Superefficient estimation and Le Cam's rescue

Paul Delatte
delatte@usc.edu
University of Southern California

Last updated: 08 November, 2022

1 Efficient estimation

1.1 A little history of the Gauss–Fisher–Le Cam efficiency

There exists a number of criteria to evaluate estimators $\hat{\theta}_n$ of a "true parameter" θ_0 (picked by nature among indexes $\theta \in \Theta$ of a family of Borel probability measures $\{\mathbb{P}_\theta : \theta \in \Theta\}$). A natural objective is that $\hat{\theta}_n$ should be close to θ_0 in some sense. Many of the evaluation criteria can be grounded more or less directly in a decision-theoretic framework built up against a notion of risk. A barer approach (which can nonetheless be connected in several ways to the statistical decision-theoretic framework) proceeds directly by looking at the moments of the estimator $\hat{\theta}_n$ under $\mathbb{P}_{\theta_0}$ to see how well the estimator concentrates around θ_0. (In the one-dimensional case, it is natural to look at the mean and variance of $\hat{\theta}_n$ under $\mathbb{P}_{\theta_0}$. The same idea generalizes to higher dimensions up to infinity.[1]) In general, finite sample comparison is complicated, so we look at what happens asymptotically, that is, when the number n of observations grow to $+\infty$. In this case, it is natural to look at the moments of the limiting distribution of $\hat{\theta}_n$ (under $\mathbb{P}_{\theta_0}$).

"In 1922, Fisher conjectured that, for [one-dimensional] regular [enough] models [...]:

(i) the ML estimator converges with asymptotic variance $I(\theta_0)^{-1}$;

(ii) if, for a sequence of estimators $(\hat{\theta}_n)$, we have $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow_{\mathbb{P}_{\theta_0}} N(0, v(\theta_0))$, then $v(\theta_0) \geq I(\theta_0)^{-1}$." (*trans. from Moulines*)

Fisher was right for (i). If Fisher was right for (ii), then it would have been possible to easily define a notion of asymptotic efficiency for estimators in regular enough models (and to show, in particular, that ML estimators are asymptotically efficient).[2] (The idea that we would a priori focus on normal limiting distributions for defining asymptotic efficiency as in (ii) is not as restrictive as could be intuitively thought: normality as a limiting distribution is in some sense "best" when it comes to concentration; this will be demonstrated formally later).

It turns out that Fisher was not exactly right for (ii). A number of works was completed in the 1930s and 1940s which increased the general belief that (ii) was true. The "sharp" asymptotic properties of MLE conjectured in (i) were formally proved between 1934 and 1949 following the work of Doob and Wald. Another important step (which

[1] In the infinite-dimensional case, we deal with stochastic processes whose concentration can be evaluated pointwise under some regularity conditions.

[2] This implied notion of Fisher efficiency in the one-dimensional case generalizes straightforwardly to higher dimensions up to infinity with the proper extensions of the Fisher information and limiting distributions (in particular, to tangent space and Gaussian processes in the infinite-dimensional case).

turned out to be unconnected) was achieved between 1943 and 1946 with the derivation of the (Fréchet–)Cramer–Rao lower bound[3] which states that the minimal variance of an unbiased estimator (under some regularity conditions) is the inverse of the Fisher information. This finite sample result was (erroneously) believed to confirm the validity of (ii) for asymptotic efficiency: an estimator with normal limit $N(0, v(\theta_0))$ is approximately unbiased, so it should hold that no estimator can be better that the one with limiting distribution $N(0, I(\theta_0)^{-1})$. (This connection between the bound and asymptotic efficiency was stated explicitly by Cramer in his book of 1946 and many times later by other statisticians; this interpretation of the bound is however misleading – as clear in the proofs, there is no explicit connection between it and the asymptotic efficiency of estimators). Based on the proved asymptotics of MLE and the newly derived information bounds, the general state of beliefs in the late 1940s was that the conjecture (ii) was true. However, in 1951, Hodges provided a counterexample to the conjecture by exhibiting a superefficient estimator of the Gaussian mean, that is, an estimator with normal limit and asymptotic variance lower than $I(\theta_0)^{-1}$. Estimator of this sort are bad (as is seen for fixed n at parameters close to θ_0), but not all superefficient estimators are bad, as proved by Stein in 1956 with the introduction of shrinkage estimators (for dimensions $d \geq 3$). If the badly behaved supperefficient estimators of Hodges's type could be easily discarded, the good performing estimators of Stein's type could not. They forced statisticians to give a definitive answer to the correctness of Fisher's conjecture.

