0 T\|_2\,\|r_\theta\|_2 + 2K|\theta|\,\|\Delta\|_2\,\|r_\theta\|_2 \qquad \text{by Cauchy-Schwarz} \\
&= o(|\theta|).
\end{aligned}$$

Notice that K need not be fixed for the last conclusion. It would suffice if we had $|T| \leq K_\theta = o(1/|\theta|)$.

If T is not bounded, the argument can be rescued by a truncation at a level K_θ that increases to infinity at a $o(1/|\theta|)$ rate. The contributions to rem(θ)

from the region where $|T| \le K_\theta$ are of order $o(|\theta|)$. We have only to show that the contributions from the region where $|T| > K_\theta$ are also of order $o(|\theta|)$ to complete the proof of differentiability. That is, we need to show that

upper.tail <25> $$\lambda(\xi_\theta^2 - \xi_0^2)T\{|T| > K_\theta\} - 2\theta\lambda\,(\xi_0 \Delta T\{|T| > K_\theta\}) = o(|\theta|).$$

On the left-hand side, the coefficient of 2θ in the second term is bounded in absolute value by

$$\lambda|\xi_0 \Delta T\{|T| > K_\theta\}| \le \|\Delta\{|T| > K_\theta\}\|_2 \, \|\xi_0 T\|_2 \le o(1)C,$$

the $o(1)$ term on the right-hand side coming via Dominated Convergence and the λ-integrability of Δ^2. For the first term on the left-hand side of <25>, factorize $\xi_\theta^2 - \xi_0^2$ as $(\theta\Delta + r_\theta)(\xi_\theta + \xi_0)$ then expand, to get

$$\begin{aligned}
&|\lambda(\xi_\theta^2 - \xi_0^2)T\{|T| > K_\theta\}| \\
&\quad \le \lambda|\theta\Delta\{|T| > K_\theta\} + r_\theta|\,|\xi_\theta T + \xi_0 T| \\
&\quad \le \big(|\theta| \cdot \|\Delta\{|T| > K_\theta\}\|_2 + \|r_\theta\|_2\big)\big(\|\xi_\theta T\|_2 + \|\xi_0 T\|_2\big),
\end{aligned}$$

the last bound following from several applications of the Cauchy-Schwarz inequality. Both terms in the leading factor are of order $o(|\theta|)$; both terms in the other factor are bounded by C. The remainder $\text{rem}(\theta)$ is of order $o(|\theta|)$, as required for differentiability. □

Remember from Section 1 that $4\lambda(\Delta^2)$ corresponds to the Fisher information at θ_0.

CRbnd <26> **Corollary.** *In addition to the conditions of the Lemma, suppose* $\lambda(\Delta^2) > 0$. *Then* $\text{var}_{\theta_0} T \ge \gamma'(\theta_0)^2 / \big(4\lambda\Delta^2\big)$.

Proof. The special case where $T \equiv 1$ gives $\lambda(\xi_0\Delta) = 0$ (or use Lemma <14>). Write γ_0 for $\gamma(\theta_0)$. From Lemma <23> deduce that $\gamma'(\theta_0)^2 = 4\langle\Delta, \xi_0(T - \gamma_0)\rangle^2$, which the Cauchy-Schwarz inequality bounds by $4\|\Delta\|_2^2 \, \mathbb{P}_{\theta_0}(T - \gamma_0)^2$. □

Variations on the information inequality lead to other useful lower bounds for variances and mean squared errors of statistics.

vanTrees <27> **Example.** Suppose $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is a family of probability densities with respect to a measure λ, with index set Θ an open subset of the real line. Suppose $\mathcal{F}$ not only has Hellinger derivative $\Delta_\theta(x)$ at each point of Θ, but also that ξ_θ, the square root of the density, has the representation

xi.rep <28> $$\xi_{\theta+\beta}(x) - \xi_\theta(x) = \int_\theta^{\theta+\beta} \Delta_t(x)\, dt \qquad \text{mod}[\lambda] \qquad \text{for all } |\beta| \le \delta,$$

as in Theorem <20>. Suppose also that $\lambda\Delta_\theta^2$ is bounded by a constant C for all θ in $[a - \delta, a + \delta]$, a bounded subinterval of Θ.

