A few notes on contiguity, asymptotics, and local asymptotic normality

John Duchi

March 1, 2021

Abstract

In this set of notes, I collect several ideas that are important for the asymptotic analysis of estimators. I try to put them in a framework that is relatively easy to understand, so that this can serve as a quick reference for further work. My treatment is based on a combination of Van Der Vaart's *Asymptotic Statistics* [3] and Le Cam and Yang's *Asymptotics in Statistics* [2].

Contents

Notation Here we collect notation. We will leave the underlying measure space implicit, though Ω will usually be the sample space. We let $\mathbb{B}$ be a complete real-vector space, usually finite dimensional though many of the results extend to Banach spaces. We say that a sequence of random variables $X_n \in \mathbb{B}$ converges in distribution to a random variable X_∞, written $X_n \xrightarrow{d} X_\infty$, if $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X_\infty)]$ for all continuous bounded functions $f : \mathbb{B} \to \mathbb{R}$. We write

$$X_n \xrightarrow[P_n]{d} X_\infty$$

if the laws of X_n are taken w.r.t. P_n.

We state one important and standard result before turning to our descriptions. In the theorem, we say that $P_n \xrightarrow{d} P$ if random variables $X_n \sim P_n$ satisfy $X_n \xrightarrow{d} X$, where $X \sim P$.

Theorem 1 (Portmanteau Lemma)**.** *Let $\{P_n\}_{n\in\mathbb{N}}, P$ be a sequence of measures defined on a common probability space $(\Omega, \mathcal{F})$. The following are all equivalent.*

(i) $X_n \xrightarrow{d} X$*, where* $X_n \sim P_n$ *and* $X \sim P$

(ii) For all bounded continuous f*,* $\mathbb{E}_{P_n}[f] \to \mathbb{E}_P[f]$

(iii) For all bounded, Lipschitz continuous f*,* $\mathbb{E}_{P_n}[f] \to \mathbb{E}_P[f]$

(iv) For all non-positive upper semicontinuous f*,* $\limsup_n \mathbb{E}_{P_n}[f] \le \mathbb{E}_P[F]$

(v) For all non-negative lower semicontinuous f*,* $\liminf_n \mathbb{E}_{P_n}[f] \ge \mathbb{E}_P[f]$

(vi) For all closed sets C*,* $\limsup_n P_n(C) \le P(C)$

(vii) For all open sets U*,* $\liminf_n P_n(U) \ge P(U)$

(viii) For all continuity sets of P*, that is, sets* A *such that* $\mathbb{P}(\mathrm{bd}\, A) = 0$*,* $\lim_n P_n(A) = P(A)$

1 Contiguity

In the first section, we discuss a number of results related to *contiguity*, which is essentially a way to change measure asymptotically, and builds on ideas of absolute continuity that apply in non-asymptotic situations. In brief, contiguity is useful for a number of ideas; the main two that we are concerned with are

(i) Calculations of asymptotic power for tests under alternate distributions in hypothesis testing problems, and locally worst-case alternatives

(ii) Asymptotic results on minimaxity and asymptotic normality

For now, this note studies mostly item (ii) above.

1.1 Absolute continuity

To begin with, we recall the definitions of absolute continuity. We say that a measure ν is *absolutely continuous* with respect to a measure μ, written $\nu \ll \mu$, if for all (measurable) sets A, we have

$\mu(A) = 0$ implies $\nu(A) = 0$. When this is the case, the Radon-Nikodym theorem guarantees that there is some density g such that, for any ν-integrable function f, we have

$$\int f d\nu = \int f g d\mu,$$

and we define the derivative $\frac{d\nu}{d\mu} = g$. Now, let P and Q be probability distributions asymptotically continuous with respect to a base measure μ (for example, $\mu = P + Q$ suffices). Let $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$. Then the Lebesgue decomposition thoerem states that we may decompose Q with respect to P, that is, into two parts $Q^{\parallel}$ and $Q^{\perp}$ where $Q = Q^{\parallel} + Q^{\perp}$ and

$$Q^{\parallel}(A) = Q(A \cap \{p > 0\}) \text{ and } Q^{\perp}(A) = Q(A \cap \{p = 0\}). \tag{1}$$

Then a standard result [3, Lemma 6.2] is that $Q^{\parallel} \ll P$, $Q^{\perp} \perp P$, and

$$Q^{\parallel}(A) = \int_A \frac{q}{p} dP \text{ for all measurable } A.$$

Thus, it is natural to define $\frac{dQ}{dP} = \frac{q}{p} = \frac{dQ^{\parallel}}{dP}$, which is defined P-almost surely. In addition, we have $Q \ll P$ if and only if $Q(\{p = 0\}) = 0$, which again occurs if and only if $\int \frac{q}{p} dP = 1$.

Note that absolute continuity allows changes of measure, in the sense that if $Q \ll P$, then

$$\int f dQ = \int f q d\mu = \int f \frac{q}{p} p d\mu = \int f \frac{dQ}{dP} dP,$$

because $Q^{\parallel} = Q$. More generally, if we let Ω be the sample space and $X : \Omega \to \mathbb{B}$ be a random vector, then the Q-law of X is available from the P-law of the pair $(X, dQ/dP)$ whenever $Q \ll P$, because

$$\mathbb{E}_Q[f(X)] = \mathbb{E}_P\left[f(X)\frac{dQ}{dP}\right].$$

Writing this in an evocative way for what follows, if we let M denote the joint measure (law) of the pair (X, V) where $V = \frac{dQ}{dP} \in \mathbb{R}_+$ under the distribution P, so that M is a measure on $\mathbb{B} \times \mathbb{R}_+$, then

$$Q(X \in A) = \mathbb{E}_P\left[1\{X \in A\}\frac{dQ}{dP}\right] = \mathbb{E}_P\left[1\{X \in A\} V\right] = \int_{A \times \mathbb{R}_+} v dM(x, v). \tag{2}$$

1.2 Contiguity basics

With the definitions of absolute continuity above, the next definitions (of contiguity) and results should feel somewhat familiar, as they are (roughly) asymptotic ways of defining absolute continuity. We first give a definition.

Definition 1.1. *Let P_n and Q_n be a sequence of probability measures on common probability spaces Ω_n. We say Q_n is* contiguous with respect to *P_n, written $Q_n \triangleleft P_n$, if for any sequence of sets A_n such that $P_n(A_n) \to 0$ we have $Q_n(A_n) \to 0$. We say P_n and Q_n are mutually contiguous, written $P_n \triangleleft\triangleright Q_n$, if $P_n \triangleleft Q_n$ and $Q_n \triangleleft P_n$.*

We make a few remarks on this definition before giving Le Cam's first lemma. If we let $\frac{dQ_n}{dP_n} = \frac{dQ_n^{\parallel}}{dP_n}$ as is our usual notational convention, where $Q_n = Q_n^{\parallel} + Q_n^{\perp}$, then

$$\mathbb{E}_{P_n}\left[\frac{dQ_n}{dP_n}\right] = Q_n(\{p_n > 0\}) \le 1,$$

so that the sequence $\frac{dQ_n}{dP_n} \geq 0$ is uniformly tight under the sequence of laws P_n. In particular, Prohorov's theorem guarantees that for any subsequence of $\frac{dQ_n}{dP_n}$, there is a further subsequence (denoted by $n(k)$) and random variable $V \geq 0$ such that

$$\frac{dQ_{n(k)}}{dP_{n(k)}} \xrightarrow[P_{n(k)}]{d} V.$$

The next result should be somewhat intuitive, and is known as Le Cam's first lemma. See van der Vaart [3, Lemma 6.4] for a full proof, which relies on the Portmanteau and Prohorov theorems.

Lemma 1.1 (Le Cam's first lemma)**.** *Let P_n and Q_n be sequences of probability measures on spaces Ω_n. Then the following are all equivalent.*

(i) $Q_n \triangleleft P_n$.

(ii) If $\frac{dP_n}{dQ_n} \xrightarrow{d}_{Q_n} U$ *along a subsequence, then* $\mathbb{P}(U > 0) = 1$.

(iii) If $\frac{dQ_n}{dP_n} \xrightarrow{d}_{P_n} V$ *along a subsequence, then* $\mathbb{E}[V] = 1$.

(iv) For any random variables T_n, we have $T_n \xrightarrow{p}_{P_n} 0$, then $T_n \xrightarrow{p}_{Q_n} 0$.

The result (i) if and only if (iv) is essentially definitional. For intuition that (i) and (ii) are equivalent, note that if $Q_n \triangleleft P_n$, then any "asymptotically" zero mass set under P_n must also have zero mass under Q_n, which is to say, we must have $\frac{dP_n}{dQ_n} > 0$ with Q_n probability arbitrarily near 1. Thus $U > 0$ with probability 1. Similarly, if $\frac{dP_n}{dQ_n} \xrightarrow{d}_{Q_n} U$ and $\mathbb{P}(U > 0) = 1$, then we have that there are no "asymptotic" sets with zero P_n mass but positive Q_n mass, that is, $Q_n \triangleleft P_n$. The heuristic argument for (i) iff (iii) is similar: if $Q_n \ll P_n$, then defining $V_n = \frac{dQ_n}{dP_n}$ we must have $\mathbb{E}_{P_n}[V_n] = 1$ by definition, and similarly, $Q_n(\{p_n > 0\}) = 1$ if and only if $Q_n \ll P_n$. Thus Lemma 1.1 shows that contiguity is truly an asymptotic version of absolute continuity.

Proof We only show (i) if and only if (iv). Let $T_n \xrightarrow{p} 0$ under P_n. Then if (i) holds and we define $A_n = \{|T_n| \geq \epsilon\}$, where $\epsilon > 0$ is arbitrary, we have $P_n(A_n) \to 0$ and thus $Q_n(A_n) \to 0$. That is, $T_n \xrightarrow{p} 0$ under Q_n. The converse is similarly clear: let A_n be a sequence such that $P_n(A_n) \to 0$. Define $T_n = 1$ if A_n occurs and 0 otherwise, so that $T_n \xrightarrow{p}_{P_n} 0$ and thus $T_n \xrightarrow{p}_{Q_n} 0$, whence $Q_n(A_n) \to 0$. □

The standard example application of Le Cam's first lemma is to asymptotically Gaussian measures. As we see presently, this has strong connections to asymptotic normality and optimality of verious estimation procedures.

Example 1 (Asymptotic log-normality)**:** Let us suppose that P_n and Q_n are sequences of measures such that

$$\frac{dP_n}{dQ_n} \xrightarrow[Q_n]{d} e^Z \quad \text{where} \quad Z \sim \mathsf{N}(\mu, \sigma^2).$$

Then we claim that $Q_n \triangleleft P_n$, and also that $P_n \triangleleft\triangleright Q_n$ if and only if $\mu = -\frac{1}{2}\sigma^2$.

To see this, note that we certainly have $Q_n \triangleleft P_n$ by Lemma 1.1(ii), because $e^Z > 0$ with probability 1 for $Z \sim \mathsf{N}(\mu, \sigma^2)$. Now, we have $\mathbb{E}[e^Z] = e^{\mu + \frac{1}{2}\sigma^2}$, and this is 1 if and only if $\mu = -\frac{1}{2}\sigma^2$, and thus $P_n \triangleleft Q_n$ by part (iii) of Lemma 1.1. ◇

1.3 Pairs of random variables and contiguity

As noted above, absolute continuity allows changes of measure from the pair of random variables $(X, \frac{dQ}{dP})$ under the measure P (recall expression (2)). This is also possible in an asymptotic sense, which the following version of Le Cam's Third Lemma makes precise.

Lemma 1.2. *Let $Q_n \triangleleft P_n$ and let X_n be a sequence of random variables satisfying*

$$\left(X_n, \frac{dQ_n}{dP_n}\right) \xrightarrow[P_n]{d} (X, V) \in \mathbb{B} \times \mathbb{R}_+.$$

Let M be the joint measure of the pair (X, V) on $\mathbb{B} \times \mathbb{R}_+$. Then the set function

$$L(A) := \mathbb{E}[1\{X \in A\} V] = \int_{A \times \mathbb{R}_+} v dM(x, v)$$

defines a probability measure, and $X_n \xrightarrow[Q_n]{d} Z$ for $Z \sim L$.