It appeared, notably through the work of Le Cam, that the conjecture (ii) of Fisher could be largely salvaged. This is true in particular for parametric estimation. In this case, two broad solutions were devised. A first solution worked out by Hajek and Le Cam is to consider local maximum risk (LAM) (but at the cost of hiding differences – see vdW in The Statistical Work of Lucien Le Cam). A second solution worked out by Le Cam from 1953 is to show that in the parametric case superefficiency only happens on negligible sets. Thereafter, the idea that best estimators have normal distribution with asymptotic variance the inverse of the Fisher information almost holds in practice (and this can be made precise).

If all is (almost) good with the Gauss–Fisher–Le Cam approach to efficiency in parametric estimation, things are slightly more complicated in semiparametric and nonparametric settings. In particular, superefficiency becomes a much bigger problem which does not disappear as in parametric estimation (hence the necessity to consider some uniform criteria such as LAM). Using the LAM approach (with limit experiments), a lot can still be salvaged for regular enough models. In this context, a lot of work has been done but many questions remain unanswered: in particular, lower bounds have been worked out in many cases, but their sharpness is often an issue. For fully nonparametric models (which are not regular enough), the approach proves limited (or difficult): this justifies considering directly a full minimax approach which makes comparing estimators much more manageable.

[3] Mauriche Fréchet is the first to have proved the bound in 1943. The same result was independently obtained in 1945 by Rao and then in 1946 by Cramer.

1.2 Fisher's conjecture and superefficiency

1.2.1 Asymptotics of maximum likehood estimators

Proposition 1 (Consistency of MLE (1949))**.** *Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be Borel probability measures with densities $\{p_\theta : \theta \in \Theta\}$. Suppose there exists $\hat{\theta}_n^{\mathrm{ML}} \in \arg\max_\theta L(\theta; X_1, \ldots, X_n)$ where $X_1, \ldots, X_n \sim_{iid} \mathbb{P}_{\theta_0}$. If*

1. Θ is compact;

2. $\mathbb{E}_{\theta_0}\| \log \frac{p_\theta}{p_{\theta_0}} \|_\infty < +\infty$;

3. for x a.e., $\theta \mapsto p_\theta(x)$ is continuous;

4. $\mathbb{P}_\theta \neq \mathbb{P}_{\theta_0}$ for all $\theta \neq \theta_0$;

then

$$\hat{\theta}_n^{\mathrm{ML}} \xrightarrow{\mathbb{P}_{\theta_0}} \theta_0.$$

Proof. T.9.9. in Keener TS p.157 or T.2.3.1. in Moulines p.129 (where the condition of existence of an MLE is even relaxed to asymptotic existence of an MLE). For a slightly different version, see T.5.1. in Tsybakov p.117. See also S.5.2. and S.5.5. in vdW AS p.44 and p.61. □

Proposition 2 (AN of MLE (1934))**.** *Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be Borel probability measures with densities $\{p_\theta : \theta \in \Theta\}$. Suppose there exists $\hat{\theta}_n^{\mathrm{ML}} \in \arg\max_\theta L(\theta; X_1, \ldots, X_n)$ where $X_1, \ldots, X_n \sim_{iid} \mathbb{P}_{\theta_0}$. Suppose that*

1. $\Theta \subseteq \mathbb{R}^d$ open;

2. $\hat{\theta}_n^{\mathrm{ML}}$ is consistent.

