Chapter 3

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 likelihood theory.

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

SECTION 3 shows that Hellinger differentiability of marginal densities implies existence of a local quadratic approximation to the likelihood ratio for product measures.

SECTION 4 explains why Hellinger differentiability almost implies contiguity for product measures.

SECTION 5 derives the information inequality, as an illustration of the elegance brought into statistical theory by Hellinger differentiability.

SECTION 6 discusses connections between Hellinger differentiability and pointwise differentiability of densities, leading to a sufficient condition for Hellinger differentiability.

SECTION 7 explains how one can dispense with the dominating measure for the definition of Hellinger differentiability. The slightly strengthened concept—Differentiability in Quadaratic Mean—is shown to be preserved under measurable maps.

Notation: Throughout the Chapter, $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ will denote a family of probability measures, on a fixed $(\mathcal{X}, \mathcal{A})$, indexed by a subset Θ of $\mathbb{R}^k$. In all Sections except the last, f_θ will denote the density of P_θ with respect to a fixed dominating measure λ. The function $\xi_\theta(x)$ will always denote the positive square root of $f_\theta(x)$, and $\|\cdot\|_2$ will always denote the $\mathcal{L}^2(\lambda)$ norm.

Most results in the Chapter will concern behavior near some arbitrarily chosen point θ_0 of Θ. For simplicity of notation, I will usually assume $\theta_0 = 0$, except in a few basic definitions. Thus an expression such as $(\theta - \theta_0)'\dot{\xi}_{\theta_0}$ will simplify to $\theta'\dot{\xi}_0$, a form that is easier to read and occupies less space on the page. The simplification involves no loss of theoretical generality, because the same effect could always be achieved by a reparametrization, $\theta := t + \theta_0$.

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, $\tfrac{1}{2}\exp(-|x-\theta|)$, 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 soon see, 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.

<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 $\dot{\tau}_{\theta_0}$ if $\tau(\theta) = \tau(\theta_0) + (\theta-\theta_0)'\dot{\tau}_{\theta_0} + r(\theta)$, where $\|r(\theta)\| = o(|\theta-\theta_0|)$ as $\theta \to \theta_0$. The derivative $\dot{\tau}_{\theta_0}$ is a k-vector $(v_1,\ldots,v_k)$ of elements from $\mathcal{V}$, and $t'\dot{\tau}_{\theta_0} = \sum_i t_i v_i$.*

The family $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ (dominated by λ) is said to be Hellinger differentiability at θ_0 if the map $\theta \mapsto \xi_\theta(x) := \sqrt{f_\theta(x)}$ is differentiable in $\mathcal{L}^2(\lambda)$ norm at θ_0. That is, $\mathcal{P}$ is Hellinger differentiable at θ_0 if there exists a vector $\dot{\xi}_{\theta_0}(x)$ of functions in $\mathcal{L}^2(\lambda)$ such that

<2> $$\xi_\theta(x) = \xi_{\theta_0}(x) + (\theta-\theta_0)'\dot{\xi}_{\theta_0}(x) + r_\theta(x) \qquad \text{with } \|r_\theta\|_2 = o(|\theta-\theta_0|) \text{ as } \theta \to \theta_0.$$

REMARK. Some authors (for example, Bickel, Klaassen, Ritov & Wellner (1993, page 202)) adopt a slightly different definition,

$$\sqrt{f_\theta(x)} = \sqrt{f_{\theta_0}(x)} + \tfrac{1}{2}(\theta-\theta_0)'\Delta(x)\sqrt{f_{\theta_0}(x)} + r_\theta(x),$$

replacing the Hellinger derivative $\dot{\xi}_{\theta_0}$ by $\tfrac{1}{2}\Delta(x)\sqrt{f_{\theta_0}(x)}$. As explained in Section 7, the modification very cleverly adds an extra regularity assumption to the definition. The two definitions are not completely equivalent.

Classical statistical theory, especially when dealing with independent observations from a P_θ, makes heavy use of the function $\ell_\theta(x) := \log f_\theta(x)$. The variance matrix $\mathbb{I}_\theta$ of the ***score function*** (the vector $\dot{\ell}_\theta(x)$ of partial derivatives with respect to θ) is called the ***Fisher information matrix*** for the model. The classical regularity conditions justify differentiation under the integral sign to get

<3> $$P_\theta\dot{\ell}_\theta(x) = \lambda\dot{f}_\theta(x) = \frac{\partial}{\partial\theta}\lambda f_\theta(x) = 0,$$

whence $\mathbb{I}_\theta := \mathrm{var}_\theta\left(\dot\ell_\theta\right) = P_\theta\left(\dot\ell_\theta \dot\ell_\theta'\right)$.

Under assumptions of Hellinger differentiability, the derivative $\dot\xi_\theta$ 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

$$\frac{2\dot\xi_\theta(x)}{\xi_\theta(x)} \overset{?}{=} \frac{2}{\sqrt{f_\theta(x)}}\frac{\partial}{\partial\theta}\sqrt{f_\theta(x)} = \frac{1}{f_\theta(x)}\frac{\partial f_\theta(x)}{\partial\theta} = \dot\ell_\theta(x).$$

The equality <3> corresponds to the assertion $P_\theta\left(\dot\xi_\theta/\xi_\theta\right) = \lambda\left(\xi_\theta\dot\xi\right) = 0$, which Section 2 will show to be a consequence of Hellinger differentiability and the identity $\lambda f_\theta \equiv 1$. The Fisher information $\mathbb{I}_\theta$ at θ corresponds to the matrix

$$P_{\theta_0}\left(\dot\ell_\theta\dot\ell_\theta'\right) \overset{?}{=} 4P_{\theta_0}\left(\dot\xi_\theta\dot\xi_\theta'/\xi_\theta^2\right) \overset{?}{=} 4\lambda\left(\dot\xi_\theta\dot\xi_\theta'\right).$$

Here I flag both equalities as slightly suspect, not just for the unsupported assumption of equivalence between pointwise and Hellinger differentiabilities, but also because of a possible $0/0$ cancellation. Perhaps it would be better to insert an explicit indicator function, $\{\xi_\theta > 0\}$, as a factor, to protect against $0/0$. To avoid possible ambiguity or confusion, I will write $\tilde{\mathbb{I}}_\theta$ for $4\lambda(\dot\xi_\theta\dot\xi_\theta')$ and $\tilde{\mathbb{I}}_\theta^\circ$ for $4\lambda(\dot\xi_\theta\dot\xi_\theta'\{\xi_\theta > 0\})$, to hint at equivalent forms for $\mathbb{I}_\theta$ without yet giving precise conditions under which all three exist and are equal.

The classical assumptions also justify further interchanges of integrals and derivatives, to derive an alternative representation $\mathbb{I}_\theta = -\mathbb{P}_\theta\ddot\ell_\theta$ for the information matrix. It might seem obvious that there can be no analog of this representation for 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. As shown in Section 3, the fact that $\|\xi_\theta\|_2 \equiv 1$ leads to a quadratic approximation for a log-likelihood ratio—a sort of Taylor expansion to quadratic terms without the usual assumption of twice continuous differentiability. Remarkable.

2. Differentiability of unit vectors

Suppose τ is a map from $\mathbb{R}^k$ into some inner product space $\mathcal{H}$ (such as $\mathcal{L}^2(\lambda)$). Suppose also that τ is differentiable (in norm) at θ_0,

$$\tau_\theta = \tau_{\theta_0} + (\theta-\theta_0)'\dot\tau_{\theta_0} + r_\theta \qquad \text{with } \|r_\theta\| = o(|\theta-\theta_0|) \text{ near } \theta_0.$$

For simplicity of notation, suppose $\theta_0 = 0$.