Proof We know that $V \geq 0$, and the joint convergence postulated in the lemma guarantees that $\frac{dQ_n}{dP_n} \xrightarrow[P_n]{d} V$. Lemma 1.1(iii) guarantees that $\mathbb{E}[V] = 1$, so the function L is certainly a probability measure. Now, let $f \geq 0$ be a lower semicontinuous function. Then by the fact that $f \geq 0$, we have by definition of dQ_n/dP_n that

$$\mathbb{E}_{Q_n}[f(X_n)] \geq \mathbb{E}_{P_n}\left[f(X_n)\frac{dQ_n}{dP_n}\right].$$

The Portmanteau lemma (Theorem 1) then implies that

$$\liminf_n \mathbb{E}_{P_n}\left[f(X_n)\frac{dQ_n}{dP_n}\right] \geq \mathbb{E}_M[f(X)V] = \int f(x) v dM(x, v) \overset{(\star)}{=} \int f(z) dL(z),$$

where equality $(\star)$ follows by definition of the measure L by integration against simple functions $1\{x \in A\}$. Thus $\liminf_n \mathbb{E}_{Q_n}[f] \geq \mathbb{E}[f(Z)]$ for $Z \sim L$, and applying the Portmanteau lemma again gives this lemma. □

The powerful consequence of Lemma 1.2 is that it allows changes of measure asymptotically, whcih has applications both asymptotically optimal estimators and tests. As the special case that will be most important to us, we state an example (simply a special case of Lemma 1.2) what is normally called Le Cam's third lemma.

Example 2 (Le Cam's Third Lemma)**:** Let P_n, Q_n be a sequence of measures and $X_n \in \mathbb{B}$ a sequence of random variables, and assume that

$$\left(X_n, \log\frac{dQ_n}{dP_n}\right) \xrightarrow[P_n]{d} \mathsf{N}\left(\begin{bmatrix}\mu \\ -\frac{1}{2}\sigma^2\end{bmatrix}, \begin{bmatrix}\Sigma & \tau \\ \tau^\top & \sigma^2\end{bmatrix}\right).$$

Then we claim that $X_n \xrightarrow[Q_n]{d} \mathsf{N}(\mu + \tau, \Sigma)$.

To see this claim, note that the continuous mapping theorem implies

$$\left(X_n, \frac{dQ_n}{dP_n}\right) \xrightarrow[P_n]{d} (X, V) \quad \text{where} \quad V = e^Z \text{ and } (X, Z) \sim \mathsf{N}\left(\begin{bmatrix}\mu \\ -\frac{1}{2}\sigma^2\end{bmatrix}, \begin{bmatrix}\Sigma & \tau \\ \tau^\top & \sigma^2\end{bmatrix}\right).$$

Thus Lemma 1.1(iii) (and the example 1 following) imply $P_n \lhd\rhd Q_n$. Making the asymptotic change of measure in Lemma 1.2, we have $X_n \xrightarrow[Q_n]{d} Y$ for a random variable Y with law L defined by $L(A) = \mathbb{E}[1\{X \in A\} e^Z]$. It thus remains to find the distribution of L, for which we use the characteristic function. Indeed, by definition of L, we have for $t \in \mathbb{B}^*$ and $i = \sqrt{-1}$ that

$$\begin{aligned}\int e^{i\langle t,y\rangle} dL(y) = \mathbb{E}\left[e^{i\langle t,X\rangle} V\right] = \mathbb{E}\left[e^{i\langle t,X\rangle} e^Z\right] &= \mathbb{E}\left[\exp\left(\left\langle \begin{bmatrix} it \\ 1 \end{bmatrix}, \begin{bmatrix} X \\ Z \end{bmatrix} \right\rangle\right)\right] \\ &= \exp\left(i\langle t, \mu\rangle - \frac{1}{2}\sigma^2 - \frac{1}{2}\begin{bmatrix} t \\ -i \end{bmatrix}^\top \begin{bmatrix} \Sigma & \tau \\ \tau^\top & \sigma^2 \end{bmatrix} \begin{bmatrix} t \\ -i \end{bmatrix}\right) \\ &= \exp\left(i\langle t, \mu + \tau\rangle - \frac{1}{2} t^\top \Sigma t\right),\end{aligned}$$

which is evidently the characteristic function of a $\mathsf{N}(\mu + \tau, \Sigma)$-distributed random vector. $\diamond$

This example is important, because it shows that if we (as is often the case) have a particular type of asymptotic normality under a sequence of null distributions P_n, then we can view the limiting random vector under alternatives Q_n as being distributed $\mathsf{N}(\mu + \tau, \Sigma)$, with the same covariance.

1.4 Distances on probability distributions

As a bit of an aside, which may help to motivate our discussion of quadratic mean differentiability, we make here a few remarks on optimal testing errors and other techniques for proving *bounds* on performance—i.e. guaranteeing that estimators or tests cannot distinguish between certain alternatives. Our starting point is a standard result on the summed probabilities of error in a hypothesis test, but first we define a few notions of distance between probability distributions.

Definition 1.2 (Distances between probability distributions)**.** *Let P and Q be probability distributions with densities p and q with respect to a measure μ. The* total variation distance *between P and Q is*

$$\|P - Q\|_{\mathrm{TV}} := \sup_A |P(A) - Q(A)|.$$

The Hellinger distance *between P and Q is defined by its square*

$$d_{\mathrm{hel}}^2(P, Q) := \frac{1}{2}\int (\sqrt{p} - \sqrt{q})^2 \, d\mu = 1 - \int \sqrt{pq} d\mu.$$

There are a number of relationships between the variation distance and Hellinger distance (some of which we explore in exercises), and we state a few as a lemma, leaving the proof to the reader.

Lemma 1.3 (Relationships between distances)**.** *Let P and Q be as in Definition 1.2. Then we have the following equalities.*

(a) $\|P - Q\|_{\mathrm{TV}} = \frac{1}{2}\int |p - q| d\mu$

(b) $\|P - Q\|_{\mathrm{TV}} = \int (p \vee q) d\mu - 1 = 1 - \int (p \wedge q) d\mu$

(c) $\sup_{\|f\|_\infty \le 1} \int f(dP - dQ) = 2\|P - Q\|_{\mathrm{TV}}$.

In addition, the Hellinger distance and variation distance satisfy

$$d_{\mathrm{hel}}^2(P, Q) \le \|P - Q\|_{\mathrm{TV}} \le d_{\mathrm{hel}}(P, Q)\sqrt{2 - d_{\mathrm{hel}}^2(P, Q)}.$$

It is often more convenient to use the Hellinger distance for i.i.d. sampling schemes, because it behaves substantially more nicely than does variation distance on product distributions. Indeed, let P^n and Q^n denote the n-fold products of P and Q. Then

$$\begin{aligned} d_{\rm hel}^2(P^n, Q^n) &= 1 - \int \sqrt{p(x_1) \cdots p(x_n)} \sqrt{q(x_1) \cdots q(x_n)} d\mu^n \\ &= 1 - \left(\int \sqrt{pq} d\mu \right)^n \\ &= 1 - \left(1 - d_{\rm hel}^2(P, Q)\right)^n , \end{aligned} \tag{3}$$

where the final inequality follows because $d_{\rm hel}^2(P, Q) = 1 - \int \sqrt{pq} d\mu$. Equality (3) makes clear that if we know $d_{\rm hel}(P, Q)$, then we can immediately calculate $d_{\rm hel}(P^n, Q^n)$.

With these definitions, we can give a few results on optimality of tests, as well as impossibility results, which motivate a few of our coming local alternatives and asymptotic calculations. First, we present a standard result, which we state as a lemma for convenient reference.

Lemma 1.4 (Le Cam)**.** *Let P_0 and P_1 be arbitrary distributions on a space $\mathcal{X}$ and consider the simple hypothesis test of P_0 against P_1. For any test $\psi : \mathcal{X} \to \{0, 1\}$, we have*

$$P_0(\psi \neq 0) + P_1(\psi \neq 1) \geq 1 - \|P_0 - P_1\|_{\rm TV} ,$$

with equality if $\psi(x) = 1\{p_1(x) \geq p_0(x)\}$.

Proof Associated with any test is a rejection region, so let $A = \{x : \psi(x) = 0\}$. Then

$$P_0(A^c) + P_1(A) = 1 - P_0(A) + P_1(A) = 1 - (P_0(A) - P_1(A)) \geq 1 - \sup_A |P_0(A) - P_1(A)|.$$

The equality is immediate from Lemma 1.3. □

If we consider a sequence of testing problems then, indexed by n, $\{P_{0,n}\}_{n \in \mathbb{N}}$ against $\{P_{1,n}\}_{n \in \mathbb{N}}$, then we see that the best possible error rate for testing $P_{1,n}$ against $P_{0,n}$ is governed by the limits of $\|P_{0,n} - P_{1,n}\|_{\rm TV}$. Now, note that the function $x \mapsto x\sqrt{2 - x^2}$ is increasing on $[0, 1]$, and increases to 1. Then by Lemma 1.3, if all the cluster points of $d_{\rm hel}(P_{0,n}, P_{1,n})$ lie in the open interval $(0, 1)$, then evidently we have

$$1 > \limsup_n \inf_\psi \{P_{0,n}(\psi \neq 0) + P_{1,n}(\psi \neq 1)\} \geq \liminf_n \inf_\psi \{P_{0,n}(\psi \neq 0) + P_{1,n}(\psi \neq 1)\} > 0.$$

That is, there is no perfect test, but there are tests whose average error rate is better than 50%.

1.5 Quadratic mean differentiability and local alternatives

As another heuristic, consider the case of local alternatives, where we compare testing the value of a parameter θ, where the null is $H_0 : \theta = \theta_0$ and we have the sequence of alternatives $H_1 : \theta = \theta_0 + h/\sqrt{n}$ for some perturbation h. If we can show that $d_{\rm hel}(P_{\theta_0}^n, P_{\theta_0 + h/\sqrt{n}}^n)$ has some limit as $n \to \infty$, then this would provide concrete bounds on the probability of error in tests, by Lemma 1.4.

To make this a bit more concrete, we consider Taylor expansions of $\sqrt{p_\theta}$ that allow us to compute such limits. First, recall that $\sqrt{a + \delta} = \sqrt{a} + \frac{\delta}{2\sqrt{a}} + O(\delta^2)$, and suppose that p_θ is sufficiently differentiable that we can write

$$p_{\theta+h} = p_\theta + \nabla_\theta p_\theta^\top h + O(\|h\|^2) \text{ and } \sqrt{p_{\theta+h}} = \sqrt{p_\theta + \nabla_\theta p_\theta^\top h + O(\|h\|^2)} = \sqrt{p_\theta} + \frac{\nabla p_\theta^\top h}{2p_\theta} \sqrt{p_\theta} + O(\|h\|^2).$$

Note that if $\ell_\theta = \log p_\theta$ is the log-likelihood, then $\dot{\ell}_\theta = \frac{\nabla p_\theta}{p_\theta}$ is the score function and we have that $\sqrt{p_{\theta+h}} = \sqrt{p_\theta} + \frac{1}{2} h^\top \dot{\ell}_\theta \sqrt{p_\theta} + O(\|h\|^2)$ as long as the derivatives exist. Then, continuing with our heuristics, we also have

$$d_{\rm hel}^2(P_\theta, P_{\theta+h}) = \frac{1}{2} \int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta}\right)^2 d\mu = \frac{1}{8} \int (\dot{\ell}_\theta^\top h)^2 p_\theta d\mu + o(\|h\|^2) = \frac{1}{8} h^\top I_\theta h + o(\|h\|^2)$$

where I_θ is the Fisher Information for the model P_θ. That is, at least heuristically, we have that

$$d_{\rm hel}^2(P_\theta, P_{\theta+h/\sqrt{n}}) = \frac{1}{8n} h^\top I_\theta h + o(\|h\|^2 / n).$$

These calculations—and the optimality in testing they imply—motivate the following definition, which makes the preceding calculations rigorous, but also allows us to require a Taylor expansion of $\sqrt{p_\theta}$ to exist only in mean-square (which is natural, because all we care about are distances between distributions).