Fix!

Let $\mathcal{Q} = \{q_\alpha : -\delta < \alpha < \delta\}$ be a family of probability densities with respect to Lebesgue measure μ on the real line, each concentrated on $[a, b]$. To make things easy, suppose that each q_α is bounded by a fixed constant K. Write $\eta_\alpha(\theta)$ for $\sqrt{q_\alpha(\theta)}$. Suppose $\mathcal{Q}$ has Hellinger derivative $\dot\eta$ at $\alpha = 0$: that is, the remainder term $\rho_\alpha(\theta) = \eta_\alpha(\theta) - \eta_0(\theta) - \alpha\dot\eta(\theta)$ has $\mu(\rho_\alpha^2) = o(\alpha^2)$ as $\alpha \to 0$.

Consider once more the two-parameter of family of densities constructed in Example <8>. The subfamily of densities $m_\alpha(x, \theta) = q_\alpha(\theta) \, f_{\theta-\alpha}(x)$ has Hellinger derivative $\dot\eta\xi_\theta - \eta_0\Delta_\theta$ at $\alpha = 0$. Take $q_\alpha(\theta) = q(\theta - \alpha)$, a translation of a fixed absolutely continuous density q with compact support, for which $\mu\dot q^2/q < \infty$. Write $\mathbb{M}_\alpha$ for the probability measure corresponding to m_α.

Let $T(x)$ be a statistic, an estimator for the unknown parameter θ in the model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ defined by the f_θ densities. Write $\gamma(\theta)$ for the

expected value $\mathbb{P}_\theta T = \lambda f_\theta(x)T(x)$. Under $\mathbb{M}_\alpha$, the random variable $T(x) - \theta$ has expected value

$$\mathbb{M}_\alpha^{x,\theta}(T(x) - \theta) = \mu^\theta q(\theta - \alpha)\,(\gamma(\theta - \alpha) - \theta) = \mu^t q(t)\,(\gamma(t) - t) + \alpha$$

The squared norm of the Hellinger derivative equals

$$\lambda \otimes \mu(\dot\eta\xi_\theta - \eta_0\Delta_\theta)^2 = \mu\dot\eta^2 + \mu q_0(\theta)\lambda\Delta_\theta^2,$$

because $\lambda\xi_\theta\Delta_\theta = 0$. The information inequality for the one-parameter family takes an elegant form,

$$\mu^\theta q(\theta)\mathbb{P}_\theta\,(T(x) - \theta)^2 \geq \frac{1}{\mathbb{I}_q + \mu^\theta q(\theta)\mathbb{I}(\theta)},$$

where $\mathbb{I}_q = 4\mu\dot\eta^2 = \mu\dot q^2/q$ denotes the information function for the shift family, and $\mathbb{I}(\theta) = \lambda\Delta_\theta^2$ denotes the information function for the $\mathcal{P}$ model.

The inequality is known as the ***van Trees inequality***. It has many statistical applications. See Gill & Levit (1995) for details. □

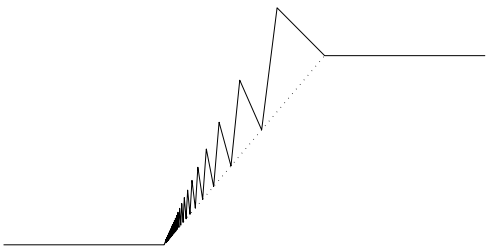
[§] 6. Problems

Prob.sqrt.nonac [1] (Construction of an absolutely continuous density whose square root is not absolutely continuous.) For $i \geq 3$ define

$$\alpha_i = \frac{1}{i(\log i)^2} \qquad \text{and} \qquad \beta_i = \frac{1}{i(\log i)^5},$$

Define $B_i = 2\sum_{j\geq i}\beta_j$. Define functions

$$H_i(t) = \alpha_i\,(1 - |t - B_i - \beta_i|/\beta_i)^+ \qquad \text{and} \qquad H(t) = (1 \wedge t)^+ + \sum_{i\geq 3} H_i(t).$$



(i) Show that B_i decreases like $(\log i)^{-4}$.