The Cauchy-Schwarz inequality gives $|\langle\tau_0, r_\theta\rangle| \le \|\tau_0\|\,\|r_\theta\| = o(|\theta|)$. It would usually be a blunder to assume naively that the bound must therefore be of order $O(|\theta|^2)$; typically, higher-order differentiability assumptions are needed to derive approximations with smaller errors. However, if $\|\tau_\theta\|$ is constant—that is, if τ_θ 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, it is easy to show that in general $\langle\tau_0, r_\theta\rangle$ equals a quadratic in θ plus an error of order $o(|\theta|^2)$. The sequential form of the assertion will be more convenient for the calculations in Section 3.

la théorème du chat mort

<4> **Lemma.** *Let $\{\alpha_n\}$ be a sequence of constants tending to zero. Let $\tau_0, \tau_1, \ldots$ be elements of norm one for which $\tau_n = \tau_0 + \alpha_n W + \rho_n$, with W a fixed element of $\mathcal{H}$ and $\|\rho_n\| = o(\alpha_n)$. Then $\langle \tau_0, W\rangle = 0$ and $2\langle \tau_0, \rho_n\rangle = -\alpha_n^2 \|W\|^2 + o(\alpha_n^2)$.*

Proof. Because both τ_n and τ_0 have unit length,

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

The $o(\alpha_n)$ and $o(\alpha_n^2)$ rates of convergence in the second and fourth lines come from the Cauchy-Schwarz inequality. The exact zero on the left-hand side of the equality exposes the leading $2\alpha_n\langle \tau_0, W\rangle$ as the only $O(\alpha_n)$ term on the right-hand side. It must be of smaller order, $o(\alpha_n)$ like the other terms, which can happen only if $\langle \tau_0, W\rangle = 0$, leaving

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

□ as asserted.

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

<5> **Corollary.** *If $\mathcal{P}$ has a Hellinger derivative $\dot{\xi}_{\theta_0}$ at 0, and if 0 is an interior point of Θ, then $\lambda\left(\xi_0 \dot{\xi}_0\right) = 0$ and $8\lambda\left(\xi_0 r_\theta\right) = -\theta' \tilde{\mathbb{I}}_0 \theta + o(|\theta|^2)$ near 0.*

Proof. Start with the second assertion, in its equivalent form for sequences $\theta_n \to 0$. Write θ_n as $|\theta_n| u_n$, with u_n a unit vector in $\mathbb{R}^k$. By a subsequencing argument, we may assume that $u_n \to u$, in which case,

$$\xi_{\theta_n} = \xi_0 + |\theta_n| u_n' \dot{\xi}_0 + r_{\theta_n} = \xi_0 + |\theta_n| u' \dot{\xi}_0 + \left(r_{\theta_n} + |\theta_n|(u_n - u)' \dot{\xi}_0\right).$$

Invoke the Lemma (with $W = u'\dot{\xi}_0$) to deduce that $u'\lambda\left(\xi_0 \dot{\xi}_0\right) = 0$ and

$$
\begin{aligned}
-4|\theta_n|^2 \lambda\left(u'\dot{\xi}_0\right)^2 + o(|\theta_n|^2) &= 8\lambda\left(\xi_0\left(r_{\theta_n} + |\theta_n|(u_n - u)'\dot{\xi}_0\right)\right)\\
&= 8\lambda\left(\xi_0 r_{\theta_n}\right) + 8|\theta_n|(u_n - u)'\lambda\left(\xi_0 \dot{\xi}_0\right).
\end{aligned}
$$

Because 0 is an interior point, for every unit vector u there are sequences $\theta_n \to 0$ through Θ for which $u = \theta_n/|\theta_n|$. Thus $u'\lambda\left(\xi_0 \dot{\xi}_0\right) = 0$ for every unit vector u, implying that $\lambda\left(\xi_0 \dot{\xi}_0\right) = 0$. The last displayed equation reduces the sequential □ analog of the asserted approximation.

REMARK. If 0 were not an interior point of the parameter space, there might not be enough directions u along which $\theta_n \to 0$ through Θ, and it might not follow that $\lambda(\xi_0 \dot{\xi}_0) = 0$. 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 0 is an interior point. See Le Cam & Yang (1988, Section 6.2) and Le Cam (1986, page 575) for further details. See also the discussion in Section 4.

3. Quadratic approximation for log likelihood ratios

Suppose observations $\{x_i\}$ are drawn independently from the distribution P_0. Under the classical regularity conditions, the log of the likelihood ratio $dP_\theta^n/dP_0^n = \prod_{i\le n} f_\theta(x_i)/f_0(x_i)$ has a local quadratic approximation in $1/\sqrt{n}$ neighborhoods of 0, under P_0^n. (Remember that, in general, $d\mathbb{Q}/d\mathbb{P}$ denotes the density with respect to $\mathbb{P}$ of the part of $\mathbb{Q}$ that is absolutely continuous with respect to $\mathbb{P}$.) For example, the following result (for one dimension) was proved in Section 2.2.

<6> **Theorem.** *Let $\mathbb{P}_n := P_0^n$ and $\mathbb{Q}_n := P_{\theta_n}^n$, for $\theta_n := \delta_n/\sqrt{n}$ with $\{\delta_n\}$ bounded. Suppose the map $\theta \mapsto f_\theta$ is twice differentiable in a neighborhood U of 0 with:*

(i) $\theta \mapsto \ddot{f}_\theta(x)$ *is continuous at* 0;

(ii) there exists a λ-integrable function $M(x)$ with $\sup_{\theta\in U} |\ddot{f}_\theta(x)| \le M(x)$ a.e. $[P_0]$;

(iii) $P_0^x\left(\dot{f}_\theta(x)/f_0(x)\right)^2 \to P_0^x\left(\dot{f}_0(x)/f_0(x)\right)^2 =: \mathbb{I}_0 < \infty$ *as* $\theta \to 0$;

(iv) $P_\theta\{f_0 = 0\} = o(\theta^2)$ *as* $\theta \to 0$.

Then $P_0\dot{\ell}_0(x) = 0 = P_0\left(\ddot{f}(x)/f_0(x)\right)$ *and, under* $\{\mathbb{P}_n\}$,

$$\frac{d\mathbb{Q}_n}{d\mathbb{P}_n} = \left(1 + o_p(1)\right)\exp\left(\delta_n Z_n - \tfrac{1}{2}\delta_n^2\mathbb{I}_0\right),$$

where $Z_n := \sum_{i\le n}\dot{\ell}_0(x_i)/\sqrt{n} \rightsquigarrow N(0,\mathbb{I}_0)$. *Consequently,* $\mathbb{Q}_n \lhd \mathbb{P}_n$.

The method of proof consisted of writing the likelihood ratio as

$$\prod_{i\le n}\left(1 + \epsilon_n(x_i)\right) \qquad \text{where } \epsilon_n(x) := \{f_0(x) > 0\}\left(f_{\theta_n}(x) - f_0(x)\right)/f_0(x),$$

then showing that, under $\mathbb{P}_n$,

(a) $\max_{i\le n}|\epsilon_n(x_i)| = o_p(1)$,

(b) $\sum_{i\le n}\epsilon_n(x_i) = \delta_n Z_n + o_p(1)$,

(c) $\sum_{i\le n}\epsilon_n(x_i)^2 = \delta_n^2\mathbb{I}_0 + o_p(1)$.