Definition 1.3. *The family $\{P_\theta\}_{\theta \in \Theta}$ of distributions on $\mathcal{X}$ is* quadratic mean differentiable (QMD) *at $\theta \in \mathbb{R}^d$ if there exists a score function $\dot{\ell}_\theta : \mathcal{X} \to \mathbb{R}^d$ such that*

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^\top \dot{\ell}_\theta \sqrt{p_\theta}\right)^2 d\mu = o(\|h\|^2).$$

For a QMD family, we define $I_\theta = \int \dot{\ell}_\theta \dot{\ell}_\theta^\top dP_\theta$ to be the Fisher information, which follows the usual conventions for nicely structured probability distributions (such as exponential families; see van der Vaart [3, Chapters 6–8]).

A calculation then shows the following lemma, which again we leave as an exercise for the reader, but which shows that under i.i.d. sampling, it is hard to test between distributions getting close to one another at a rate of $1/\sqrt{n}$.

Lemma 1.5. *Let the family $\{P_\theta\}_{\theta \in \Theta}$ be quadratic mean differentiable and $\theta_0 \in \operatorname{int} \Theta$. Then for any $h \in \mathbb{R}^d$,*

$$\lim_{n \to \infty} d_{\rm hel}^2(P_{\theta_0}^n, P_{\theta_0 + h/\sqrt{n}}^n) = 1 - \exp\left(-\frac{1}{8} h^\top I_{\theta_0} h\right).$$

Returning to Lemma 1.4, we then see that for *any* sequence $\{\psi_n\}$ of tests, if the perturbation h is bounded, so that $\exp(-\frac{1}{8} h^\top I_{\theta_0} h) < 1$, we have for a QMD family that

$$\liminf_n \left\{P_{\theta_0}^n(\psi_n \neq 0) + P_{\theta_0 + h/\sqrt{n}}^n(\psi_n \neq 1)\right\} > 0.$$

Again, we leave verification of this as an exercise for the reader.

Quadratic mean differentiable families also generalize Fisher information in a natural way beyond pure differentiability of the likelihood function.

Proposition 1. *Let the family $\{P_\theta\}_{\theta \in \Theta}$ be quadratic mean differentiable. Then the score function $\dot{\ell}_\theta$ is mean zero, $P_\theta \dot{\ell}_\theta = 0$, and the Fisher Information $I_\theta := P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^\top$ exists.*

Proof Letting $p_n = p_{\theta + h/\sqrt{n}}$ and $p = p_\theta$ be the respective densities, and defining for simplicity $g = h^\top \dot{\ell}_\theta$, we have by the definition of QMD that

$$\int n \left(\sqrt{p_n} - \sqrt{p} - \frac{1}{2} g \sqrt{p} / \sqrt{n}\right)^2 d\mu = o(1) \quad \text{or} \quad \int \left(\sqrt{n}(\sqrt{p_n} - \sqrt{p}) - \frac{1}{2} g \sqrt{p}\right)^2 d\mu = o(1).$$

That is, we have $\sqrt{n}(\sqrt{p_n} - \sqrt{p}) \stackrel{L_2}{\to} \frac{1}{2}g$ (in $L_2(\mu)$-norm) and so $(\sqrt{p_n} - \sqrt{p}) \stackrel{L_2}{\to} 0$. Thus, we find that

$$Pg = \int gpd\mu = \int g\sqrt{p}\sqrt{p}d\mu = 2\lim_n \int \sqrt{n}(\sqrt{p_n}-\sqrt{p})\sqrt{p}d\mu = \lim_n \int \sqrt{n}(\sqrt{p_n}-\sqrt{p})(\sqrt{p}+\sqrt{p_n})d\mu,$$

the final equality a consequence of the $L_2(\mu)$ limits of $\sqrt{p_n}-\sqrt{p}$. But $(\sqrt{p_n}-\sqrt{p})(\sqrt{p}+\sqrt{p_n}) = p_n - p$, and this integrates to zero always. So $Pg = P_\theta h^\top \dot{\ell}_\theta = 0$ for all h, or $P_\theta \dot{\ell}_\theta = 0$.

The existence of $P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^\top$ is simpler. We have by the triangle inequality that

$$\begin{aligned}
o(\|h\|) &\geq \left(\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}\dot{\ell}_\theta^\top h \sqrt{p_\theta}\right)^2 d\mu\right)^{\frac{1}{2}} \\
&\geq \left(\frac{1}{4}\int (\dot{\ell}_\theta^\top h)^2 p_\theta d\mu\right)^{\frac{1}{2}} - \left(\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta}\right)^2 d\mu\right)^{\frac{1}{2}} \\
&= \left(\frac{1}{4}\int (\dot{\ell}_\theta^\top h)^2 p_\theta d\mu\right)^{\frac{1}{2}} - \sqrt{2d_{\rm hel}^2(P_{\theta+h}, P_\theta)}.
\end{aligned}$$

As $d_{\rm hel} \leq 1$ and h is arbitrary, this gives the result. □

2 Local Asymptotic Normality

Now we begin to explore concepts that build off of the results on contiguity, but which make stronger connections to normality theory. First, we begin by providing a definition of local asymptotic normality; we give a slightly stronger definition than Le Cam's classical definitions, but our efforts to make things concrete allow (to us) more intuitive and somewhat easier proofs.

In the statement of the definition, we assume there is a sequence of spaces Ξ^n, where $\xi^n \in \Xi^n$, and that for each n we have a family of probability measures $P_{\theta,n}$ indexed by $\theta \in \mathbb{R}^d$ (or $\theta \in \Theta \subset \mathbb{R}^d$). We then consider local perturbations h of θ, where the local perturbations are at a scale of $1/\sqrt{n}$, so that they become "more" local at the rate $n^{-\frac{1}{2}}$.

Definition 2.1. *The family $\{P_{\theta,n}\}$, defined for $\theta \in \Theta$ and $n \in \mathbb{N}$, is* locally asymptotically normal *at the point θ if $\theta \in \operatorname{int}\Theta$ and there exists a mapping $\Delta_n : \Xi^n \to \mathbb{R}^d$ and matrix $K \succeq 0$ such that for all $h \in \mathbb{R}^d$ and n large enough that $\theta + h/\sqrt{n} \in \Theta$,*

$$\log \frac{dP_{\theta+h/\sqrt{n},n}(\xi^n)}{dP_{\theta,n}(\xi^n)} = h^\top \Delta_n(\xi^n) - \frac{1}{2}h^\top K h + o_{P_{\theta,n}}(\|h\|)$$

and

$$\Delta_n(\xi^n) \underset{P_{\theta,n}}{\stackrel{d}{\to}} \mathsf{N}(0, K).$$

We call K the precision matrix *for the family $\{P_{\theta,n}\}$.*

In Definition 2.1, we use $o_{P_{\theta,n}}(\|h\|)$ to mean a quantity that converes to zero in $P_{\theta,n}$ probability as long as $\|h\|$ is bounded.

2.1 Examples

Example 3 (Shrinking i.i.d. Gaussian locations)**:** As a standard example of local asymptotic normality, we may consider the simple model

$$Y_i = \frac{1}{\sqrt{n}}h + \xi_i, \quad \xi_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \Sigma),$$

for $i = 1, \ldots, n$. Letting $\theta = 0$ for simplicity, we let $P_{h/\sqrt{n},n}$ be the distribution of Y_i with mean $h/\sqrt{n}$, and by inspection we have that

$$\log \frac{dP_{h,n}(\overline{Y}_n)}{dP_{0,n}(\overline{Y}_n)} = h^\top \Sigma^{-1} n^{\frac{1}{2}} \overline{Y}_n - \frac{1}{2} h^\top \Sigma^{-1} h,$$

where $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Certainly under $P_{0,n}$, we have $n^{\frac{1}{2}}\overline{Y}_n \stackrel{d}{\to} \mathsf{N}(0, \Sigma)$, and thus this model satisfies Definition 2.1 with $K = \Sigma^{-1}$ and $\Delta_n(\overline{Y}_n) = \Delta_n(Y_1, \ldots, Y_n) = \sqrt{n}\Sigma^{-1}\overline{Y}_n$. $\diamond$

Example 4 (Quadratic-mean-differentiable families)**:** Recall that a family $\{P_\theta\}_{\theta \in \Theta}$ is quadratic-mean-differentiable if each element has density p_θ with respect to some base measure μ, and there is a score function $\dot{\ell}_\theta$ such that

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}\dot{\ell}_\theta^\top h \sqrt{p_\theta}\right)^2 d\mu = o(\|h\|^2).$$

By Propoosition 1, a QMD family has a Fisher information matrix, and the score function is mean zero. We show that it is also locally asymptotically normal.

For notational simplicity, let $P = P_\theta$ and $P_n = P^n_{\theta+h/\sqrt{n}}$, that is, the base distribution P_θ and its local alternative $P_{\theta+h/\sqrt{n}}$ (under i.i.d. sampling). We claim that

$$\log \frac{dP_n(X_1, \ldots, X_n)}{dP(X_1, \ldots, X_n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n h^\top \dot{\ell}_\theta(X_i) - \frac{1}{2} h^\top I_\theta h + o_P(1), \tag{4}$$

where $I_\theta = \mathbb{E}_P[\dot{\ell}_\theta \dot{\ell}_\theta^\top]$ denotes the Fisher information. That is, QMD families under local alternatives are locally asymptotically normal with precision matrix I_θ, because $n^{-\frac{1}{2}}\sum_{i=1}^n \dot{\ell}_\theta(X_i) \stackrel{d}{\to}_{P_\theta} \mathsf{N}(0, I_\theta)$.

Let us prove the claim (4); we use a Taylor expansion of the log-likelihood and definition of QMD families to do so (our proof follows van der Vaart [3, Ch. 7]). First, we have

$$\log \prod_{i=1}^n \frac{p_n(X_i)}{p(X_i)} = \sum_{i=1}^n 2 \log \frac{\sqrt{p_n}}{\sqrt{p}}(X_i) = \sum_{i=1}^n 2 \log \left(1 + \frac{1}{2}\left(2\sqrt{\frac{p_n}{p}}(X_i) - 1\right)\right).$$

Now, define $W_{n,i} = 2\sqrt{\frac{p_n}{p}}(X_i) - 1$. Then a Taylor expansion of $\log(1+x) = x - \frac{1}{2}x^2 + x^2 r(x)$ where $\limsup_{x\to 0} |x^{-1} r(x)| < \infty$ shows that

$$\sum_{i=1}^n \log \frac{p_n}{p}(X_i) = 2\sum_{i=1}^n \log\left(1 + \frac{1}{2}W_{n,i}\right) = \sum_{i=1}^n \left[W_{n,i} - \frac{1}{4}W_{n,i}^2 + 2W_{n,i}^2 r(W_{n,i})\right]. \tag{5}$$

We consider each of these terms in turn.