Suppose further that there exists an open neighborhood $\Theta_0 \subseteq \Theta$ of θ_0 such that:

3. the set $A = \{x : p_\theta(x) > 0\}$ is the same for all $\theta \in \Theta_0$;

4. for a.e. x, $\theta \in \Theta_0 \mapsto p_\theta(x)$ is twice continuously differentiable;

5. there exist positive measurable functions g and h such that for all $\theta \in \Theta_0$ and x (μ-)a.e., $\|\nabla l(\theta; x)\|^2 < h(x)$, $\|H_l(\theta; x)\| \leq h(x)$, and $p_\theta(x) \leq g(x)$, and such that $\int_E (1 + h(x))g(x)\, d\mu(x) < \infty$ (where μ is dominating $\mathbb{P}_\theta$ for all θ).

Suppose finally that

5. the matrix $I(\theta_0) := -\mathbb{E}_{\theta_0}(H_l(\theta_0; X))$, which is well defined from (4-5), is invertible.

Then

$$\sqrt{n}(\hat{\theta}_n^{\mathrm{ML}} - \theta_0) \rightsquigarrow_{\mathbb{P}_0} N(0, I(\theta_0)^{-1}).$$

Proof. T.9.14. in Keener TS p.158 or T.2.3.3. in Moulines p.131 or T.5.2. in Tsybakov notes p.126 or T.14 in Loubes p.33. The conditions stated are sufficient, but not necessary (weaker versions exist, in particular by Le Cam – see T.5.39 p.65 in S.5.5. in vdW AS). Our version is simply localized at θ_0 (and we use that matrix inversion is continuous as in Wellner). The condition that $\hat{\theta}_n^{\mathrm{ML}} \in \arg\max_\theta L(\theta; X_1, \ldots, X_n)$ can be weakened as in Moulines. Condition (5) implies, in particular, that for all $\theta \in \Theta_0$, $x \mapsto \nabla_\theta l(\theta; x)$ is ($\mathbb{P}_\theta$-) square integrable and $x \mapsto H_l(\theta; x)$ is ($\mathbb{P}_\theta$-) integrable and that $\int p_\theta(x)\, d\mu(x)$ can be differentiated twice with respect to $\theta \in \Theta_0$ under the integral sign, but it is also required to apply the dominated convergence theorem as used in the proof. See S.5.2. in Pfanzagl p.113 (for details on conditions and explanation of simpler versions using integral expansion, as in Moulines or Tsybakov, than proofs using Taylor–Lagrange expansions, as in Loubes or Keener, following Cramer (1946)). □

History. The idea of MLE was known at least from 1760. Fisher is the one who gave MLE its central place in statistics with an important advocacy endeavor in 1922. The asymptotic distribution of MLE starts with Pearson and Filon (1986), efficiency of MLE was already suggested by Edgeworth (1908-9), Fisher gave the name of ML in 1922 and conjectured in a broader context the good asymptotic properties of MLE. The hard part in the CAN proof is the consistency, which (was not investigated by Fisher and) had to wait for Wald (1949) for a correct proof (earlier attempts were all faulty, in particular Cramer (1946)). Asymptotic normality given consistency is more straightforward and was proved correctly by Doob (1934) under some restrictive hypotheses. (More general results were derived later.) (See Johann Pfanzagl Mathematical Statistics: Essays on History and Methodology.)