(ii) Use the fact that $\sum_i \alpha_i < \infty$ to prove that H is absolutely continuous.

(iii) Show that $\alpha_i/B_i \to 0$, then deduce that H has derivative 1 at 0.

(iv) Show that

$$\sqrt{H(B_{i-1} - \beta_i)} - \sqrt{H(B_{i-1})} = \frac{\alpha_i - \beta_i}{\sqrt{H(B_{i-1} + \beta_i)} + \sqrt{H(B_{i-1})}},$$

which decreases like $1/i$, then deduce that

$$\sum_{i=k}^{k+m} |\sqrt{H(B_{i-1} - \beta_i)} - \sqrt{H(B_{i-1})}|$$

can be made arbitrarily large while keeping $\sum_{i=k}^{k+m}|\beta_i|$ arbitrarily small. Deduce that $\sqrt{H}$ is not absolutely continuous.

(v) Show, by an appropriate "rounding off of the corners" at each point where H has different left and right derivatives followed by some smooth truncation and rescaling, that there exists an absolutely continuous,

everywhere differentiable probability density function f for which $\sqrt{f}$ is not absolutely continuous.

dbl.exp [2] Let $f_\theta(x) = \frac{1}{2}\exp(-|x-\theta|)$, for $\theta \in \mathbb{R}$ (the double-exponential location family of densities with respect to Lebesgue measure).

(i) Show that $\int \sqrt{f_\theta(x)\, f_{\theta+\delta}(x)}\, dx = (1+\delta/2)\exp(-\delta/2)$.

(ii) Deduce that the density f_θ is Hellinger differentiable at every θ.

(iii) Show that $\theta \mapsto f_\theta(x)$ is not differentiable, for each fixed x, at $\theta = x$.

(iv) Prove Hellinger differentiability by a direct Dominated Convergence argument, without the explicit calculation from (i).

(v) Prove Hellinger differentiability by an appeal to Example <22>, without the explicit calculation from (i).

deriv.vanish [3] Suppose $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is a familiy of densities indexed by a subset Θ of $\mathbb{R}^k$. Suppose 0 is an interior point of Θ and that $\mathcal{F}$ is Hellinger differentiable at $\theta = 0$, with derivative Δ. Show that $\Delta(x) = 0$ almost everywhere on $\{f_0 = 0\}$. Hint: Approach 0 from each direction in $\mathbb{R}^k$. Deduce that both $\mathbb{P}_0\Delta\{f_0 = 0\}$ and $\mathbb{P}_0\Delta^2\{f_0 = 0\}$ equal zero.

HD.product [4] Suppose $\mathcal{F} = \{f_t(x) : t \in T\}$ is a family of probability densities with respect to a measure λ, $\mathcal{G} = \{g_s(x) : s \in S\}$ is a family of probability densities with respect to a measure μ. Suppose $\mathcal{F}$ is Hellinger differentiable at $t = 0$ and $\mathcal{G}$ is Hellinger differentiable at $s = 0$. Show that the family of densities $\{f_s(x)g_t(y) : (s,t) \in S \otimes T\}$ with respect to $\lambda \otimes \mu$ is Hellinger differentiable at $(s,t) = (0,0)$. Hint: Use Cauchy-Schwarz to bound contributions from most of the cross-product terms in the expansion of $\sqrt{f_t(x)g_s(y)}$.