Result (a) plus the fact that $\sum_{i\le n}\epsilon_n(x_i)^2 = O_p(1)$ implied that

<7>
$$\prod_{i\le n}\left(1 + \epsilon_n(x_i)\right) = \left(1 + o_p(1)\right)\exp\left(\sum_{i\le n}\epsilon_n(x_i) - \tfrac{1}{2}\epsilon_n(x_i)^2\right),$$

from which the final assertion followed.

Le Cam (1970) established a similar quadratic approximation under an assumption of Hellinger differentiability. The method of proof is very similar to the method just outlined, but with a few very subtle differences. Remember that $\tilde{\mathbb{I}}_0 := 4\lambda\left(\dot{\xi}_\theta\dot{\xi}_\theta'\right)$ and $\tilde{\mathbb{I}}_0^\circ := 4\lambda\left(\dot{\xi}_0\dot{\xi}_0'\{\xi_\theta > 0\}\right)$.

<8> **Theorem.** *Suppose $\mathcal{P}$ is Hellinger differentiable at 0, with $\mathcal{L}^2(\lambda)$ derivative $\dot{\xi}_0$. Let $\mathbb{P}_n := P_0^n$ and $\mathbb{Q}_n := P_{\theta_n}^n$, with $\theta_n := \delta_n/\sqrt{n}$ for a bounded sequence $\{\delta_n\}$. Then, under $\{\mathbb{P}_n\}$,*

$$\frac{d\mathbb{Q}_n}{d\mathbb{P}_n} = \left(1 + o_p(1)\right)\exp\left(\delta_n' Z_n - \tfrac{1}{4}\delta_n'(\tilde{\mathbb{I}}_0 + \tilde{\mathbb{I}}_0^\circ)\delta_n\right),$$

where

$$Z_n := 2n^{-1/2}\sum_{i\le n}\{\xi_0(x_i) > 0\}\dot{\xi}_0(x_i)/\xi_0(x_i) \rightsquigarrow N(0,\tilde{\mathbb{I}}_0^\circ).$$

REMARK. It is traditional to absorb the $1 + o_p(1)$ factor for the likelihood ratio into the exponent. One then has some awkwardness with the right-hand side of the approximation at samples for which the left-hand side is zero. The awkwardness occurs with positive $\mathbb{P}_n$ probability if $P_{\theta_0}\{f_{\theta_n} = 0\} > 0$.

Proof. I will give the proof only for the one-dimensional case. The proof for the multi-dimensional case is analogous.

Write τ_n for ξ_{θ_n}, and ρ_n for r_{θ_n}, and L_n for $d\mathbb{Q}_n/d\mathbb{P}_n$. By Hellinger differentiability,

$$\tau_n(x) = \xi_0(x) + n^{-1/2}\delta_n\dot{\xi}_0(x) + \rho_n(x) \qquad \text{with } \lambda\rho_n^2 = o(\theta_n^2).$$

Define

<9> $$\eta_n(x) := \{\xi_0(x) > 0\}\frac{\tau_n(x) - \xi_0(x)}{\xi_0(x)} = \frac{\delta_n}{\sqrt{n}}D(x) + R_n(x),$$

where

$$D(x) := \{\xi_0(x) > 0\}\dot{\xi}_0(x)/\xi_0(x) \qquad \text{and} \qquad R_n(x) := \{\xi_0(x) > 0\}\rho_n(x)/\xi_0(x).$$

The indicator functions have no effect within the set $A_n := \cap_{i\le n}\{\xi_0(x_i) > 0\}$, which has $\mathbb{P}_n$-probability one, but they will protect against $0/0 \overset{?}{=} 1$ when converting from P_0- to λ-integrals. On the set A_n ,

$$\sqrt{L_n} = \prod_{i\le n}\tau_n(x_i)/\xi_0(x_i) = \prod_{i\le n}\left(1 + \eta_n(x_i)\right).$$

For almost the same reason as in the proof of Theorem <6>, we need to show that

(i) $\max_{i\le n}|\eta_n(x_i)| = o_p(1)$,

(ii) $\sum_{i\le n}\eta_n(x_i) = \frac{1}{2}\delta_n Z_n - \frac{1}{8}\delta_n^2\tilde{\mathbb{I}}_0 + o_p(1)$,

(iii) $\sum_{i\le n}\eta_n(x_i)^2 = \frac{1}{4}\delta_n^2\tilde{\mathbb{I}}_0^\circ + o_p(1)$.

The analog of <7>, with η_n replacing ϵ_n, will then give

$$\sqrt{L_n} = \left(1 + o_p(1)\right)\exp\left(\sum_{i\le n}\eta_n(x_i) - \frac{1}{2}\sum_{i\le n}\eta_n(x_i)^2\right),$$

from which the assertion of the Theorem follows by squaring both sides.

REMARK. Notice that (ii) differs significantly from its analog (b) for the proof of Theorem <6>, through the addition of a constant term. However, the difference is compensated by a halving of the corresponding constant in (iii), as compared with (c). The differences occur because, on the set $\{f_0(x) > 0\}$,

$$\epsilon_n(x) = \frac{\tau_n(x)^2 - \xi_0(x)^2}{\xi_0(x)^2} = \frac{\tau_n(x) - \xi_0(x)}{\xi_0(x)}\frac{2\xi_0(x) + \tau_n(x) - \xi_0(x)}{\xi_0(x)} = 2\eta_n(x) + \eta_n(x)^2.$$

Thus

$$\sum_{i\le n}\epsilon_n(x_i) = 2\sum_{i\le n}\eta_n(x_i) + \sum_{i\le n}\eta_n(x_i)^2 = \delta_n Z_n - \tfrac{1}{4}\delta_n^2\tilde{\mathbb{I}}_0 + \tfrac{1}{4}\tilde{\mathbb{I}}_0^\circ + o_p(1).$$

As you will see in Section 4, the conditions of Theorem <6> actually imply $\tilde{\mathbb{I}}_0 = \tilde{\mathbb{I}}_0^\circ$, a condition equivalent to the contiguity $\mathbb{Q}_n \lhd \mathbb{P}_n$.

Assertions (i), (ii), and (iii) will follow from <9>, via simple probability facts, including: if $Y_1, Y_2, \ldots$ are independent, identically distributed random variables with $\mathbb{P}|Y_1|^r < \infty$ for some constant $r \ge 1$ then $\max_{i\le n}|Y_i| = o_p(n^{1/r})$. (The proof appeared as a Problem to Chapter 2.)

First note that

$$P_0 D(x) = \lambda\left(\xi_0(x)^2 \frac{\dot\xi_0(x)}{\xi_0(x)}\{\xi_0(x) > 0\}\right) = \lambda\left(\xi_0\dot\xi_0\right) = 0 \qquad \text{by Corollary <5>,}$$

$$P_0 D(x)^2 = \lambda\left(\xi_0(x)^2 \frac{\dot\xi_0(x)^2}{\xi_0(x)^2}\{\xi_0(x) > 0\}\right) = \lambda\left(\dot\xi_0^2\{\xi_0(x) > 0\}\right) = \tfrac{1}{4}\tilde{\mathbb{I}}_0^\circ,$$

$$8P_0 R(x) = 8\lambda\left(\xi_0(x)\rho_n(x)\right) = -\delta_n^2\tilde{\mathbb{I}}_0/n + o(1/n),$$

$$P_0 R(x)^2 \le \lambda\rho_n(x)^2 = o(1/n).$$

From the expressions involving D we get

$$Z_n = 2\sum\nolimits_{i\le n} D(x_i)/\sqrt{n} \rightsquigarrow N(0, \tfrac{1}{4}\tilde{\mathbb{I}}^\circ),$$

$$n^{-1}\sum\nolimits_{i\le n} D(x_i)^2 = \tfrac{1}{4}\tilde{\mathbb{I}}^\circ + o_p(1),$$

$$\max_{i\le n}|D(x_i)| = o_p(n^{1/2}).$$

From the expressions involving R_n we get

$$\mathbb{P}_n\left(\sum\nolimits_{i\le n} R_n(x_i)\right) = -\delta_n^2\tilde{\mathbb{I}}_0 + o(1),$$

$$\operatorname{var}\left(\sum\nolimits_{i\le n} R_n(x_i)\right) \le \sum\nolimits_{i\le n}\mathbb{P}_n R_n(x_i)^2 \to 0,$$

which together imply that

$$\sum\nolimits_{i\le n} R(x_i) = -\tfrac{1}{8}\delta_n^2\tilde{\mathbb{I}}_0 + o_p(1),$$

$$\left(\max\nolimits_{i\le n}|R_n(x_i)|\right)^2 \le \sum\nolimits_{i\le n} R_n(x_i)^2 = o_p(1).$$