For notational convenience, define $g(x) = h^\top \dot{\ell}_\theta(x)$. First, we have that

$$\mathrm{Var}_P\left(\sum_{i=1}^n W_{n,i} - n^{-\frac{1}{2}}\sum_{i=1}^n g(X_i)\right) \le n\mathbb{E}_P\left[(W_{n,1} - n^{-\frac{1}{2}} g(X_1))^2\right] = no(\|h\|^2/n) \to 0$$

as $n \to \infty$, by the definition of QMD. Thus, we can write $nW_{n,i}^2 = g^2(X_i) + E_{n,i}$ where $\mathbb{E}[|E_{n,i}|] \to 0$ as $n \to \infty$, whence $n^{-1} \sum_{i=1}^n E_{n,i} \xrightarrow{p} 0$. Moreover, we have

$$n\mathbb{E}_P[W_{n,1}] = 2n \left(\int \sqrt{p_n p} d\mu - 1 \right) = -2n d_{\rm hel}^2(p_n, p)^2 \to -\frac{1}{4} \mathbb{E}_P[g(X)^2],$$

again by the definition of QMD. In particular, $\sum_{i=1}^n W_{n,i} = n^{-\frac{1}{2}} \sum_{i=1}^n g(X_i) - \frac{1}{4} h^\top I_\theta h + o_P(1)$, because $\mathbb{E}_P[g^2] = h^\top \mathbb{E}_P[\dot{\ell}_\theta \dot{\ell}_\theta^\top] h = h^\top I_\theta h$. The law of large numbers, as a consequence of the variance bound on $\sum_{i=1}^n W_{n,i}$, implies

$$\sum_{i=1}^n W_{n,i}^2 \xrightarrow{p} P g^2 = h^\top I_\theta h.$$

The last quantity is to control the remainders in expression (5). Fix a constant $\epsilon > 0$. We have that

$$\begin{aligned} \mathbb{P}(\max_{i \le n} |W_{n,i}| \ge 2\epsilon) \le n\mathbb{P}(|W_{n,1}| \ge 2\epsilon) &\le n\mathbb{P}(g(X_i)^2 \ge n\epsilon^2) + n\mathbb{P}(|E_{n,i}| \ge n\epsilon^2) \\ &\le \epsilon^{-2} P(g^2 1\left\{g^2 \ge n\epsilon\right\}) + \epsilon^{-2} \mathbb{E}[|E_{n,i}|] \to 0 \end{aligned}$$

as $n \to \infty$, by the fact that $Pg^2 < \infty$. Thus, we have $\max_{i \le n} |W_{n,i}| = o_P(1)$ and so $\sum_{i=1}^n W_{n,i}^2 r(W_{n,i}) \le \max_i r(W_{n,i}) \sum_{i=1}^n W_{n,i}^2 = o_P(1) O_P(1) = o_P(1)$. In combination with the expression (5), this yields our desired claim (4). $\diamond$

Example 5 (Tilted distributions)**:** We may provide a somewhat more complex example than the preceding one. Suppose we have a base distribution P_0 on a set Ξ, and let the random vector $X = X(\xi) \in \mathbb{R}^d$ have mean-zero under P_0 and finite second moment, where $\Sigma = \mathbb{E}_{P_0}[XX^\top]$. Now, for $h \in \mathbb{R}^d$, define the tilted measure

$$dP_h(\xi) = \frac{\left[1 + h^\top X(\xi)\right]_+}{C_h} dP_0(\xi) \quad \text{where} \quad C_h = \int \left[1 + h^\top X\right]_+ dP_0.$$

Moreover, for $\xi^n = (\xi_1, \ldots, \xi_n) \in \Xi^n$, define the i.i.d. sampling measures

$$dP_{h/\sqrt{n},n}(\xi^n) = \prod_{i=1}^n dP_{h/\sqrt{n}}(\xi_i).$$

We claim that this family is Locally Asymptotically Normal.

First, we claim that $C_h = 1 + o(\|h\|_2^2)$. To see this, note that $C_h \ge \int (1 + h^\top X) dP_0 = 1$, and if we define $\Xi_h = \{\xi : h^\top X(\xi) \le -1\}$, then

$$\begin{aligned} C_h &= \int \left[1 + h^\top X\right]_+ dP_0 = \int (1 + h^\top X) dP_0 - \int_{\Xi_h} (1 + h^\top X) dP_0 \\ &= 1 + \mathbb{E}_{P_0}[1\{X^\top h \le -1\} |1 + h^\top X|]. \end{aligned}$$

Notably, we have that if $x^\top h \le -1$, then $|1 + x^\top h| \le |x^\top h| \le (x^\top h)^2 \le \|x\|^2 \|h\|^2$, whence

$$\frac{1}{\|h\|^2} 1\{X^\top h \le -1\} |1 + X^\top h| \le \|X\|^2 1\{X^\top h \le -1\}.$$

Lebesgue's dominated convergence theorem thus implies that

$$\limsup_{h \to 0} \frac{1}{\|h\|^2} \mathbb{E}_{P_0}[1\{X^\top h \le -1\}|1 + h^\top X|] = 0,$$

because the term inside the expectation certainly converges to 0 as $h \to 0$.

Now we note the standard result that if a vector X has k moments, that is, $\mathbb{E}[\|X\|^k] < \infty$, then this is equivalent to $\sum_{n=1}^\infty \mathbb{P}(\|X_n\|^k \ge n) < \infty$ where X_n are i.i.d. copies of X, so that

$$\max_{i \le n} n^{-\frac{1}{2}} \|X(\xi_i)\| \stackrel{a.s.}{\to}_{P_0} 0$$

by the Borel-Cantelli theorem if $\Sigma = \mathbb{E}_{P_0}[XX^\top]$ exists. Thus, for all $h \in \mathbb{R}^d$, we have (uniformly in $\|h\| \le c$ for any constant $c < \infty$) that

$$\begin{aligned}
&\log \frac{dP_{h/\sqrt{n},n}(\xi^n)}{dP_{0,n}(\xi^n)} \\
&= \sum_{i=1}^n \log \left[1 + n^{-\frac{1}{2}} h^\top X(\xi_i)\right]_+ - n \log C_{h/\sqrt{n}} \\
&= \sum_{i=1}^n \left[n^{-\frac{1}{2}} h^\top X(\xi_i) - \frac{1}{2} n^{-1} h^\top X(\xi_i) X(\xi_i)^\top h + o(n^{-1} \|h\|^2 \|X(\xi_i)\|^2)\right] - n \log \left(1 + o(n^{-1}\|h\|^2)\right) \\
&= h^\top \left(n^{-\frac{1}{2}} \sum_{i=1}^n X(\xi_i)\right) - \frac{1}{2} h^\top \Sigma h + o_{P_{0,n}}(\|h\|^2).
\end{aligned}$$

Because $n^{-\frac{1}{2}} \sum_{i=1}^n X(\xi_i) \stackrel{d}{\to}_{P_{0,n}} \mathsf{N}(0, \Sigma)$, we see that the tilted family is indeed LAN. $\diamond$

2.2 Connections to contiguity and heuristics for normality

Now, let us make a few connections between Definition 2.1 and the contiguity results of the preceding section. We assume without loss of generality that $\theta = 0$ in the remainder of our results, because everything simply shifts by this amount. Then Definition 2.1 shows that

$$\left(\Delta_n(\xi^n), \log \frac{dP_{h/\sqrt{n},n}}{dP_{0,n}}(\xi^n)\right) \stackrel{d}{\underset{P_{0,n}}{\to}} \mathsf{N}\left(\begin{bmatrix} 0 \\ -\frac{1}{2} h^\top K h \end{bmatrix}, \begin{bmatrix} K & Kh \\ (Kh)^\top & h^\top K h \end{bmatrix}\right), \tag{6}$$

because $\mathbb{E}[ZZ^\top h] = Kh$ for $Z \sim \mathsf{N}(0, K)$. By inspection, the convergence (6) is completely identical to that in Le Cam's Third Lemma, Example 2. Thus, we see that under the local alternative distributions $P_{h/\sqrt{n},n}$, we have the asymptotic shift

$$\Delta_n(\xi^n) \stackrel{d}{\underset{P_{h/\sqrt{n},n}}{\to}} \mathsf{N}(Kh, K).$$

By defining the random variables

$$Z_n(\xi^n) := K^{-1} \Delta_n(\xi^n), \tag{7}$$

we have the similar asymptotic shift

$$Z_n(\xi^n) \stackrel{d}{\underset{P_{h/\sqrt{n},n}}{\to}} \mathsf{N}(h, K^{-1}).$$

As this convergence occurs for *any* choice of the local perturbation h (because eventually, $h/\sqrt{n}$ is near enough zero that $P_{h/\sqrt{n},n}$ is well-defined), it is intuitive that there be some more general type of "limiting normality" for problems involving estimation in $P_{0,n}$. We can indeed make this very rigorous in a number of ways. But before proceeding, we simply note the following simple heuristic. Assuming that we would like to estimate the perturbation h in the sequence of local models $P_{h/\sqrt{n},n}$, we use $Z_n(\xi^n)$ to see that asymptotically, we would like to estimate the mean (location) of a single normally distributed random vector $\mathsf{N}(h, K^{-1})$.

3 Limiting Experiments and Posterior Distributions under Local Asymptotic Normality

With the preliminaries and definitions of local asymptotic normality presented, we now make rigorous the heuristic that the local alternatives $P_{h/\sqrt{n},n}$ should somehow result in asymptotically "normal" estimation problems, or asymptotically normal distributions. We give two approaches, one the more classical approach due to Le Cam (cf. [2]), which actually constructs asymptotic posterior distributions for the parameter h in various Bayesian models, and the other due to Van Der Vaart, which shows how the limits are a type of normal location experiment.

3.1 Asymptotic distributions—the Le Cam approach

Our first set of results study local asymptotic normality by providing explicit limiting distributions and estimates of the posterior on h in various Bayesian settings, which we can then transform into optimality guarantees.

We begin with a lemma, which is a simplification of a result due to Le Cam and Yang [2], to allow more explicit distributional calculations.

Lemma 3.1 (Le Cam and Yang [2], Proposition 6.3.2)**.** *Let Z_n be defined as in expression (7). Fix $c \in (0, \infty)$ and $\epsilon > 0$, and define*

$$A_{n,b} := \{\xi^n \in \Xi^n : \|Z_n(\xi^n)\| \le b\}.$$

There exist $B = B(c, \epsilon)$ and $N = N(c, \epsilon)$ such that $b \ge B$ and $n \ge N$ imply that

(i) If $\|h\| \le c$, then $P_{h/\sqrt{n},n}(A_{n,b}) \ge 1 - \epsilon$.

(ii) If we define the tilted measure

$$dQ_{h,n}(\xi^n) = \exp\left(-\frac{1}{2}\left[(Z_n(\xi^n) - h)^\top K(Z_n(\xi^n) - h) - Z_n(\xi^n)^\top K Z_n(\xi^n)\right]\right) dP_{0,n}(\xi^n),$$

then

$$\lim_n \sup_{h:\|h\|\le b} \int 1\{\xi^n \in A_{n,b}\} \left|dQ_{h,n} - dP_{h/\sqrt{n},n}\right| = 0.$$

Proof We have by the joint contiguity $P_{h/\sqrt{n},n} \triangleleft\triangleright P_{0,n}$ that for suitably large B and N we have $P_{h/\sqrt{n}}(A_{n,b}) \ge 1 - \epsilon$, because $P_{0,n}(A_{n,b}) \to 1$ as $b, n \to \infty$.

Let $z_n = Z_n(\xi^n)$ for short, and let us use the implicit understanding that $dP = dP(\xi^n)$ throughout. For the second result, we note that by the local asymptotic normality assumption, we have

$$\log \frac{dP_{h/\sqrt{n},n}}{dP_{0,n}} = h^\top K z_n - \frac{1}{2} h^\top K h + o_{P_{0,n}}(\|h\|) \quad \text{and} \quad \log \frac{dQ_{h,n}}{dP_{0,n}} = h^\top K z_n - \frac{1}{2} h^\top K h.$$

Thus

$$\left|\frac{dQ_{h,n}}{dP_{0,n}} - \frac{dP_{h/\sqrt{n},n}}{dP_{0,n}}\right| = \exp\left(h^\top K z_n - \frac{1}{2}h^\top K h\right)\left|1 - \exp\left(o_{P_{0,n}}(\|h\|)\right)\right|$$

as $n \to \infty$, and since the first term is $O(1)$ under $P_{0,n}$, the previous display converges in probability to zero under $P_{0,n}$. Moreover, on the event $A_{n,b}$, we know that z_n is bounded, and thus

$$\int 1\left\{\xi^n \in A_{n,b}\right\}\left|dQ_{h,n} - dP_{h/\sqrt{n},n}\right| = \int 1\left\{\xi^n \in A_{n,b}\right\}\left|\frac{dQ_{h,n}}{dP_{0,n}} - \frac{dP_{h/\sqrt{n},n}}{dP_{0,n}}\right| dP_{0,n} \to 0$$

as $n \to \infty$, and the convergence is uniform so long as $\|h\|$ is bounded. □

We now state our first major theorem, which is essentially equivalent to Proposition 6.4.4 of Le Cam and Yang [2], though we give a few minor modifications that make its application, statement, and proof simpler. In the theorem, we require a bit of notation. For matrices $K, \Gamma \succeq 0$ with $\Gamma \succ 0$, define the conditional distribution $\mathsf{G}_{K,\Gamma}(\cdot \mid z)$ as the normal distribution with mean $(K+\Gamma^{-1})^{-1}Kz$ and covariance $(K+\Gamma^{-1})^{-1}$ (i.e. precision matrix $K+\Gamma^{-1}$), that is,

$$\mathsf{G}_{K,\Gamma}(A \mid z) = \mathbb{P}(W \in A) \;\text{ for }\; W \sim \mathsf{N}\left((K+\Gamma^{-1})^{-1}Kz, (K+\Gamma^{-1})^{-1}\right).$$

We shall see that in certain Bayesian settings, if the model family $\{P_{h/\sqrt{n},n}\}$ is locally asymptotically normal with precision K, then the posterior distribution of h conditional on the data ξ^n is approximately $\mathsf{G}_{K,\Gamma}(\cdot \mid Z_n(\xi^n))$, where $Z_n(\xi^n) = K^{-1}\Delta_n(\xi^n)$ as in expression (7).