DQM.L1 [5] Suppose $\mathcal{F} = \{f_\theta : \theta \in \mathbb{R}^k\}$ has Hellinger derivative Δ at θ_0. Show that $\mathcal{F}$ is also differentiable in $\mathcal{L}^1$ norm with derivative $\Delta_1 = 2\sqrt{f_{\theta_0}}\Delta$, that is, show

$$\lambda|f_\theta - f_{\theta_0} - (\theta-\theta_0)'\Delta_1| = o(|\theta-\theta_0|) \qquad \text{near } \theta_0$$

converse [6] If $\mathcal{F}$ is $\mathcal{L}^1$ differentiable and $\lambda \dot{f}^2/f_0 < \infty$ is $\mathcal{F}$ also Hellinger differentiable? [Expand.]

hellinger.euclid [7] Let $\mathbb{P}_\theta$ be the probability measure defined by the density $f_\theta(\cdot)$. A simple application of the Cauchy-Schwarz inequality shows that

$$H(\mathbb{P}_\theta, \mathbb{P}_{\theta_0})^2 = (\theta-\theta_0)'\lambda\left(\dot{\xi}(x)\dot{\xi}(x)'\right)(\theta-\theta_0) + o(|\theta-\theta_0|^2).$$

Provided the matrix $\Gamma = \lambda\left(\Delta(x)\Delta(x)'\right)$ is nonsingular, it then follows that there exist nonzero constants C_1 and C_2 for which

$$C_1|\theta-\theta_0| \le H(\mathbb{P}_\theta, \mathbb{P}_{\theta_0}) \le C_2|\theta-\theta_0| \qquad \text{near } \theta_0.$$

If such a pair of inequalities holds, with fixed strictly positive constants C_1 and C_2, throughout some subset of Θ, then Hellinger distance plays the same role as ordinary Euclidean distance on that set.

bayes.hdiff [8] Suppose $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is a family of probability densities with respect to a measure λ, with index set Θ a subset of the real line. As in Theorem <20>, suppose

$$\sqrt{f_{\theta+\beta}(x)} - \sqrt{f_\theta(x)} = \int_\theta^{\theta+\beta} \Delta_t(x)\, dt \qquad \text{mod}[\lambda], \quad \text{for } |\beta| \le \delta,\ a \le \theta \le b,$$

with $\sup_t \lambda\Delta_t^2 = C < \infty$, where $[a-\delta, b+\delta] \subseteq \Theta$. Let $\mathcal{Q} = \{q_\alpha : -\delta < \alpha < \delta\}$ be a family of probability densities with respect to Lebesgue measure μ on $[a,b]$, each bounded by a fixed constant K, and with Hellinger derivative $\dot{\eta}$ at $\alpha = 0$.

Create a new family $\mathcal{P} = \{p_{\alpha,\beta}(x,\theta) : \max(|\alpha|,|\beta|) < \delta\}$ of probability densities $p_{\alpha,\beta}(x,\theta) = q_\alpha(\theta)\, f_{\theta+\beta}(x)$ with respect to $\lambda \otimes \mu$.

(i) Show that $\mathcal{P}$ is Hellinger differentiable at $\alpha = 0, \beta = 0$ with derivative having components $\dot{\eta}\sqrt{f_\theta}$ and $\sqrt{f_0}\Delta_\theta$.

(ii) Try to relax the assumptions on Ω.

[§] 7. Notes

I borrowed the exposition for the last three Sections from Pollard (1997). The essential argument is fairly standard, but the interpretation of some of the details is novel. Compare with the treatments of Le Cam (1970, and 1986 Section 17.3), Ibragimov & Has'minskii (1981, page 114), Millar (1983, page 105), Le Cam & Yang (1990, page 101), or Strasser (1985, Chapter 12).

Theorem <17> is due to Lebesgue. See Hawkins (1979), and in particular the remarks on page 145, for the long history of the Theorem and a discussion of exactly what was proved by Lebesgue. See also the footnote on page 188 of Lebesgue (1973), which apparently dates from the 1928 second edition of Lebesgue's published lectures.