Assertions (i), (ii), and (iii) now follow easily.
For (i):

$$\max\nolimits_{i\le n}|\eta_n(x_i)| \le |\delta_n|\max\nolimits_{i\le n}\frac{|D(x_i)|}{\sqrt{n}} + \max\nolimits_{i\le n}|R_n(x_i)| = o_p(1).$$

For (ii):

$$\sum\nolimits_{i\le n}\eta_n(x_i) = \tfrac{1}{2}\delta_n\sum\nolimits_{i\le n}\frac{D(x_i)}{\sqrt{n}} + \sum\nolimits_{i\le n} R_n(x_i) = \tfrac{1}{2}\delta_n Z_n - \tfrac{1}{8}\delta_n^2\tilde{\mathbb{I}}_0 + o_p(1).$$

For (iii):

$$\left|\left(\sum\nolimits_{i\le n}\eta_n(x_i)^2\right)^{1/2} - \left(\delta_n^2\sum\nolimits_{i\le n}\frac{D(x_i)^2}{n}\right)^{1/2}\right| \le \left(\sum\nolimits_{i\le n} R_n(x_i)^2\right)^{1/2} = o_p(1),$$

implying that

$$\sum\nolimits_{i\le n}\eta_n(x_i)^2 = \delta_n^2\sum\nolimits_{i\le n}\frac{D(x_i)^2}{n} + o_p(1) = \tfrac{1}{4}\delta_n^2\tilde{\mathbb{I}}_0^\circ + o_p(1).$$

□ The asserted quadratic approximation follows.

4. Contiguity

Consider once more the product measures $\mathbb{P}_n := P_0^n$ and $\mathbb{Q}_n := P_{\theta_n}^n$, as defined in Theorem <8>, under the assumption of Hellinger differentiability at $\theta = 0$. For contiguity we need $\mathbb{P}L = 1$ for every limit in distribution of a subsequence of $L_n := d\mathbb{Q}_n/d\mathbb{P}_n$. Along a further subsequence $\delta_n \to \delta \in \mathbb{R}^k$, so we must have L of the form

$$\exp\left(\delta' Z - \tfrac{1}{4}\delta'\left(\tilde{\mathbb{I}}_0 + \tilde{\mathbb{I}}_0^\circ\right)\delta\right) \qquad \text{with } Z \text{ distributed } N(0, \tilde{\mathbb{I}}_0^\circ).$$

This random variable has expected value equal to 1 if and only if

$$\delta'\tilde{\mathbb{I}}_0\delta = \tfrac{1}{2}\delta'\left(\tilde{\mathbb{I}}_0 + \tilde{\mathbb{I}}_0^\circ\right)\delta,$$

which rearranges to the condition

$$\lambda\left((\delta'\dot{\xi}_0)^2\{\xi_0 = 0\}\right) = 0$$

or, equivalently $(\delta'\dot{\xi}_0)\{\xi_0 = 0\} = 0$ a.e. $[\lambda]$.

We also have a necessary condition for contiguity, from Section 2.2, namely, $nP_{\theta_n}\{f_0 = 0\} = o(1/n)$. With Hellinger differentiability, we have another way to express this condition. If $\theta_n := \delta_n/\sqrt{n}$ then $\xi_{\theta_n}(x) = n^{-1/2}\delta_n\dot{\xi}_0(x) + r_{\theta_n}(x)$ when $\xi(x) = 0$, so that

$$nP_{\theta_n}\{f_0 = 0\} = n\lambda\left(\xi_{\theta_n}^2\{\xi_0 = 0\}\right) = \delta_n'\lambda\left(\dot{\xi}_0\dot{\xi}_0'\{\xi_0 = 0\}\right)\delta_n + o(1).$$

If $\delta_n \to \delta$, the necessary condition for contiguity becomes $\delta'\dot{\xi}_0\{\xi_0 = 0\} = 0$ a.e. $[\lambda]$. Putting the two arguments together we get a neater form of Theorem <8>.

<10> **Corollary.** *If $\mathcal{P}$ is Hellinger differentiable at 0, and if $\theta_n = \delta_n/\sqrt{n}$, with $\{\delta_n\}$ bounded, then $P_{\theta_n}^n \lhd P_0^n$ if and only if $nP_{\theta_n}\{f_0 = 0\} \to 0$. In that case,*

$$dP_{\theta_n}^n/dP_0^n = \left(1 + o_p(1;\mathbb{P}_n)\right)\exp\left(\delta_n' Z_n - \tfrac{1}{2}\delta_n'\tilde{\mathbb{I}}_0\delta_n\right).$$

Hellinger differentiability alone does not imply contiguity, as shown by a simple counterexample.

<11> **Example.** Define $\mathcal{P} := \{P_\theta : 0 \le \theta \le 1\}$ via the densities

$$f_\theta(x) = \xi_\theta(x)^2 := (1-\theta^2)\,(1-|x|)^+ + \theta^2\,(1-|x-2|)^+$$

with respect to Lebesgue measure λ on $[-1, 3]$, The densities f_0 and f_1 have disjoint support, and $\xi_\theta = \left(1-\theta^2\right)^{1/2}\xi_0 + \theta\xi_1$. By direct calculation

$$\lambda\,|\xi_\theta(x) - \xi_0(x) - \theta\xi_1(x)|^2 = \left(\sqrt{1-\theta^2} - 1\right)^2 = O(\theta^4).$$

Thus $\mathcal{P}$ is Hellinger differentiable at $\theta = 0$ with $\mathcal{L}^2(\lambda)$ derivative $\dot{\xi}_0 := \sqrt{f_1}$, but $P_\theta\{f_0 = 0\} = \theta^2$. The random variable Z_n is equal to zero a.e. $[P_0^n]$, and $\tilde{\mathbb{I}}_0 = 1$, and $\tilde{\mathbb{I}}_0 = 0$. What happens to the likelihhood ratio when $\theta = \delta/\sqrt{n}$? □

You might suspect that the extreme behavior in the previous Example is caused by the fact that the parameter value 0 lies on the boundary of Θ. That suspicion would be well founded. At interior points the nonnegativity of the density forces the $\mathcal{L}^2(\lambda)$ to behave well, leading to a sistuation where Corollary <10> applies.