For some intuition for why this result might appear, let us revisit Example 3, where we observe $Y_i \overset{\text{iid}}{\sim} \mathsf{N}(h/\sqrt{n}, \Sigma)$ for $i \in [n]$. This model is locally asymptotically normal with $K = \Sigma^{-1}$ and $\Delta_n = n^{\frac{1}{2}}\Sigma^{-1}\overline{Y}_n$, or $z_n = n^{\frac{1}{2}}\overline{Y}_n$. Then if we put the prior π on h defined by $\mathsf{N}(0,\Gamma)$, then the posterior density

$$\begin{aligned}
p(h \mid Y_{1:n}) &\propto \exp\left(-\frac{1}{2}h^\top\Gamma^{-1}h - \frac{n}{2}(h/\sqrt{n} - \overline{Y}_n)^\top K(h/\sqrt{n} - \overline{Y}_n)\right) \\
&= \exp\left(-\frac{1}{2}(h - z_n)^\top K(h - z_n) - \frac{1}{2}h^\top\Gamma^{-1}h\right) \\
&\propto \exp\left(-\frac{1}{2}(h - (K+\Gamma^{-1})^{-1}Kz_n)^\top(K+\Gamma^{-1})(h - (K+\Gamma^{-1})^{-1}Kz_n)\right),
\end{aligned}$$

which is to say,

$$h \mid Y_{1:n} \sim \mathsf{N}\left((K+\Gamma^{-1})^{-1}Kz_n, (K+\Gamma^{-1})^{-1}\right) = \mathsf{G}_{K,\Gamma}(\cdot \mid z_n).$$

Now, let us make rigorous the more general intuition that asymptotically, the posterior distribution of h should be Gaussian in locally asymptotically normal models. Let $\pi^{\Gamma,c}$ denote the Gaussian distribution with mean 0 and variance Γ truncated to the region $\{h : \|h\| \le c\}$. Let $\pi_n^{\Gamma,c}(\cdot \mid \xi^n)$ denote the posterior distribution of h under the conditional model

$$h \sim \pi^{\Gamma,c} \;\text{ and }\; \xi^n \mid h \sim P_{h/\sqrt{n},n}.$$

Assuming that $P_{h/\sqrt{n},n}$ is locally asymptotically normal with vector Δ_n and precision K (Definition 2.1), we then have the following theorem.

Theorem 2. *Under the conditions of the above paragraph, let $Z_n = K^{-1}\Delta_n$. Define the marginal distributions*

$$\overline{P}_n := \int P_{h/\sqrt{n},n} d\pi^{\Gamma,c}(h)$$

on Ξ^n. Then for any $\epsilon > 0$, there exists $C = C(\epsilon)$ and $N = N(\epsilon)$ such that for $c \geq C$ and $n \geq N$,

$$\int \left\| \mathsf{G}_{K,\Gamma}(\cdot \mid Z_n(\xi^n)) - \pi_n^{\Gamma,c}(\cdot \mid \xi^n) \right\|_{\mathrm{TV}} d\overline{P}_n(\xi^n) \leq \epsilon.$$

Theorem 2 is one particular way of saying that, asymptotically, Locally Asymptotically Normal families have Gaussian *posterior* distributions, which means that estimation in an LAN family is eventually identical to estimation of a parameter h from the Gaussian shift family $\mathsf{N}(h, K^{-1})$ as h varies over $\mathbb{R}^d$.

Proof Before beginning the proof proper, we state a convenient lemma, whose purely technical proof we defer to Appendix A.1. To state the lemma, let M_1 and M_2 be finite positive measures on the product space $\Xi \times \Theta$. Then under standard conditions, we may define the dis-integration $M_i(d\xi, d\theta) = \nu_i(d\xi) M_i(d\theta \mid \xi)$, which is to say, simply writing the regular conditional probability if M_i are probability measures, where $M_i(\cdot \mid \xi)$ is a probability measure for ν_i-almost all ξ. Then it is possible to give bounds on the differences between the conditional (dis-integrated) measures $M_i(\cdot \mid \xi)$ in terms of the differences between the joint measures, as the next lemma shows.

Lemma 3.2 (Le Cam and Yang [2], Lemma 6.4.2)**.** *Let the conditions of the previous paragraph hold. Then*

$$\int \|M_1(\cdot \mid \xi) - M_2(\cdot \mid \xi)\|_{\mathrm{TV}} \left(d\nu_1(\xi) + d\nu_2(\xi)\right) \leq 4 \|M_1 - M_2\|_{\mathrm{TV}}.$$

Now we may proceed with the proof of our result. We define a few restricted probability measures based on Lemma 3.1. Let $c > 0$ be as in the statement of the theorem. Also let $A_{n,b}$ be the high-probability sets in Lemma 3.1, so that $P_{h/\sqrt{n},n}(A_{n,b}) \geq 1 - \epsilon$ for all sufficiently large b and n whenever $\|h\| \leq c$, and we know that $z_n = Z_n(\xi^n)$ satisfies $\|z_n\| \leq b$ on $A_{n,b}$. Then define the restricted measures

$$P^{\mathsf{rest}}_{h/\sqrt{n},n}(A) := P_{h/\sqrt{n},n}(A \cap A_{n,b}).$$

These are nearly the same as $P_{h/\sqrt{n},n}$, except they allow us to assume various random variables are bounded.

We now define a series of joint distributions and tilted measures that approximate the true distributions of ξ^n and h. For our joint measures—which we call M_0, M_1, M_2, M_3—we suppress dependence on n for notational convenience. For all h we define the tilted measures

$$dQ_{h,n}(\xi^n) := \exp\left(-\frac{1}{2}\left[(Z_n(\xi^n) - h)^\top K (Z_n(\xi^n) - h) - Z_n(\xi^n)^\top K Z_n(\xi^n)\right]\right) dP^{\mathsf{rest}}_{0,n}(\xi^n),$$

which are (by Lemma 3.1(ii)) essentially equivalent to $dP_{h/\sqrt{n},n}$: they satisfy

$$\lim_n \sup_{\|h\| \leq c} \left\| Q_{h,n} - P^{\mathsf{rest}}_{h/\sqrt{n},n} \right\|_{\mathrm{TV}} = 0. \tag{8}$$

However—as we shall see—the posterior distributions of h under the "sampling" scheme $Q_{h,n}$ for Z_n are eventually Gaussian, which will yield the theorem. Let M_0 be the true joint distribution on the pair (h, ξ^n) under our truncated Gaussian prior, defined by

$$dM_0(\xi^n, h) = dP_{h/\sqrt{n},n}(\xi^n) d\pi^{\Gamma,c}(h).$$

In addition, let M_1 and M_2 be the joint distributions on the pair (h, ξ^n) defined by

$$dM_1(\xi^n, h) = dQ_{h,n}(\xi^n) d\pi^{\Gamma,c}(h) \quad \text{and} \quad dM_2(\xi^n, h) = dP^{\mathsf{rest}}_{h/\sqrt{n},n}(\xi^n) d\pi^{\Gamma,c}(h).$$

Because the support $\operatorname{supp} \pi^{\Gamma,c} \subset \{h \in \mathbb{R}^d : \|h\| \le c\}$, the limit (8) immediately implies that M_1 and M_2 are close:

$$\limsup_n \|M_1 - M_2\|_{\mathrm{TV}} \le \limsup_n \int \left\| Q_{h,n} - P^{\mathsf{rest}}_{h/\sqrt{n},n} \right\|_{\mathrm{TV}} d\pi^{\Gamma,c}(h) = 0.$$

Moreover, we have that

$$\|M_0 - M_2\|_{\mathrm{TV}} \le \int \left\| P^{\mathsf{rest}}_{h/\sqrt{n},n} - P_{h/\sqrt{n},n} \right\|_{\mathrm{TV}} d\pi^{\Gamma,c}(h) = \int P_{h/\sqrt{n},n}(\Xi^n \setminus A_{n,b}) d\pi^{\Gamma,c}(h) \le \epsilon$$

for all sufficiently large n, by definition of the restricted measures and sets $A_{n,b}$ from Lemma 3.1(i).

For our final joint measure, we consider the true Gaussian prior $\pi^{\Gamma,\infty}$, which is $\mathsf{N}(0, \Gamma)$, defining

$$\begin{aligned} dM_3(\xi^n, h) &= dQ_{h,n}(\xi^n) d\pi^{\Gamma,\infty}(h) \qquad (9) \\ &= \exp\left(-\frac{1}{2}(h - (K + \Gamma^{-1})^{-1} K z_n)^\top (K + \Gamma^{-1})(h - (K + \Gamma^{-1})^{-1} K z_n) + \frac{1}{2} z_n^\top K z_n \right) dP^{\mathsf{rest}}_{0,n}(\xi^n), \end{aligned}$$

where we used the shorthand $z_n = Z_n(\xi^n) = K^{-1}\Delta_n(\xi^n)$. For any $\epsilon > 0$, there is certainly a C large enough that $c \ge C$ implies $\|\pi^{\Gamma,c} - \pi^{\Gamma,\infty}\|_{\mathrm{TV}} \le \epsilon$. As a consequence, we have

$$\begin{aligned} 2\|M_1 - M_3\|_{\mathrm{TV}} &= \int |dQ_{h,n}(\xi^n) d\pi^{\Gamma,c}(h) - dQ_{h,n}(\xi^n) d\pi^{\Gamma,\infty}(h)| \\ &\le \left(\int_{\Xi^n} \sup_h dQ_{h,n}(\xi^n) \right) \int |d\pi^{\Gamma,\infty}(h) - d\pi^{\Gamma,c}(h)| \\ &= \int_{\Xi}^n \exp\left(\frac{1}{2} Z_n(\xi^n)^\top K Z_n(\xi^n) \right) dP^{\mathsf{rest}}_{0,n}(\xi^n) \cdot 2 \left\| \pi^{\Gamma,c} - \pi^{\Gamma,\infty} \right\|_{\mathrm{TV}}. \end{aligned}$$

But of course, by the definition of the restricted measure P^{rest}, we know that $Z_n(\xi^n)$ is bounded on the support of $P^{\mathsf{rest}}_{0,n}$, and thus the final expression has upper bound

$$\|M_1 - M_3\|_{\mathrm{TV}} \le O(1) \left\| \pi^{\Gamma,c} - \pi^{\Gamma,\infty} \right\|_{\mathrm{TV}},$$

where the $O(1)$ term depends only on the constant b in the high probability sets $A_{n,b}$ of Lemma 3.1 (i.e. $O(1) \le \exp(\frac{1}{2} b^2 \|\!|K|\!\|)$). In particular, for any $\epsilon > 0$, we may choose c large enough in the prior $\pi^{\Gamma,c}$ that $\limsup_n \|M_1 - M_3\|_{\mathrm{TV}} \le \epsilon$. Summarizing all of our preceding derivations, we see that for any $\epsilon > 0$, we may choose c large enough that

$$\limsup_n \|M_k - M_l\|_{\mathrm{TV}} \le \epsilon \quad \text{for } k, l \in \{0, \ldots, 3\}. \qquad (10)$$

Now, by inspection, we see that the posterior distribution of h under the joint measure M_3 is

$$M_3(\cdot \mid \xi^n) = \mathsf{N}\left((K + \Gamma^{-1})^{-1} K Z_n(\xi^n), (K + \Gamma^{-1})^{-1} \right) = \mathsf{G}_{K,\Gamma}(\cdot \mid Z_n(\xi^n)).$$

Let $\pi^{\Gamma,c}(\cdot \mid \xi^n)$ denote the posterior distribution on h as described in the theorem statement. Then $dM_0(\xi^n, h) = d\overline{P}_n(\xi^n) d\pi^{\Gamma,c}(h \mid \xi^n)$ by construction, and $dM_3(\xi^n, h) = d\overline{M}_3(\xi^n) d\mathsf{G}_{K,\Gamma}(h \mid Z_n(\xi^n))$. Then the dis-integration result of Lemma 3.2 shows that

$$\begin{aligned} &\int \left\| \mathsf{G}_{K,\Gamma}(\cdot \mid Z_n(\xi^n)) - \pi^{\Gamma,c}(\cdot \mid \xi^n) \right\|_{\mathrm{TV}} d\overline{P}_n(\xi^n) \\ &\quad \le \int \left\| \mathsf{G}_{K,\Gamma}(\cdot \mid Z_n(\xi^n)) - \pi^{\Gamma,c}(\cdot \mid \xi^n) \right\|_{\mathrm{TV}} (d\overline{P}_n(\xi^n) + d\overline{M}_3(\xi^n)) \le 4 \|M_0 - M_3\|_{\mathrm{TV}}. \end{aligned}$$

But we know that the final quantity is eventually less than ϵ by inequality (10). $\square$

3.2 Limiting Gaussian experiments

JCD Comment: I will probably write this eventually. For now, van der Vaart [3, Chapter 7] is a good reference, and gives an alternative approach to ours.