Hájek (1962) used Hellinger differentiability to establish limit behaviour of rank tests for shift families of densities. Most of results in Section 4 are adapted from the Appendix to Hájek (1972), which in turn drew on Hájek & Šidák (1967, page 211) and earlier work of Hájek. For a proof of the multivariate version of Theorem <20> see Bickel et al. (1993, page 13). A reader who is puzzled about all the fuss over negligible sets, and behaviour at points where the densities vanish, might consult Le Cam (1986, pages 585–590) for a deeper discussion of the subtleties.

The proof of the information inequality (Lemma <23>) is adapted from Ibragimov & Has'minskii (1981, Section 1.7), who apparently gave credit to Blyth & Roberts (1972), but I could find no mention of Hellinger differentiability in that paper.

Cite van der Vaart (1988, Appendix A3) and Bickel et al. (1993, page 461) for Theorem <10>. Ibragimov & Has'minskii (1981, page 70) asserted that the result follows by "direct calculations". Indeed my proof uses the same truncation trick as in the proof of Lemma <23>, which is based on the argument of Ibragimov & Has'minskii (1981, page 65). Le Cam & Yang (1988, Section 7) deduced an analogous result (preservation of DQM under restriction to sub-sigma-fields) by an indirect argument using equivalence of DQM with the existence of a quadratic approximation to likelihood ratios of product measures (an LAN condition).

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, Baltimore.

Blyth, C. & Roberts, D. (1972), On inequalities of Cramér-Rao type and admissibility proofs, *in* L. Le Cam, J. Neyman & E. L. Scott, eds, 'Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 17–30.

Gill, R. & Levit, B. (1995), 'Applications of the van Trees inequality: a Bayesian Cramér-Rao bound', *Bernoulli* **1**, 59–79.

Hájek, J. (1962), 'Asymptotically most powerful rank-order tests', *Annals of Mathematical Statistics* **33**, 1124–1147.

Hájek, J. (1972), Local asymptotic minimax and admissibility in estimation, *in* L. Le Cam, J. Neyman & E. L. Scott, eds, 'Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 175–194.

Hájek, J. & Šidák, Z. (1967), *Theory of Rank Tests*, Academic Press. Also published by Academia, the Publishing House of the Czechoslavak Academy of Sciences, Prague.

Hawkins, T. (1979), *Lebesgue's Theory of Integration: Its Origins and Development*, second edn, Chelsea, New York.

Ibragimov, I. A. & Has'minskii, R. Z. (1981), *Statistical Estimation: Asymptotic Theory*, Springer, New York.

Le Cam, L. (1970), 'On the assumptions used to prove asymptotic normality of maximum likelihood estimators', *Annals of Mathematical Statistics* **41**, 802–828.

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

Le Cam, L. & Yang, G. L. (1988), 'On the preservation of local asymptotic normality under information loss', *Annals of Statistics* **16**, 483–520.

Le Cam, L. & Yang, G. L. (1990), *Asymptotics in Statistics: Some Basic Concepts*, Springer-Verlag.

Lebesgue, H. (1973), *Leçons sur l'integration et la recherche des fonctions primitives*, third edn, Chelsea, New York. First edition published in 1904 by Gauthiers-Villars, Paris.

Millar, P. W. (1983), 'The minimax principle in asymptotic statistical theory', *Springer Lecture Notes in Mathematics* pp. 75–265.

Pollard, D. (1997), Another look at differentiability in quadratic mean, *in* D. Pollard, E. Torgersen & G. L. Yang, eds, 'A Festschrift for Lucien Le Cam', Springer-Verlag, New York, pp. 305–314.

Royden, H. L. (1974), *Real and Complex Analysis*, second edn, McGraw-Hill, New York.

Strasser, H. (1985), *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*, De Gruyter, Berlin.

van der Vaart, A. (1988), *Statistical estimation in large parameter spaces*, Center for Mathematics and Computer Science.