1.2.2 (Fréchet–)Cramer–Rao lower bound

Proposition 3 ((Fréchet)–Cramer–Rao Lower Bound (1943-1946))**.** *Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be Borel probability measures with densities $\{p_\theta : \theta \in \Theta\}$. Let $T\colon E \to \mathbb{R}^d$ be a ($\mathbb{P}_\theta$-) square integrable statistic such that $\mathbb{E}_\theta(T)$ is differentiable. Suppose that:*

1. $\Theta \subseteq \mathbb{R}^d$ open;

2. the set $A = \{x : p_\theta(x) > 0\}$ is the same for all $\theta \in \Theta$;

3. for x a.e., $\theta \mapsto p_\theta(x)$ is differentiable, and $x \mapsto \nabla_\theta l(\theta; x)$ is ($\mathbb{P}_\theta$-) square integrable;

4. for all $\theta \in \Theta$, $I(\theta) := \mathbb{E}_\theta(\nabla_\theta l(\theta; X)\nabla_\theta l(\theta; x)^T)$, which is well-defined from (3), is invertible;

5. $\int p_\theta(x)\, d\mu(x)$ and $\int T(x) p_\theta(x)\, d\mu(x)$ can be differentiated with respect to θ under the integral sign.

Then, for all $\theta \in \Theta$,

$$\mathrm{Var}_\theta(T) \geq \nabla_\theta \mathbb{E}_\theta(T)^T I(\theta)^{-1} \nabla_\theta \mathbb{E}_\theta(T).$$

Proof. T.2.2. in Wellner notes p.83 or T.5. in Barra NFSM p.38 or T.5.10. in Lehmann&Casella TPE p.120. (See also T.4.9. In Keener TS p.92 and T.1.4.28. in Moulines p.70, both for unbiased estimators. See Tsybakov and Moulines for sufficient integrability conditions under which (5) is true.) The first condition in (5) is not necessary for the existence of the Fisher information (as we define it in the proposition), but it is required to express it as a variance, and so proved the result by using the covariance inequality (derived from Cauchy–Schwarz). □

2 Superefficient estimation

Given Borel probability measures $\{\mathbb{P}_\theta : \theta \in \Theta\}$, an estimator with limit distribution $N(0, V^2(\theta))$ such that $V^2(\theta) < 1/I(\theta)$ for at least one $\theta \in \Theta$ is said to be **(Fisher) superefficient**. The definition naturally generalizes in higher dimensions.

Proposition 4 (Hodges's Superefficient Estimator)**.** *Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, 1)$ so that $I(\theta) = 1$. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, $|a| < 1$, and define*

$$T_n = \begin{cases} \bar{X}_n & \text{if} \quad |X_n| > n^{-1/4}, \\ a\bar{X}_n & \text{if} \quad |X_n| \leq n^{-1/4}. \end{cases}$$

Then

$$\sqrt{n}(T_n - \theta) \rightsquigarrow N(0, V^2(\theta))$$

where

$$V^2(\theta) = \begin{cases} 1 & \textit{if} \quad \theta \neq 0, \\ a^2 & \textit{if} \quad \theta = 0, \end{cases}$$

and so $V^2(\theta) < 1/I(\theta)$ *at* $\theta = 0$*, that is,* T_n *is Fisher superefficient.*

Proof. E.3.1. in Wellner notes p.98 or E.16.1. in Keener TS p.320 or E.8.1. in vdW AS p.109. □

Hodges's estimator is thus supperefficient, which seems to contradict the many optimality properties of the sample mean in the GLM. However, superefficieny in this case comes at a price (see E.16.1. in Keener TS). Let us consider the (n rescaled) risk under squared error loss, namely $nR_n(T_n, \theta) = n\mathbb{E}_\theta[(T_n - \theta)^2]$. It is easily seen that $nR_n(\bar{X}_n, \theta) = 1$. We also have $\lim_{n\to\infty} nR_n(T_n, \theta) = V^2(\theta)$, but, for fixed n, the scaled risk $nR_n(T_n, \theta)$ is close to a^2 at $\theta = 0$ and close to 1 everywhere else except on each side of $\theta = 0$ where it peaks to $+\infty$ as $n \longrightarrow \infty$. "The conclusion is that T_n "buys" its better asymptotic behavior at $\theta = 0$ at the expense of erratic behavior close to 0. Because the values of θ at which T_n is bad differ from n to n, the erratic behavior is not visible in the pointwise limit distributions under fixed θ." (vdW)