4 Efficiency of estimators and asymptotic minimax results

It is possible to provide a number of asymptotic efficiency results based on our characterizations of local asymptotic normality. There are a few approaches to such results. One approach, following Le Cam [2], provides Bayesian lower bounds that use strongly the asymptotic normality guarantees above, as we can apply standard normality theory. The other, a slightly more abstract formulation than ours, uses Van Der Vaart's treatment of asymptotically normal families as limiting to normal experiments.

For each of these techniques, we require a classical and important result due to Anderson [2, 3] on optimal estimation of a Gaussian mean. Recall that a function $L : \mathbb{R}^k \to \mathbb{R}$ is *quasi-convex* if it is bowl-shaped, that is, the sublevel sets

$$\mathrm{sub}_\alpha(L) := \{x \in \mathbb{R}^k : L(x) \le \alpha\}$$

are convex sets. A function L is symmetric if $L(x) = L(-x)$ for all x. Let $L : \mathbb{R}^k \to \mathbb{R}$ be a symmetric and quasi-convex function. A simple version of Anderson's lemma is as follows.

Lemma 4.1 (Anderson)**.** *Let $A \in \mathbb{R}^d \times \mathbb{R}^k$ and let $X \sim \mathsf{N}(\mu, \Sigma)$, where $X \in \mathbb{R}^d$. Then*

$$\inf_{v \in \mathbb{R}^k} \mathbb{E}[L(AX - v)] = \mathbb{E}[L(A(X - \mu))] = \mathbb{E}[L(A\Sigma^{\frac{1}{2}}W)]$$

for $W \sim \mathsf{N}(0, I_{d\times d})$, so that $v = A\mu$ minimizes $\mathbb{E}[L(AX - v)]$.

A trivial consequence of Anderson's lemma is that $\mathbb{E}[X]$ minimizes $\mathbb{E}[\|X - v\|_2^2]$ over v, though of course this is provable by other means. A less trivial consequence, however, is to take the function $L(x) = \frac{1}{2}\|x\|_2^2 \wedge B = \min\{\frac{1}{2}\|x\|_2^2, B\}$, which is evidently quasi-convex and symmetric. Then again, Anderson's lemma shows that $\inf_{v\in\mathbb{R}^d} \mathbb{E}[\|X - v\|_2^2 \wedge B] = \mathbb{E}[\|X - \mathbb{E}[X]\|_2^2 \wedge B]$.

4.1 The Bayesian approach to asymptotic lower bounds

For our first theorem, we use the asymptotic posterior argument of Theorem 2 to apply Anderson's lemma in an estimation problem.

Theorem 3 (Local Asymptotic Minimax Theorem)**.** *Let $L : \mathbb{R}^d \times \mathbb{R}$ be a symmetric, quasi-convex, and bounded function, and let $P_{\theta_0+u/\sqrt{n},n}$ be a locally asymptotically normal family of distributions with precision matrix K. Then for any sequence of estimators $\widehat{\theta}_n : \Xi^n \to \Theta$*

$$\liminf_{c\to\infty} \liminf_n \sup_{\|\theta-\theta_0\| \le \frac{c}{\sqrt{n}}} \mathbb{E}_{P_{\theta,n}}\left[L(\sqrt{n}(\widehat{\theta}_n - \theta))\right] \ge \mathbb{E}\left[L\left(K^{-\frac{1}{2}}W\right)\right],$$

where $W \sim \mathsf{N}(0, I_{d\times d})$.

In fact, an inspection of the proof shows that we may replace the supremum in the above limit with an integral over prior measures $\pi_{n,c}$ that are absolutely continuous with respect to Lebesgue measure and supported on $\{\theta : \|\theta - \theta_0\| \le c/\sqrt{n}\}$. Even more, we may take $\pi_{n,c}$ to be the Gaussian distribution with mean θ_0 and covariance $\frac{1}{\epsilon(c)n} I$, where $\epsilon(c) \to 0$ as $c \to \infty$, truncated to values $\|\theta - \theta_0\| \le c/\sqrt{n}$. That is,

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \inf_{\widehat{\theta}_n} \int \mathbb{E}_{P_{\theta,n}} \left[L(\sqrt{n}(\widehat{\theta}_n - \theta)) \right] d\pi_{n,c}(\theta) \ge \mathbb{E}\left[L\left(K^{-\frac{1}{2}} W \right) \right], \tag{11}$$

which is a stronger result than Theorem 3.

Proof Assume without loss of generality that L takes values in $[0, 1]$. We begin by noting that we may replace the supremum over θ such that $\|\theta - \theta_0\| \le c/\sqrt{n}$ with a supremum over $h : \|h\| \le c$, defining $\theta = \theta_0 + h/\sqrt{n}$. Without loss of generality we take $\theta_0 = 0$. Then if we take priors $\pi^{\Gamma,c}$ to be normal distributions $\mathsf{N}(0, \Gamma)$ truncated to $\|h\| \le c$, then

$$\sup_{\|\theta - \theta_0\| \le \frac{c}{\sqrt{n}}} \mathbb{E}_{P_{\theta,n}} \left[L(\sqrt{n}(\widehat{\theta}_n - \theta)) \right] \ge \int \mathbb{E}_{P_{h/\sqrt{n},n}} \left[L(\sqrt{n}(\widehat{\theta}_n - h/\sqrt{n})) \right] d\pi^{\Gamma,c}(h).$$

If we define $\overline{P}_n = \int P_{h/\sqrt{n},n} d\pi(h)$ and $d\pi(h \mid \xi^n)$ to be the posterior over h conditional on the observation ξ^n under the prior $\pi^{\Gamma,c}$, then the preceding expression has the further lower bound

$$\int \mathbb{E}\left[L\left(\sqrt{n}\widehat{\theta}_n(\xi^n) - h \right) \mid \xi^n \right] d\overline{P}_n(\xi^n) \ge \int \inf_{\widehat{h}} \mathbb{E}\left[L(\widehat{h} - h) \mid \xi^n \right] d\overline{P}_n(\xi^n). \tag{12}$$

Now we apply Theorem 2 to the bound (12). We know that if $\mathsf{G}_{K,\Gamma}(\cdot \mid \xi^n)$ is the Gaussian distribution $\mathsf{N}((K + \Gamma^{-1})^{-1} K z_n, (K + \Gamma^{-1})^{-1})$, then for any $\epsilon > 0$ and suitably large c and all sufficiently large n (we may choose n after c), we have

$$\int_{\Xi^n} \left\| \mathsf{G}_{K,\Gamma}(\cdot \mid \xi^n) - \pi^{\Gamma,c}(\cdot \mid \xi^n) \right\|_{\mathrm{TV}} d\overline{P}_n(\xi^n) \le \epsilon.$$

Adding and subtracting the expectation of $L(\widehat{h} - h)$ under the distribution $\mathsf{G}_{K,\Gamma}(\cdot \mid \xi^n)$ in expression (12), we have

$$\begin{aligned}
&\int \inf_{\widehat{h}} \mathbb{E}\left[L(\widehat{h} - h) \mid \xi^n \right] d\overline{P}_n(\xi^n) \\
&\ge \int \inf_{\widehat{h}} \mathbb{E}_{\mathsf{G}_{K,\Gamma}}\left[L(\widehat{h} - h) \mid \xi^n \right] d\overline{P}_n(\xi^n) - \int \sup_{h,\widehat{h}} |L(\widehat{h} - h)| \left\| \mathsf{G}_{K,\Gamma}(\cdot \mid \xi^n) - \pi^{\Gamma,c}(\cdot \mid \xi^n) \right\|_{\mathrm{TV}} d\overline{P}_n(\xi^n) \\
&\ge \int \inf_{\widehat{h}} \mathbb{E}_{\mathsf{G}_{K,\Gamma}}\left[L(\widehat{h} - h) \mid \xi^n \right] d\overline{P}_n(\xi^n) - \sup_{h,\widehat{h}} |L(\widehat{h} - h)| \epsilon,
\end{aligned}$$

where we have applied Theorem 2. Applying Anderson's Lemma 4.1 to the Gaussian $\mathsf{G}_{K,\Gamma}$, we obtain

$$\int \inf_{\widehat{h}} \mathbb{E}\left[L(\widehat{h} - h) \mid \xi^n \right] d\overline{P}_n(\xi^n) \ge \mathbb{E}\left[L\left((K + \Gamma^{-1})^{-\frac{1}{2}} W \right) \right] - \epsilon \tag{13}$$

where $W \sim \mathsf{N}(0, I_{d\times d})$, for all sufficiently large c and n.

Lastly, we note that if we take $\Gamma = \frac{1}{\epsilon} K^{-1}$, then we have $(K + \Gamma^{-1}) = (1 + \epsilon)K$, and as $\epsilon \to 0$, we certainly have $((1 + \epsilon)K)^{-\frac{1}{2}} W \to W$. Recalling that L is lower semi-continuous (see the Portmanteau lemma) and that $\epsilon > 0$ in expression (13) is arbitrary, taking $\Gamma = \frac{1}{\epsilon} K^{-1}$ gives the final result. $\square$

4.2 Estimating functionals

In some cases, we may wish to estimate a function of the distribution at hand, which is a somewhat different problem than just estimating the parameter. In the next section, we consider this in a non-parametric sense, but in this section, we shall content our selves with estimating a smooth function of the parameter θ.

Consider a function $\psi : \Theta \to \mathbb{R}^k$, where $\Theta \subset \mathbb{R}^d$, and let $\dot{\psi}(\theta)$ be its derivative matrix in the sense that $\psi(\theta) = \psi(\theta_0) + \dot{\psi}(\theta_0)(\theta - \theta_0) + o(\|\theta - \theta_0\|)$ for θ near θ_0. Then we may claim the following result, which is a minor extension of Theorem 3. In the corollary, we assume in that the loss L is Lipschitz continuous, which is no real loss of generality but makes our proof simpler. We also adopt the notation $\pi_{n,c}$ to be truncated Gaussian distributions centered at θ_0 truncated to $\|\theta - \theta_0\| \le c/\sqrt{n}$, where we show how to choose the covariance in the proof.

Corollary 4.1. *In addition to the conditions of Theorem 3 around θ_0, assume that L is Lipschitz continuous. Then for any sequence of estimators $\widehat{\psi}_n : \Xi^n \to \mathbb{R}^k$, we have*

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \inf_{\widehat{\psi}_n} \int \mathbb{E}_{P_{\theta,n}} \left[L(\sqrt{n}(\psi(\theta) - \widehat{\psi}_n)) \right] d\pi_{n,c}(\theta) \ge \mathbb{E}\left[L(\dot{\psi}(\theta_0) K^{-\frac{1}{2}} W) \right]$$

for $W \sim \mathsf{N}(0, I_{d\times d})$.