<12> **Corollary.** *Let $\mathcal{P}$ be Hellinger differentiable at 0, with $\mathcal{L}^2(\lambda)$ derivative $\dot{\xi}_0$. If $0 \in \text{int}(\Theta)$, then $\dot{\xi}_0\{\xi_0 = 0\} = 0$ a.e. $[\lambda]$, which implies that $P_{\theta_n}\{f_0 = 0\} = o(1/n)$ if $\theta_n := \delta_n/\sqrt{n}$, with $\{\delta_n\}$ bounded.*

Proof. Let u be a unit vector in $\mathbb{R}^k$. Consider $\theta_n := \alpha_n u$, with α_n decreasing to zero so fast that $\sum_n \|r_{\theta_n}\|_2/\alpha_n < \infty$, which implies $r_{\theta_n}(x)/\alpha_n \to 0$ a.e. $[\lambda]$. We can then draw the pointwise conclusion $u'\dot\xi(x)\{\xi_0(x) = 0\} \ge 0$ a.e. $[\lambda]$ from the inequality

$$0 \le \alpha_n^{-1}\xi_{\theta_n}(x)\{\xi_0(x) = 0\} = u'\dot\xi(x)\{\xi_0(x) = 0\} + r_{\theta_n}(x)/\alpha_n \qquad \text{a.e. } [\lambda].$$

□ The conclusion holds for every unit vector. The assertion of the Corollary follows.

5. Information inequality

The information inequality for the model $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ bounds the variance of an estimator $T(x)$ from below by an expression involving the expected value of the statistic and the Fisher information: under suitable regularity conditions,

$$\text{var}_\theta(T) \ge \dot\gamma_\theta' \mathbb{I}_\theta^{-1}\dot\gamma_\theta \qquad \text{where } \gamma_\theta := P_\theta T(x).$$

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 are captured by the following Lemma. Once again, with no loss of generality I consider only behavior at $\theta = 0$.

REMARK. The measure P_θ might itself be a product measure, representing the joint distribution of a sample of independent observations from some distribution μ_θ. As shown by Problem [4], Hellinger differentiability of $\theta \mapsto \mu_\theta$ at $\theta = 0$ would then imply Hellinger differentiability of $\theta \mapsto P_\theta$ at $\theta = 0$. We could substitute an explicit product measure for P_θ in the next Lemma, but there would be no advantage to doing so.

<13> **Lemma.** *Suppose $\mathcal{P}$ is Hellinger differentiable at 0 with $\mathcal{L}^2(\lambda)$ derivative $\dot\xi_0$. Suppose $\sup_{\theta\in U} P_\theta T(x)^2 < \infty$, for some neighborhood U of 0. Then the expected value, $\gamma_\theta := P_\theta^x T(x)$, has derivative $\dot\gamma_0 = 2\lambda(\xi_0\dot\xi_0 T)$ at 0.*

REMARK. Notice that $P_\theta T$ is well defined throughout U, because of the bound on the second moment. Also $\left(\lambda|\xi_0\dot\xi_0 T|\right)^2 \le \left(\lambda\xi_0^2 T^2\right)\left(\lambda|\dot\xi_0|^2\right) < \infty$.

Proof. Write C^2 for $\sup_{\theta\in U} P_\theta T(x)^2$, so that $\|\xi_\theta T\|_2 \le C$ for each θ in U. For simplicity, I consider only the one-dimensional case. The proof for $\mathbb{R}^k$ differs only notationally.

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

<14>
$$
\begin{aligned}
|\gamma_\theta - \gamma_0 - 2\theta\lambda(\xi_0\dot\xi_0 T)| &= |\lambda\left(\xi_\theta^2 - \xi_0^2 - 2\theta\xi_0\dot\xi_0\right) T| \\
&= \lambda\left|\theta^2\dot\xi_0^2 + r_\theta^2 + 2\xi_0 r_\theta + 2\theta\dot\xi_0 r_\theta\right| |T| \\
&\le K\theta^2\|\dot\xi_0\|_2^2 + K\|r_\theta\|_2^2 \\
&\qquad + 2K\left(\|\xi_0\|_2\,\|r_\theta\|_2 + |\theta|\|\dot\xi_0\|_2\,\|r_\theta\|_2\right) \qquad \text{by Cauchy-Schwarz} \\
&= o(|\theta|).
\end{aligned}
$$

Notice that K need not be fixed for the last conclusion. It would suffice if we had $|T| \le K_\theta = o(1/|\theta|)$, which suggests a truncation argument to handle the case of unbounded T. Let K_θ increase to ∞ as $\theta \to 0$, in such a way that $|\theta|K_\theta \to 0$. The contributions to the remainder from $T\{|T| \le K_\theta\}$ are of order $o(|\theta|)$. To complete the proof we have only to show that

<15>
$$
\lambda(\xi_\theta^2 - \xi_0^2)T\{|T| > K_\theta\} - 2\theta\lambda\left(\xi_0\dot\xi_0 T\{|T| > K_\theta\}\right) = o(|\theta|).
$$

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

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

the $o(1)$ term on the right-hand side coming via Dominated Convergence and the λ-integrability of $\dot\xi_0^2$. For the first term on the left-hand side of <15>, factorize $\xi_\theta^2 - \xi_0^2$ as $(\theta\dot\xi_0 + r_\theta)(\xi_\theta + \xi_0)$ then expand, to get

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

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 contribution to the remainder is of order $o(|\theta|)$, as required for differentiability. □

Remember from Section 1 that $4\lambda(\dot\xi_0\dot\xi_0')$ corresponds to the Fisher information matrix $\tilde{\mathbb{I}}_0$.

<16> **Corollary.** *In addition to the conditions of the Lemma, suppose $\tilde{\mathbb{I}}_0$ is nonsingular. Then* $\mathrm{var}_{\theta_0} T \ge \dot\gamma_0\tilde{\mathbb{I}}_0^{-1}\dot\gamma_0$.

interior point?

Proof. The special case where $T \equiv 1$ gives $\lambda(\xi_0\dot\xi_0) = 0$ (or use Lemma <4>). Let α be a fixed vector in $\mathbb{R}^k$. From Lemma <13> deduce that

$$
\begin{aligned}
(\alpha'\dot\gamma_0)^2 &= 4\left(\lambda\left(\alpha'\dot\xi_0\right)(T - \gamma_0)\xi_0\right)^2 \\
&\le 4\alpha'\lambda(\dot\xi_0\dot\xi_0')\alpha\,\lambda\left(\xi_0^2(T - \gamma_0)^2\right) \qquad \text{by Cauchy-Schwarz} \\
&= \alpha'\tilde{\mathbb{I}}_0\alpha\mathbb{P}_{\theta_0}(T - \gamma_0)^2
\end{aligned}
$$

Choose $\alpha := \tilde{\mathbb{I}}_0^{-1}\dot\gamma_0$ to complete the proof. □

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

<17> **Example.** Van Trees inequality—needs to be reworked.

$\vdots$

The information inequality for the one-parameter family takes an elegant form,

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

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. A sufficient condition for Hellinger differentiability

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

Consider the case where Θ is one-dimensional, with θ as an interior point. Suppose $\mathcal{P}$ is hellinger differentiable at 0, with $\mathcal{L}^2(\lambda)$ derivative $\dot{\xi}_0$. That is,

$$\xi_\theta(x) = \xi_0(x) + \theta\dot{\xi}_0(x) + r_\theta(x) \qquad \text{with } \|r_\theta\|_2 = o(|\theta|).$$

If a sequence $\{\theta_n\}$ tends to zero fast enough, then $\sum_n \|r_{\theta_n}\|_2/|\theta_n| < \infty$, from which it follows that $|r_{\theta_n}(x)| = o(|\theta_n|)$ a.e. $[\lambda]$. Unfortunately the aberrant negligible set of x might depend on $\{\theta_n\}$, so we cannot immediately invoke the usual subsequencing argument to deduce that $|r_\theta(x)| = o(|\theta|)$ a.e. $[\lambda]$. That is, it does not follow immediately that $\theta \mapsto \xi_\theta(x)$ is differentiable at $\theta = 0$ for λ-almost all x. However, if by some means we can show that the pointwise derivative $\xi_0'(x)$ does exist then we must have $\dot{\xi}(x) = \xi_0'(x)$ a.e. $[\lambda]$.