The next example due to Stein (1956) exhibits an estimator that is not Fisher supperefficient (from non-normal distribution – see R.7.2. Wellner notes p.188) but that satisfies another form of superefficiency (which, defined as a non-asymptotic notion, takes care of the "fixed n deficiencies" of Fisher superefficiency). An estimator is said[4] to be **Stein superefficient** if it is not Fisher efficient and its squared error loss risk (MSE) is lower than that of a Fisher efficient estimator for all $\theta \in \Theta$ and the inequality is strict for at least one $\theta \in \Theta$. This thus translates in terms of domination and admissibility[5] (which, unless stated, are always taken with respect to MSE): an estimator that is not Fisher efficient is Stein supperefficient if and only if it dominates a Fisher efficient estimator; if there exists a Stein supperefficient estimator, then there is at least one Fisher efficient estimator that is inadmissible; a Stein superefficient estimator need not be admissible (as it can be dominated by another Stein superefficient estimator). In some sense, Stein superefficient estimators are much more attractive than (some of) the pathological Fisher superefficient estimators, for they improve without trade-off upon an (almost noncontroversial) integrated measure of efficiency (whereas Hodges's estimator does not).

Proposition 5 ((James–)Stein's Skrinkage Estimator)**.** *Let* $X_1, \ldots, X_n$ *be i.i.d.* $N(\theta, I_d)$ *where* $d \geq 3$*. Let* $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$*, and define*

$$T_n = \bar{X}_n\left(1 - \frac{d-2}{n\|\bar{X}_n\|^2}\right).$$

[4] This is our own convention: one limitation of the definition is that Fisher efficient estimators are not unique and may have different MSE, so that a Stein superefficient estimator may exist and still be dominated by a Fisher efficient estimator. (Note that this is a conjecture in need of an example.)

[5] For a given risk function R, a rule δ is said to be inadmissible if there is a rule δ' such that $R(\delta, \theta) \geq R(\delta', \theta)$ for all $\theta \in \Theta$ and $R(\delta, \theta) > R(\delta', \theta)$ for some $\theta \in \Theta$. A rule such as δ' is said to be dominating the rule δ.

Then $\bar{X}_n$ is UMVU and asymptotically efficient in the sense of Gauss–Fisher, but $\bar{X}_n$ is inadmissible with respect to MSE. Indeed, T_n dominates $\bar{X}_n$ with respect to MSE, and so T_n is Stein superefficient. In particular,

$$n\mathbb{E}_\theta[\|T_n - \theta\|^2] = d - \mathbb{E}_\theta\left[\frac{d-2}{n\|\bar{X}_n\|^2}\right] < d = n\mathbb{E}_\theta[\|\bar{X}_n - \theta\|^2].$$

Proof. T.7.4. in Wellner notes p.185 or T.3.3. in Tsybakov INE p.162 or E.8.12. in vdW AS p.119. or T.11.3. in Keener TS p.210. Note that the result is often stated for $n = 1$, in which case $\bar{X}_n = X_1$ (see S.5.4. in Tsybakov INE p.155 for the equivalence in terms of risk between the models – this is also the link between parametric and nonparametric in the interpretation of the Gaussian sequence model). □

Remark. Even if the James–Stein estimator is Stein superefficient, it is itself inadmissible. (see Wellner R.7.3. p.188).

It can be further proved (see Wellner) that

$$\frac{\mathbb{E}_\theta[\|T_n - \theta\|^2]}{\mathbb{E}_\theta[\|\bar{X}_n - \theta\|^2]} = 1 - \frac{d-2}{d}\mathbb{E}_0\left[\frac{d-2}{\chi^2_d(n\|\theta\|^2/2)}\right] \begin{cases} = 2/d & \text{if} \quad \theta = 0, \\ \longrightarrow 1 & \text{as} \quad n \longrightarrow \infty \text{ for fixed } \theta \neq 0, \\ \longrightarrow 1 & \text{as} \quad \|\theta\| \longrightarrow \infty \text{ for fixed } n. \end{cases}$$

This shows that the improvement of the James–Stein estimator over the Fisher efficient estimator is asymptotically negligible (see Tsybakov INE p.162). This suggests that looking asymptotically at uniform (in θ) measure of performance should rescue the Gauss–Fisher notion of efficiency. (Note, however, that in light of what happens in the non-asymptotic regime, the existence of Stein superefficiency remains a major conundrum for the design of good estimation procedures!)