Proof Let $\mathfrak{M}(n,c) = \inf_{\widehat{\psi}_n} \int \mathbb{E}_{P_{\theta,n}} \left[L(\sqrt{n}(\psi(\theta) - \widehat{\psi}_n)) \right] d\pi_{n,c}(\theta)$ denote the minimax quantity on the left side of the corollary. Without loss of generality, we let $\theta_0 = 0$, and as in the proof of Theorem 3, we let the priors $\pi^{\Gamma,c}$ denote Gaussian distributions $\mathsf{N}(0,\Gamma)$ truncated to $\|h\| \le c$. In this case we have

$$\mathfrak{M}(n,c) = \inf_{\widehat{\psi}_n} \int \mathbb{E}_{P_{h/\sqrt{n},n}} \left[L(\sqrt{n}(\psi(h/\sqrt{n}) - \widehat{\psi}_n)) \right] d\pi^{\Gamma,c}(h)$$

Using that L is $\mathrm{Lip}(L)$-Lipschitz and writing $\psi(h/\sqrt{n}) = \psi(0) + \dot{\psi}(0)h/\sqrt{n} + [\psi(h/\sqrt{n}) - \psi(0) - \dot{\psi}(0)h/\sqrt{n}]$, we obtain

$$\begin{aligned}\mathfrak{M}(n,c) \ge \inf_{\widehat{\psi}_n} \int & \mathbb{E}_{P_{h/\sqrt{n},n}} \left[L(\dot{\psi}(0)h - \widehat{\psi}_n) \right] d\pi^{\Gamma,c}(h) \\ & - \mathrm{Lip}(L) \mathbb{E}_{\pi^{\Gamma,c}} \left[\left\| \sqrt{n}(\psi(h/\sqrt{n}) - \psi(0) - \dot{\psi}(0)h/\sqrt{n}) \right\| \wedge \frac{1}{\mathrm{Lip}(L)} \right].\end{aligned}$$

As $n \to \infty$, we have $\sqrt{n}(\psi(h/\sqrt{n}) - \psi(0) - \dot{\psi}(0)h/\sqrt{n}) = \sqrt{n} o(\|h\|/\sqrt{n}) = o_P(1)$, so the last term converges to 0. The remainder of the proof is identical to the proof of Theorem 3, as the function $v \mapsto L(\dot{\psi}(0)v)$ is symmetric, bounded, and quasi-convex. □

4.3 Non-parametric estimation

In more general statistical problems, it may not make sense to assume the existence of a parameter uniquely identifying the distribution, in which case the previous derivations and notions of optimality make less sense. In this case, it is a bit of a matter of taste exactly what "optimality" means in terms of estimation, but there are a number of possible approaches. We follow one based on choosing "hardest" (parametric) sub-models of a given family, which is also the approach taken

by Bickel et al. [1] and in the final chapter of van der Vaart [3], much of it based on joint work with Susan Murphy. Our approach will be to consider (roughly) tilts of the distribution P by function $g \in L^2(P)$ with $Pg = 0$, which is one way of slightly perturbing the underlying distribution and asking for optimality with respect to these small tilts.

As an example to motivate our considerations, consider the set $\mathcal{P}$ of all distributions on a sample space $\mathcal{X} \subset \mathbb{R}^d$, and assume we observe data $X_i \stackrel{\text{iid}}{\sim} P \in \mathcal{P}$ and wish to estimate $\mathbb{E}_P[X]$. Certainly the mean does not uniquely specify P, so the parametric results in the preceding sections say little about this situation, but we intuitively believe that the empirical mean $\frac{1}{n}\sum_{i=1}^n X_i$ should be efficient and (at least asymptotically) optimal for estimation of $\mathbb{E}_P[X]$. A purely minimax analysis, however, is not sufficient: let $Q = (1-\epsilon)P + \epsilon\delta_x$, where δ_x denotes the point mass at x. Then $d_{\rm hel}^2(P^n, Q^n) = 1 - (1 - d_{\rm hel}^2(P,Q))^n$ and $d_{\rm hel}^2(P,Q)^2 \le \|P - Q\|_{\rm TV} \le 2\epsilon$, so that we may take $\epsilon = \frac{1}{n}$ and obtain $d_{\rm hel}^2(P^n, Q^n) < 1 - e^{-1}$, no procedure can consistently distinguish P and Q given n i.i.d. observations, but the mean difference $\mathbb{E}_P[X] - \mathbb{E}_Q[X] = \epsilon(\mathbb{E}_P[X] - x)$ can be arbitrarily large by taking x large.

With this motivation, we consider a different approach and attempt to define information appropriately. We have a distribution P known to belong to some set $\mathcal{P}$ of distributions on $\mathcal{X}$, and we wish to estimate the value $\psi(P)$ of the function $\psi : \mathcal{P} \to \mathbb{R}^d$ at P. Our idea is to consider smooth enough parametric sub-models $\mathcal{P}_0 \subset \mathcal{P}$, where we let $\mathcal{P}_0 = \{P_\theta \mid \theta \in \Theta\}$ for some parameters θ, and then apply the results of the preceding sections. By looking at the least favorable or "hardest" sub-models, we can provide a theory of asymptotic normality.

4.3.1 Score functions and quadratic mean differentiability

Following van der Vaart [3], we consider 1-dimensional submodels that are appropriately quadratic mean differentiable (QMD), as in Section 1.5 and Definition 1.3. Let the map $t \mapsto P_t$, for $0 \le t < \infty$ (where $P_0 = P$) be QMD at $t = 0$ with score function $g : \mathcal{X} \to \mathbb{R}$, that is, satisfying

$$\int \left(\sqrt{dP_t} - \sqrt{dP} - \frac{1}{2} t g \sqrt{dP}\right)^2 = o(t^2), \tag{14}$$

or, similarly, $\int (\frac{dP_t^{1/2} - dP^{1/2}}{t} - \frac{1}{2} g dP^{1/2})^2 \to 0$ as $t \downarrow 0$. A slightly more general version of Eq. (14) is to consider $g : \mathcal{X} \to \mathbb{R}^k$ and $h \in \mathbb{R}^k$ near zero, and assuming

$$\int \left(\sqrt{dP_h} - \sqrt{dP} - \frac{1}{2} h^\top g \sqrt{dP}\right)^2 = o(\|h\|^2). \tag{15}$$

An immediate consequence of the convergence (15) is based on Proposition 1.

Lemma 4.2. *Let the family $\{P_h\}$ be QMD (15) for $h \in \mathbb{R}^k$. Then $Pg = 0$, $P\|g\|^2 < \infty$, and the family is locally asymptotically normal (Definition 2.1) with precision matrix $K = Pgg^\top$, that is,*

$$\sum_{i=1}^n \log \frac{dP_{h/\sqrt{n}}}{dP}(X_i) = \frac{1}{\sqrt{n}} h^\top \sum_{i=1}^n g(X_i) - \frac{1}{2} h^\top P g g^\top h + o_P(\|h\|).$$

We call a function g above a *score function*, as in the parametric QMD situation, and by considering a variety of sub-models, we obtain a collection of score functions that we denote by $\dot{\mathcal{P}}_P$ at P. Each function in this family belongs to $L^2(P)$ by Lemma 4.2.

Example 6 (Fully non-parametric models)**:** The typical example of such a family is the fully non-parametric setting, where we take the tangent set $\dot{\mathcal{P}}_P$ at P to be all functions $g \in L^2(P)$ with

$Pg = 0$. This is evidently the "maximal" tangent set, in that every score function must belong to $L^2(P)$ and satisfy $Pg = 0$. Concretely, we let the models be defined by *tilts* of the distribution P. To motivate, the case in which $\sup_x |g(x)| < \infty$ allows us to define the tilted distribution

$$dP_t(x) := (1 + tg(x))dP(x),$$

which (for small $t > 0$) is a valid distribution and satisfies $\int dP_t = 1$. To handle all of $L^2(P)$, we tweak this slightly. Let $\phi : \mathbb{R} \to [0, 2]$ be any function, $\mathcal{C}^2$ near 0, which we assume to be 1-Lipschitz, with $\phi(0) = 1$ and $\phi'(0) = 1$, so that $\phi(t) = 1 + t + o(t)$. As a simple example of such a ϕ, consider $\phi(t) = \min\{2, \max\{1 + t, 0\}\}$. Then we set

$$dP_t(x) := \frac{1}{c(t)}\phi(tg(x))dP(x),$$

where $c(t)$ is the normalizing constant. By Lebesgue's dominated convergence theorem, we have

$$\lim_{t \downarrow 0} \frac{1}{t} \int \left[\phi(tg(x)) - (1 + tg(x))\right] dP(x) = \phi'(0) \int g(x)dP(x) = 0,$$

as $|\phi(tg(x)) - (1 + tg(x))|/t \le 2|g(x)|$, so that $c(t) = 1 + o(t)$. We also have that

$$\frac{\partial}{\partial t} \log dP_t(x)\Big|_{t=0} = g(x),$$

and similarly that the score function of P_t is g in the QMD sense (14). $\diamond$

4.3.2 Influence functions and derivatives

Now, we consider the problem of actually estimating $\psi(P)$. We assume that $\psi(P_t)$ is differentiable at $t = 0$, as we cannot define appropriate notions of information for non-smooth functionals (in a sense, if for some sub-model $t \mapsto P_t$, the function $t \mapsto \psi(P_t)$ is not differentiable at $t = 0$, then in a sense it is not smooth enough to estimate at typical $1/\sqrt{n}$ rates; for an exploration of this phenomenon, see Exercise 11.2). More precisely, let us say that $\psi : \mathcal{P} \to \mathbb{R}^k$ is *differentiable at the point P relative to a tangent set $\dot{\mathcal{P}}_P$* (recall that $\dot{\mathcal{P}} \subset \{g : \mathcal{X} \to \mathbb{R}, Pg = 0, Pg^2 < \infty\}$) if there exists a continuous linear map $\dot{\psi}_P : L^2(P) \to \mathbb{R}^d$ with

$$\lim_{t \to 0} \frac{\psi(P_t) - \psi(P)}{t} = \dot{\psi}_P(g)$$

whenever the model $t \mapsto P_t$ has score function g. Because $L^2(P)$ is a Hilbert space and the bounded linear functions on $L^2(P)$ are isomorphic to $L^2(P)$, the Riesz Representation Theorem implies there must exist a vector-valued function function $\nabla\psi_P : \mathcal{X} \to \mathbb{R}^d$ with

$$\dot{\psi}_P(g) := \int \nabla\psi_P(x)g(x)dP(x).$$

It is no loss of generality to assume that $\nabla\psi_P$ is mean-zero, as each g is mean zero; we do this without comment from this point on. There is a unique version of $\nabla\psi_P$ contained in the closure of the linear span of $\dot{\mathcal{P}}_P$, but we shall not concern ourselves with such issues except to assume that $\nabla\psi_P$ belongs to this span from now on.

Example 7 (Mean estimation)**:** Let us consider the (perhaps) simplest problem of estimation of the mean of a distribution P with $\mathrm{Var}_P(X) < \infty$. That is, $\psi(P) = \mathbb{E}_P[X]$. In this case, we let $dP_t(x) = \frac{1}{c(t)}\phi(tg(x))dP(x)$ as in Example 6, where $g \in L^2(P)$ and $Pg = 0$. Then we have

$$\mathbb{E}_{P_t}[X] = \frac{1}{c(t)} \int \phi(tg(x))xdP(x) = \int x(1 + tg(x))dP(x) + o(t)$$

by the dominated convergence theorem, so that

$$\frac{\psi(P_t) - \psi(P)}{t} \to \int x g(x) dP(x),$$

which is evidently linear in g. Thus, the influence function in this case is simply $\nabla\psi_P(x) = x - \mathbb{E}_P[X]$. $\diamond$

Question 11.3 discusses another example of the appropriate influence functions for regression problems with squared loss, but without making any particular modeling assumptions so that the distribution P is a nuisance parameter.