For example, if $\theta \mapsto f_\theta(x)$ has derivative $f_0'(x)$ at $\theta = 0$, and if $f_\theta(x) > 0$, then $2\xi_0'(x) = f_0'(x)/\xi_0(x)$. At points x where $f_0(x) = 0$, both derivatives $f_0'(x)$ and $\xi_0'(x)$, if they exist, must be zero, for otherwise $f_\theta(x)$ or $\xi_\theta(x)$ would be strictly negative for some small θ, either positive or negative. Thus, if the pointwise derivatives exists then $\frac{1}{2}f_0'(x)\{\xi_0(x) > 0\}/\xi_0(x)$ is, up to a λ-equivalence, the only candidate for a Hellinger derivative at $\theta = 0$.

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.

<18> **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 nonoverlapping 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 following connection between absolute continuity and integration of derivatives is one of the most celebrated results of classical analysis (UGMTP §3.4).

<19> **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 $H'(t)$ exists at Lebesgue almost all points of $[a,b]$.

(ii) The derivative H' is Lebesgue integrable

(iii) $H(t) - H(a) = \int_a^t H'(s)\,ds$ for each t 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

<20> $$H(t) = \int_a^t h(s)\,ds \qquad \text{for all } t \text{ in } [a,b]$$

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

$$\dot{H}(t) = \begin{cases} H'(t) & \text{at points } t \text{ where the derivative exists,} \\ 0 & \text{elsewhere.} \end{cases}$$

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

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

<21> **Lemma.** *Suppose a nonnegative function H is absolutely continuous on an interval $[a,b]$, with density $\dot{H}$. Let $\Delta(t) := {}^1\!/_2\dot{H}(t)\{H(t) > 0\}/\sqrt{H(t)}$. If $\int_a^b |\Delta(t)|\,dx < \infty$ then $\sqrt{H}$ is absolutely continuous, with density Δ, that is,*

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

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 <18>, 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 <19>, for each t in $[a,b]$,

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

As η decreases to zero, the left-hand side converges to $\sqrt{H(t)} - \sqrt{H(a)}$. The integrand on the right-hand side converges to $\Delta(s)$ at points where $H(s) > 0$. For almost all s in $\{H = 0\}$ the derivative $H'(s)$ exists and equals zero; the integrand converges to $0 = \Delta(s)$ at those points. By Dominated Convergence, the right-hand side converges to $\int_a^t \Delta(s)\,ds$. □

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

<22> **Theorem.** *Suppose $\mathcal{P} = \{P_\theta(x) : |\theta| < \delta\}$ for some $\delta > 0$, with each P_θ dominated by a sigma-finite measure λ. Suppose also that*

- *(i) there exist densities such that $(x, \theta) \mapsto f_\theta(x)$ is product measurable;*
- *(ii) for λ almost all x, the function $\theta \mapsto f_\theta(x)$ is absolutely continuous on $[-\delta, \delta]$, with density $\dot{f}_\theta(x)$;*
- *(iii) for λ almost all x, the function $\theta \mapsto f_\theta(x)$ is differentiable at $\theta = 0$;*
- *(iv) for each θ the function $\Delta_\theta(x) := \frac{1}{2}\dot{f}_\theta(x)\{f_\theta(x) > 0\}/\sqrt{f_\theta(x)}$ belongs to $\mathcal{L}^2(\lambda)$ and $\lambda\Delta_\theta^2 \to \lambda\Delta_0^2$ as $\theta \to 0$.*

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

> REMARK. Assumption (iii) might appear redundant, because (ii) 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 (iii).

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. Write $\mathfrak{m}$ for Lebesgue measure on $[-\delta, \delta]$

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 Tonelli, $\infty > \mathfrak{m}^s\lambda^x\Delta_s(x)^2 = \lambda^x\mathfrak{m}^s\Delta_s(x)^2$, implying $\mathfrak{m}^s\Delta_s(x)^2 < \infty$ a.e. $[\lambda]$. From Lemma <21> it then follows that

$$\frac{\xi_\theta(x) - \xi_0(x)}{\theta} = \frac{1}{\theta}\int_0^\theta \Delta_s(x)\,ds \qquad \text{a.e. } [\lambda].$$

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

<23>
$$\lambda\left|\frac{\xi_\theta(x) - \xi_0(x)}{\theta}\right|^2 \le \frac{1}{\theta}\int_0^\theta \lambda\Delta_s(x)^2\,ds \to \lambda\Delta_0^2 \qquad \text{as } \theta \to 0.$$

Define nonnegative, measurable functions

$$g_\theta(x) := 2\,|\xi_\theta(x) - \xi_0(x)|^2/\theta^2 + 2\Delta_0(x)^2 - |r_\theta(x)/\theta|^2 .$$

By (iii), $r_\theta(x)/\theta \to 0$ at almost all x where $\xi_0(x) > 0$, and hence $g_\theta(x) \to 4\Delta_0(x)^2$; and $\Delta_0(x) = 0$ when $\xi_0(x) = 0$. Thus $\liminf g_\theta(x) \ge 4\Delta_0(x)^2$ a.e. $[\lambda]$. By Fatou's Lemma (applied along subsequences), followed by an appeal to <23>,

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

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

<24> **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 P_θ to have density $f_\theta(x) := f(x-\theta)$ with respect to λ, for each θ in $\mathbb{R}$. The conditions of Theorem <22> are satisfied, with $\lambda\Delta_\theta^2 \equiv \mathbb{I}$. The family $\{P_\theta : \theta \in \mathbb{R}\}$ is Hellinger differentiable at $\theta = 0$. In fact, the same argument works at every θ; the family is everywhere Hellinger differentiable. □

7. An intrinsic characterization of Hellinger differentiability

For the definition 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. In fact, there is really no need for a single dominating λ, provided we guard against contributions from $P_\theta^\perp$, the part of P_θ that is singular with respect to P_0. As you saw in Section 4, the assumption $P_\theta^\perp \mathcal{X} = o(|\theta|^2)$ is needed to ensure contiguity. We lose little by building the assumption into the definition. Following Le Cam & Yang (2000, Section 7.2), I will call the slightly stronger property ***differentiability in quadratic mean (DQM)***, to stress that the definition requires a little more than Hellinger differentiability.

The definition makes no assumption that the family of probability measures $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ is dominated. Instead it is expressed directly in terms of the Lebesgue decomposition of P_θ with respect to P_{θ_0}, for a fixed θ_0 in Θ. As before, I will assume $\theta_0 = 0$ to simplify notation. Remember that $P_\theta = \tilde{P}_\theta + P_\theta^\perp$, where the absolutely continuous part $\tilde{P}_\theta$ has a density p_θ with respect to P_0 and the singular part $P_\theta^\perp$ concentrates on a P_0-negligible set $\mathcal{N}_\theta$,

$$P_\theta g = \tilde{P}_\theta\left(g(x)p_\theta(x)\{x \in \mathcal{N}_\theta^c\}\right) + P_\theta^\perp\left(g(x)\{x \in \mathcal{N}_\theta\}\right),$$

at least for nonnegative measurable functions g on $\mathcal{X}$.