3 Le Cam, Kaufman, and Hajek to the rescue of Fisher efficiency

The discussions following the examples suggest two intuitive solutions to salvage the Gauss–Fisher approach from the pathologies of superefficiency, either:

- to restrict the class of estimators under consideration (in particular, only consider estimators that are (locally) regular in the sense that their limit distribution does not depend on the direction of approach of θ to θ_0 – Hodges's and Stein's estimators are not locally regular in this sense as shown in vDw AS p.119 and Wellner notes p.99 and p.188); or,

- to look at (locally) uniform measure of risks (in particular, LAM) (The virtue of this approach is clear in the case of Hodges's estimator – as the improvement is only "pointwise" – but less so for the shrinkage estimator of Stein; the fact that it is also true can be seen asymptotically, by remembering that Stein superefficiency is a non-asymptotic concept).

A third solution is to investigate where supperfficiency occurs. If superefficiency is measurably negligible, why care? Le Cam showed in 1953 that for parametric problems superefficiency only happens on sets of Lebesgue measure zero.

3.1 The "null set" rescue

Proposition 6 (Le Cam, 1953)**.** *Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be Borel probability measures with densities $\{p_\theta : \theta \in \Theta\}$. Suppose that:*

1. $\Theta \subseteq \mathbb{R}^d$ is open;

2. for x a.e., $\theta \in \Theta_0 \mapsto p_\theta(x)$ is twice continuously differentiable;

3. for all $\theta_0 \in \Theta$, there exist an open neighborhood $\Theta_0 \subseteq \Theta$ of θ_0 and a positive measurable functions g and h such that for all $\theta \in \Theta_0$ and x (μ-)a.e., $\|\nabla l(\theta; x)\|^2 < h(x)$, $\|H_l(\theta; x)\| \leq h(x)$, and $p_\theta(x) \leq g(x)$, and such that $\int_E (1+h(x))g(x)\, d\mu(x) < \infty$ (where μ is dominating $\mathbb{P}_\theta$ for all θ);

4. for all $\theta \in \Theta$, the matrix $I(\theta) := -\mathbb{E}_\theta(H_l(\theta; X))$, which is well defined from (2-3), is invertible.

If T_n is a statistic such that $\sqrt{n}(T_n - \theta) \rightsquigarrow_{\mathbb{P}_\theta} N(0, v(\theta))$ for some positive semi-definite matrix $v(\theta)$, then there exists $\Theta_N \subseteq \Theta$ such that Θ_N has Lebesgue measure zero and $v(\theta) - I^{-1}(\theta)$ is positive semi-definite for all $\theta \notin \Theta_N$.

Proof. T.4.16. in Shao MS p.287 which is based on Bahadur (1964) "On Fisher's Bound for Asymptotic Variances" which is itself a relaxation of regularity conditions of Le Cam (1953). We use lightly different integrability conditions than Bahadur for consistency with previous results. □

3.2 The "convolution and LAM theorems" rescue

One solution to salvage Fisher's conjecture is to restrict the class of estimators under consideration; this is done by the introduction of regular estimators.

Definition 7 ((Locally) Regular Estimator)**.** Let $(\mathbb{P}_\theta : \theta \in \Theta)$ be Borel probability measures. An estimator T_n is said to be a **(locally) regular** estimator of θ at $\theta_0 \in \Theta$ if for every sequence $(\theta_n)_{n\in\mathbb{N}}$ in Θ with $\lim_{n\to\infty} \sqrt{n}(\theta_n - \theta_0) = h \in \mathbb{R}^k$,

$$\sqrt{n}(T_n - \theta_n) \rightsquigarrow_{\mathbb{P}_{\theta_n}} L_{\theta_0},$$

where L_{θ_0} is a distribution that depends on θ_0 but not on h.