These influence functions are the fundamental quantity governing the convergence, and limiting convergence rates, of estimates of parameters (or other quantities) in non-parametric or nearly non-parametric problems. One can develop our parametric local asymptotic minimax theory using these non-parametric influence functions, simply by considering the sub-models to be only those parameterized by a desired function θ. Indeed, assuming the family $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ is QMD and that Θ is convex, around a fixed $\theta_0 \in \operatorname{int} \Theta$ we may consider sub-models, defined for t near 0 and $h \in \mathbb{R}^d$, of the form

$$P_t = P_{th+\theta_0},$$

where for the function $\psi(P_\theta) = \theta$ we evidently have

$$\lim_{t \downarrow 0} \frac{\psi(P_t) - \psi(P_0)}{t} = \lim_{t \downarrow 0} \frac{th}{t} = h,$$

and the tangent set $\dot{\mathcal{P}}_{\theta_0} = \{h^\top \dot{\ell}_{\theta_0} \mid h \in \mathbb{R}^d\}$ is the linear span of the usual score functions of the parameter θ_0. In this case, by inspection we have the influence function $\nabla\psi(x) := I_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(x)$, where I_{θ_0} is the Fisher information, because for score $h^\top \dot{\ell}_{\theta_0}(x)$ and $P_t = P_{th+\theta_0}$, we obtain

$$\lim_{t \to 0} \frac{\psi(P_t) - \psi(P_0)}{t} = h = \mathbb{E}_{\theta_0}\left[I_{\theta_0}^{-1} \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^\top\right] h = I_{\theta_0}^{-1} I_{\theta_0} h = h.$$

In this case, Theorem 3 shows that $\mathbb{E}[L(Z)]$, where $Z \sim \mathsf{N}(0, \mathbb{E}[\nabla\psi \nabla\psi^\top])$, is the local asymptotic minimax lower bound on estimation of θ_0.

4.3.3 A more general local asymptotic minimax theorem

One might hope that the influence functions also provide lower bounds for estimation in non-parametric contexts; happily, this turns out to be the case. In the theorem, we let $\mathcal{P}$ be a collection of distributions, $\psi : \mathcal{P} \to \mathbb{R}^d$, and $P \in \mathcal{P}$. Let $\dot{\mathcal{P}}_P$ be a tangent set to P, and assume that ψ is differentiable at P relative to $\dot{\mathcal{P}}_P$. We also let $\nabla\psi : \mathcal{X} \to \mathbb{R}^d$ denote the influence function for ψ at P. For simplicity in notation and a bit of a more compact result—as well as to avoid taking a supremum that we find perhaps unsatisfying—given $g_1, \ldots, g_k \in \dot{\mathcal{P}}$ and $h \in \mathbb{R}^k$, we let P_h denote distributions based on the score function $h^\top g(x) = \sum_{j=1}^k h_j g_j(x)$ in the QMD sense (15). We have the following result.

Theorem 4. *Let $L : \mathbb{R}^d \to \mathbb{R}$ be symmetric, quasi-convex, and bounded. There exist truncated, mean-zero Gaussian distributions $\pi_{n,c,k}$ supported on $\{h \in \mathbb{R}^k \mid \|h\| \le c/\sqrt{n}\}$ such that the following holds:*

$$\sup_{k \in \mathbb{N}} \sup_{g_1, \ldots, g_k \in \dot{\mathcal{P}}_P} \liminf_{c \to \infty} \liminf_{n \to \infty} \inf_{\widehat{\psi}_n} \int \mathbb{E}_{P_h^n}\left[L\left(\sqrt{n}(\widehat{\psi}_n - \psi(P_h))\right)\right] d\pi_{n,c,k}(h) \ge \mathbb{E}[L(Z)]$$

where $Z \sim \mathsf{N}(0, \mathbb{E}_P[\nabla\psi(X)\nabla\psi(X)^\top])$. Even more, the supremum over $k \in \mathbb{N}$ and $g_1, \ldots, g_k$ may be replaced by the single choice $g(x) = \nabla\psi(x)$.

Before giving the proof of the theorem, we revisit Example 7 to prove that the sample mean is essentially an optimal estimator.

Example 8 (Mean estimation optimality)**:** Consider again the setting of Example 7, where we define $dP_t(x) = \frac{1}{c(t)}\phi(tg(x))dP(x)$ for $g \in L^2(P)$ mean-zero and $\phi(t) = \min\{2, [1+t]_+\}$. In this case, the influence function is simply the identity, $\nabla\psi_P(x) = x$, and Theorem 4 then implies that looking over all tilt perturbations of the distribution P, as in Example 6, we obtain a local asymptotic minimax lower bound of

$$\mathbb{E}[L(Z)] \quad \text{for} \quad Z \sim \mathsf{N}\left(0, \mathrm{Cov}_P(X)\right).$$

That the sample mean $\overline{X}_n$ achieves this result is clear once we observe that $\sqrt{n}(\overline{X}_n - \mathbb{E}[X]) \xrightarrow{d} \mathsf{N}(0, \mathrm{Cov}_P(X))$ and apply Le Cam's third lemma (Example 2) to obtain uniformity in perturbations P_g. $\diamond$

Proof of Theorem 4 For any fixed set of functions $g_1, \ldots, g_k$, let $g = [g_1 \cdots g_k]^\top$. Then we have as in Corollary 4.1 that the family $\{dP_h\}_{h\in\mathbb{R}^k}$ is locally asymptotically normal with precision matrix $K = Pgg^\top$, and moreover, we may redefined our function ψ by $\psi_k : \mathbb{R}^k \to \mathbb{R}^d$ with $\psi_k(h) = \psi(P_h)$. We then have the finite dimensional result

$$\psi_k(h) = \psi_k(0) + \dot{\psi}_k(0)h + o(\|h\|) \quad \text{where} \quad \dot{\psi}_k(0) = \mathbb{E}_P\left[\nabla\psi(X)g(X)^\top\right],$$

by definition of the influence function. Applying Corollary 4.1, then, we obtain that

$$\liminf_{c\to\infty}\liminf_{n\to\infty}\inf_{\widehat{\psi}_n}\int \mathbb{E}_{P_h^n}\left[L\left(\sqrt{n}(\widehat{\psi}_n - \psi(P_h))\right)\right] d\pi_{n,c,k}(h) \geq \mathbb{E}[L(\dot{\psi}_k(0)K^{-1/2}W)]$$

for $W \sim \mathsf{N}(0, I)$. (If $K = Pgg^\top$ is not invertible, we replace K^{-1} with its pseudo-inverse.)

With this result, we note that

$$\dot{\psi}_k(0)K^{-1/2}W \sim \mathsf{N}\left(0, \mathbb{E}_P[\nabla\psi(X)g(X)^\top]\mathbb{E}_P[g(X)g(X)^\top]^{-1}\mathbb{E}_P[g(X)\nabla\psi(X)^\top]\right).$$

We claim that for any random (possibly dependent) random vectors Z, Y that

$$\mathbb{E}[YZ^\top]\mathbb{E}[ZZ^\top]^\dagger\mathbb{E}[ZY^\top] \preceq \mathbb{E}[YY^\top]. \tag{16}$$

Deferring the proof of the claim (16), note that it implies the covariance

$$\mathbb{E}_P[\nabla\psi(X)g(X)^\top]\mathbb{E}_P[g(X)g(X)^\top]^{-1}\mathbb{E}_P[g(X)\nabla\psi(X)^\top] \preceq \mathbb{E}_P[\nabla\psi(X)\nabla\psi(X)^\top].$$

By taking $g(x) = \nabla\psi(x)$, which is possible because we have $\mathbb{E}[\nabla\psi(X)] = 0$ and $\nabla\psi$ belongs to the closure of the linear span of $\dot{\mathcal{P}}_P$ by assumption (recall the discussion at the beginning of the section), we find that it is possible to choose $g : \mathcal{X} \to \mathbb{R}^d$ such that

$$\liminf_{c\to\infty}\liminf_{n\to\infty}\inf_{\widehat{\psi}_n}\int \mathbb{E}_{P_h^n}\left[L\left(\sqrt{n}(\widehat{\psi}_n - \psi(P_h))\right)\right] d\pi_{n,c,d}(h) \geq \mathbb{E}[L(Z)], \quad Z \sim \mathsf{N}(0, \mathbb{E}[\nabla\psi(X)\nabla\psi(X)^\top]).$$

Using Anderson's lemma (Lemma 4.1) and Claim (16) shows that this is the "worst" possible covariance, giving the theorem except for the proof of our claim.

To see the claim (16), we may first assume without loss of generality that $\mathbb{E}[ZZ^\top] \preceq I$: indeed, letting $\Sigma = \mathbb{E}[ZZ^\top]^\dagger$, we have

$$\mathbb{E}[\Sigma^{1/2}ZZ^\top\Sigma^{1/2}] = \Sigma^{1/2}\mathbb{E}[ZZ^\top]\Sigma^{1/2} = \Sigma^{1/2}\Sigma^\dagger\Sigma^{1/2} \preceq I,$$

so to show inequality (16) it is sufficient to show that $\mathbb{E}[YZ^\top]\mathbb{E}[ZY^\top] \preceq I$ for Z such that $\mathbb{E}[ZZ^\top] \preceq I$. For this last inequality, let v be an arbitrary vector, and note by Cauchy-Schwartz that

$$\left\|\mathbb{E}[v^\top YZ]\right\|_2 = \sup_{\|u\|_2 \le 1} \mathbb{E}[(v^\top Y)(Z^\top u)] \le \sqrt{\mathbb{E}[(v^\top Y)^2]} \sup_{\|u\|_2 \le 1} \sqrt{\mathbb{E}[(Z^\top u)]^2} \le \sqrt{v^\top \mathbb{E}[YY^\top]v}.$$

This is equivalent to what we desired to show, giving the claim (16) and completing the proof. $\square$

A Proofs of technical results

A.1 Proof of Lemma 3.2

Let $a, b, c, d \geq 0$, where we implicitly take $a = dM_1(\theta \mid \xi)$, $b = d\nu_1(\xi)$, $c = dM_2(\theta \mid \xi)$, and $d = d\nu_2(\xi)$. Then the left integrand is

$$\frac{1}{2}\int_\Xi \int_\Theta |dM_1(\theta \mid \xi) - dM_2(\theta \mid \xi)|(d\nu_1(\xi) + d\nu_2(\xi)) = \frac{1}{2}\int_\Xi \int_\Theta |a - c|(b + d).$$

As

$$|a-c|(b+d) = |ab-cb|+|ad-cd| = |ab-cd+c(d-b)|+|ab-cd+a(d-b)| \leq 2|ab-cd|+(a+c)|b-d|,$$

we immediately obtain that

$$\begin{aligned}
&\int \|M_1(\cdot \mid \xi) - M_2(\cdot \mid \xi)\|_{\mathrm{TV}} \, (d\nu_1 + d\nu_2)(\xi) \\
&\leq \int_{\Xi\times\Theta} |d\nu_1(\xi)dM_1(\theta \mid \xi) - d\nu_2(\xi)dM_2(\theta \mid \xi)| + \frac{1}{2}\int_{\Xi\times\Theta} (dM_1(\theta \mid \xi) + dM_2(\theta \mid \xi))|d\nu_1(\xi) - d\nu_2(\xi)| \\
&\leq 2\,\|M_1 - M_2\|_{\mathrm{TV}} + 2\,\|\nu_1 - \nu_2\|_{\mathrm{TV}}\,,
\end{aligned}$$

because $\int dM_i(\theta \mid \xi) \leq 1$. Then noting that

$$\int_\Xi |d\nu_1 - d\nu_2| = \int_\Xi \left|\int_\Theta dM_1(\theta,\xi) - \int_\Theta dM_2(\theta,\xi)\right| \leq \int_\Xi \int_\Theta |dM_1(\theta,\xi) - dM_2(\theta,\xi)| = 2\,\|M_1 - M_2\|_{\mathrm{TV}}$$

gives the final result.

References

[1] P. Bickel, C. A. J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer Verlag, 1998.

[2] L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

[3] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.