<25> **Definition.** *Say that $\mathcal{P}$ is differentiability in quadratic mean (DQM) at 0 if*

(i) $P_\theta^\perp(\mathcal{X}) = o(|\theta|^2)$ *as* $|\theta| \to 0$*,*

(ii) there is a vector Δ of k functions from $\mathcal{L}^2(P_0)$ for which

$$\sqrt{p_\theta(x)} = 1 + \tfrac{1}{2}\theta'\Delta(x) + r_\theta(x) \qquad \text{with } P_0\left(r_\theta^2\right) = o(|\theta|^2) \text{ near } 0.$$

REMARK. Some authors (for example, Bickel et al. 1993, page 457) use the term DQM as a synonym for differentiability in L^2 norm. The factor of $1/2$ simplifies some calculations, by making the vector Δ correspond to the ***score function*** at 0.

When $\mathcal{P}$ is dominated by a sigma-finite measure, the definition agrees with the definition of Hellinger differentiability under the assumption (i), which is needed for contiguity of product measures.

<26> **Theorem.** *Suppose $\mathcal{P}$ is dominated by a sigma-finite measure λ, with corresponding densities $f_\theta(x)$.*

(i) Suppose $P_\theta\{f_0 = 0\} = o(|\theta|^2)$ *and, for some vector $\dot{\xi}$ of functions in $\mathcal{L}^2(\lambda)$,*

$$\sqrt{f_\theta(x)} = \sqrt{f_0} + \theta'\dot{\xi}(x) + R_\theta(x) \qquad \text{where } \lambda\left(R_\theta^2\right) = o(|\theta|^2) \text{ with } \theta = 0.$$

Then $\mathcal{P}$ satisfies the DQM condition at 0, with $\Delta := 2\{f_0 > 0\}\dot{\xi}/\sqrt{f_0}$ and $r_\theta := \{f_0 > 0\}R_\theta/\sqrt{f_0}$.

(ii) *If $\mathcal{P}$ satisfies the DQM condition at 0 then it is also Hellinger differentiable at 0, with $\mathcal{L}^2(\lambda)$ derivative $\dot{\xi} := \frac{1}{2}\Delta\sqrt{f_0}$.*

Proof. For the Lebesgue decomposition we can take $p_\theta := \{f_0 > 0\} f_\theta/f_0$ and $\mathcal{N}_\theta := \{f_0 = 0\}$. Thus $P_\theta^\perp \mathcal{X} = \lambda f_\theta\{f_0 = 0\}$.

If $\mathcal{P}$ is Hellinger differentiability, as in (i), the

$$
\begin{aligned}
P_0\left|\sqrt{p_\theta} - 1 - \tfrac{1}{2}\theta'\Delta\right|^2 &= \lambda f_0\left|\{f_0 > 0\}\sqrt{f_\theta/f_0} - 1 - \tfrac{1}{2}\theta'\dot{\xi}\{f_0 > 0\}/\sqrt{f_0}\right|^2 \\
&= \lambda\left(\{f_0 > 0\}\left|\sqrt{f_\theta} - \sqrt{f_0} - \theta'\dot{\xi}\right|^2\right) = o(|\theta|^2).
\end{aligned}
$$

Conversely, if $\mathcal{P}$ satisfies DQM then

$$
\begin{aligned}
\lambda\left|\sqrt{f_\theta} - \sqrt{f_0} - \theta'\dot{\xi}\right|^2 &= \lambda\{f_0 = 0\}\left(\sqrt{f_\theta} - 0\right)^2 \\
&\quad + \lambda\{f_0 > 0\}\left|\sqrt{f_0 p_\theta} - \sqrt{f_0} - \tfrac{1}{2}\theta'\Delta\sqrt{f_0}\right|^2 \\
&= o(|\theta|^2) + P_0\left|\sqrt{p_\theta} - 1 - \tfrac{1}{2}\theta'\Delta\right|^2 = o(|\theta|^2).
\end{aligned}
$$

□

REMARK. The proof of the previous Theorem is almost trivial, once one realizes that contributions from $\{f_0 = 0\}$ need separate consideration. Both Δ and $\dot{\xi}$ vanish on that set. For the definition of Hellinger differentiability it is not, a priori, necessary that $\dot{\xi}\{f_0 = 0\} = 0$. Indeed, it is contributions from that term that can upset contiguity. Some authors *define* Hellinger differentiability at $\theta = 0$ to mean

$$
\lambda|\sqrt{f_\theta} - \sqrt{f_0} - \tfrac{1}{2}\theta'\Delta\sqrt{f_0}|^2 = o(|\theta|^2) \qquad \text{with } P_0|\Delta|^2 < \infty,
$$

thereby forcing the $\mathcal{L}^2(\lambda)$ derivative to vanish on $\{f_0 = 0\}$. In effect, such a definition makes contiguity for the product measures a requirement of differentiability in the $\mathcal{L}^2(\lambda)$ sense.

The definition of DQM has some advantages over the definition of Hellinger differentiability, even beyond the elimination of the dominating measure λ. For θ near zero, $p_\theta \approx 1$, a simplification that has subtle consequences, as illustrated by the next Theorem.

The result concerns preservation of the DQM property under measurable maps. Specifically, if T is a measurable map from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$) then each P_θ induces an image measure $Q_\theta := T(P_\theta)$ on $\mathcal{B}$, defined by $Q_\theta g := P_\theta^x g(Tx)$ for each g in $\mathcal{M}^+(\mathcal{Y}, \mathcal{B})$. For each h in $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$ we can also define a measure ν_h on $\mathcal{B}$ by

$$
\nu_h(g) := P_0^x\,(h(x)g(Tx)) \qquad \text{for each } g \text{ in } \mathcal{M}^+(\mathcal{Y}, \mathcal{B}).
$$

The measure ν_h is absolutely continuous with respect to Q_0, because if $Q_0 g = 0$, for a g in $\mathcal{M}^+(\mathcal{Y}, \mathcal{B})$, then $g(Tx) = 0$ a.e. $[P]$. I will denote the density $d\nu_h/dQ_0$ by $\pi_t(h)$. That is,

<27> $$
P_0^x\left(h(x)g(Tx)\right) = Q_0^t\,(g(t)\pi_t(h)) \qquad \text{for each } g \in \mathcal{M}^+(\mathcal{Y}, \mathcal{B}), \text{ and } h \in \mathcal{M}^+(\mathcal{X}, \mathcal{A}).
$$

In fact, $\pi_t(h)$ is the Kolmogorov conditional expectation, usually denoted by $P_0(h \mid T = t)$. (Compare with UGMTP §5.6.)

As a particular case, the image measure $T(\tilde{P}_\theta)$ is absolutely continuous with respect to Q_0, with density $P_0(p_\theta \mid T = t) = \pi_t(p_\theta)$. Under DQM,

$$\sqrt{\pi_t(p_\theta)} = \left(\pi_t\left(1 + \tfrac{1}{2}\theta'\Delta + r_\theta\right)^2\right)^{1/2} = \left(1 + \theta'\pi_t\Delta + \ldots\right)^{1/2} = 1 + \tfrac{1}{2}\theta'\pi_t\Delta + \ldots$$

If all the omitted terms can be ignored, in an $\mathcal{L}^2(Q_0)$ sense, then $\{T\tilde{P}_\theta : \theta \in \Theta\}$ would be Hellinger differentiable at 0, with $\mathcal{L}^2(Q_0)$-derivative $\pi_t(\Delta)$. The image of the singular parts, $T(P_\theta^\perp)$, has total mass $o(\theta|^2)$, which does not disturb the approximation.

<28> **Theorem.** *Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is DQM at 0 with score function Δ. Suppose T is a measurable map from $(\mathcal{X}, \mathcal{A})$ into $(\mathcal{Y}, \mathcal{B})$. Then $\{TP_\theta : \theta \in \Theta\}$ is DQM at 0, with score function $P_0(\Delta \mid T = t)$.*

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

Define $Q_\theta := TP_\theta$ and $\tilde{Q}_\theta := T\tilde{P}_\theta$. Write ξ_θ for $\sqrt{p_\theta}$, so that

$$\xi_\theta(x) = 1 + \tfrac{1}{2}\theta\Delta(x) + r_\theta(x) \qquad \text{with } P_0 r_\theta^2 = o(\theta^2).$$

Use a bar to denote "averaging" with respect to π_t,

$$\bar{\Delta}_\theta(t) := \pi_t(\Delta), \qquad \bar{r}_\theta(t) := \pi_t(r_\theta), \qquad \bar{\xi}_\theta(t) := \pi_t(\xi_\theta) = 1 + \tfrac{1}{2}\theta\bar{\Delta}(t) + \bar{r}_\theta(t).$$

Define conditional variances similarly,

$$\sigma_\theta^2(t) := \pi_t\left(\xi_\theta - \bar{\xi}_\theta\right)^2, \qquad J(t) := \pi_t\left(\Delta - \bar{\Delta}\right)^2, \qquad \epsilon_\theta(t)^2 := \pi_t\left(r_\theta - \bar{r}_\theta\right)^2.$$

Notice that

$$Q_0\sigma_\theta^2 \le Q_0\left(\pi_t\xi_\theta^2\right) = P_0\xi_\theta^2 = \tilde{P}_\theta\mathcal{X} \le 1,$$

and

$$Q_0 J \le Q_0\left(\pi_t\Delta^2\right) = P_0\Delta^2 < \infty,$$

and

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

The density of $\tilde{Q}_\theta$ with respect to Q_0 equals

$$\eta_\theta^2(t) := \pi_t(\xi_\theta^2) = \pi_t(\xi_\theta - \bar{\xi}_\theta)^2 + \bar{\xi}_\theta^2.$$

Thus $\delta_\theta(t) := \eta_\theta(t) - \bar{\xi}_\theta(t)$ is nonnegative, and $\left(\bar{\xi}_\theta + \delta_\theta\right)^2 = \eta_\theta^2 = \sigma_\theta^2 + \bar{\xi}_\theta^2$, implying

<29> $$2\bar{\xi}_\theta\delta_\theta + \delta_\theta^2 = \sigma_\theta^2 = \pi_t\left(\tfrac{1}{2}\theta(\Delta - \bar{\Delta}) + (r_\theta - \bar{r}_\theta)\right)^2 \le \tfrac{1}{2}\theta^2 J(t) + 2\epsilon_\theta^2(t).$$

REMARK. The cancellation of the leading constants when $\bar{\xi}_\theta$ is subtracted from ξ_θ seems to be vital to the proof. For general Hellinger differentiability, the cancellation does not occur.

DQM for the $\tilde{Q}_\theta$ measures means $Q_0\left(\eta_\theta - 1 - \frac{1}{2}\theta\bar{\Delta}\right)^2 = o(\theta^2)$. The difference $\eta_\theta - 1 - \frac{1}{2}\theta\bar{\Delta}$ equals $\delta_\delta + \bar{r}_\theta$. The $\bar{r}_\theta$ is easily handled:

$$Q_0\bar{r}_\theta^2 = Q_0\left(\pi_t r_\theta\right)^2 \le Q_0\pi_t r_\theta^2 = P_0 r_\theta^2 = o(\theta^2).$$

For δ_θ we need to argue from <29>, using the fact that $\bar{\xi}_\theta$ should be close to 1. More precisely, let M_θ be a positive constant (depending on θ) for which $M_\theta \to \infty$ and $1/2 \ge |\theta| M_\theta \to 0$ as $\theta \to 0$. Then the set

$$\Gamma_\theta := \{t : J(t) \le M_\theta, \quad |\bar{\Delta}(t)| \le M_\theta, \quad |\bar{r}_\theta| \le \tfrac{1}{4}, \quad |\epsilon_\theta| \le 1\},$$

has Q_0 measure tending to 1, and on Γ_θ,

$$\bar{\xi}_\theta(t) \ge 1 - \tfrac{1}{2}|\theta \bar{\Delta}(t)| - |\bar{r}_\theta(t)| \ge \tfrac{1}{2}.$$

Splitting the integrand according to whether t is in Γ_θ or not, and reducing the left-hand side of <29> to $2\bar{\xi}_\theta \delta_\theta \ge \delta_\theta$ in the first case and δ_θ^2 in the second, we have

$$\begin{aligned} Q_0 \delta_\theta^2 &\le Q_0 \left(\left(\tfrac{1}{2}\theta^2 J(t) + 2\epsilon_\theta^2(t) \right)^2 \{t \in \Gamma_\theta\} \right) + Q_0 \left(\tfrac{1}{2}\theta^2 J(t) + 2\epsilon_\theta^2(t) \{t \in \Gamma_\theta^c\} \right) \\ &\le 2 \left(\tfrac{1}{2}\theta^2 M_\theta \right)^2 + \tfrac{1}{2}\theta^2 Q_0 \left(J(t) \{J(t) > M_\theta\} \right) + 6 Q_0 \epsilon_\theta^2 \\ &= o(\theta^2). \end{aligned}$$

Dominated Convergence takes care of the middle term.

The part of Q_θ that is absolutely continuous with respect to Q_0 might be slightly larger than $\tilde{Q}_\theta$, because the image measure $TP_\theta^\perp$ might also make a contribution. That is, the density for the Lebesgue decomposition of Q_θ with respect to Q_0 might actually equal $\eta_\theta^2 + s_\theta$, where $s_\theta \ge 0$ and $\mathbb{Q}_0 s_\theta \le (TP_\theta^\perp)(\mathcal{Y}) = P_\theta^\perp(\mathcal{X}) = o(\theta^2)$. The contribution from s_θ gets absorbed into the remainder term, and adds further $o(\theta^2)$ terms to the bounds in the previous paragraph. The modification has an asymptotically negligible effect on the argument. The family $\{Q_\theta : \theta \in \Theta\}$ inherits DQM from the family $\{\tilde{Q}_\theta : \theta \in \Theta\}$. □

8. Problems

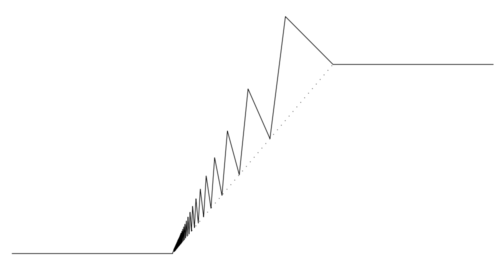
Problems not yet checked.

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

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

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

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



(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.

[2] Let $f_\theta(x) = {}^1\!/_2 \exp(-|x - \theta|)$, 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 <24>, without the explicit calculation from (i).

[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.

[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 $\sqrt{f_t(x) g_s(y)}$.

[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

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

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

[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.

[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 <22>, 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 $\mathcal{Q}$.

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

9. Notes

Incomplete

I borrowed the exposition Section 2 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).

Hájek (1962) used Hellinger differentiability to establish limit behaviour of rank tests for shift families of densities. Most of results in Section 6 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 <22> 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 <13>) 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 <28>. 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 <13>, 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.

Le Cam, L. & Yang, G. L. (2000), *Asymptotics in Statistics: Some Basic Concepts*, 2nd edn, Springer-Verlag.

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.

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. (CWI Tract 44).