Remark. 1. This is seen to be equivalent to the condition that: for every $h \in \mathbb{R}^k$,

$$\sqrt{n}\left(T_n - \theta_0 - \frac{h}{\sqrt{n}}\right) \rightsquigarrow_{\mathbb{P}_{\theta_0+h/\sqrt{n}}} L_{\theta_0}.$$

The next result, known as the convolution theorem, initiated by Kaufman and developed by Hajek, shows that among regular estimators, those with normal limit and inverse Fisher asymptotic variance are "best". The theorem is stated by Hajek under weak regularity conditions which have become standard in the literature. We thus reproduce (some version of) them. (As already pointed out, many of the previous results requiring twice differentiability can be restated under these weaker conditions or similar ones.)

Proposition 8 (**Convolution Theorem**)**.** *Let $(\mathbb{P}_\theta : \theta \in \Theta)$ Borel probability measures where $\Theta \subseteq \mathbb{R}^d$ is open. If $(\mathbb{P}_\theta : \theta \in \Theta)$ is differentiable in quadratic mean at $\theta_0 \in \Theta$ with invertible Fisher information matrix $I(\theta_0)$ and T_n is a regular estimator of θ at θ_0 with scaled limit distribution L_{θ_0}, then there exists a probability measure M_{θ_0} such that*

$$L_{\theta_0} = N(0, I^{-1}(\theta_0)) * M_{\theta_0}.$$

In particular, if L_{θ_0} has variance $v(\theta_0)$, then $v(\theta_0) - I^{-1}(\theta_0)$ is positive semidefinite.

Proof. T.8.8. in vdW AS p.115 or T.4.1. in Wellner notes p.110 or T.2.3.1. in Bickel&Wellner EAESM p.24. The conditions can be relaxed, in particular to LAN. □

This equivalently says, if we denote $\mathbb{Z}_\theta \sim L_{\theta_0}$ the scaled weak limit of T_n, that there are random variables Z_{θ_0} and Δ_{θ_0}, independent of one another, such that $\mathbb{Z}_{\theta_0} = Z_{\theta_0} + \Delta_{\theta_0}$ where $Z_{\theta_0} \sim N(0, I^{-1}(\theta_0))$ and $\Delta_{\theta_0} \sim M_{\theta_0}$. "In words, [this] says that the [scaled] limiting distribution of any regular estimator T_n of θ must be at least as “spread out” as the $N(0, I^{-1}(\theta_0))$ distribution of Z_{θ_0}." (Wellner p.110) Thus among regular estimators we could legitimately define an (asymptotically) efficient estimator as one for which the scaled limiting distribution is exactly $N(0, I^{-1}(\theta_0))$. This can be restated in terms of asymptotic optimality with respect to bowl-shaped loss functions, by applying Anderson’s lemma.

Remark. See S.3.8. in Pfanzagl MS p.59 for why the interpretation of Wellner is somehow precarious, and we need to rely on Anderson’s lemma for a better interpretation than the "spreading out" of convolutions.

Definition 9 (Bowl/Bridge-Shaped Function)**.** Let $l \colon \mathbb{R}^d \to \mathbb{R}^+$ be a function such that $l(x) = l(-x)$ for all x. If $\{x : l(x) \leq c\}$ is convex for every $c \geq 0$, then l is said to be bowl-shaped. If $\{x : l(x) \geq c\}$ is convex for every $c \geq 0$, then l is said to be bridge-shaped.

Lemma 10 (Anderson’s Lemma)**.** *If X is a random variable with values in $\mathbb{R}^d$ and bridge-shaped density and if $l \colon \mathbb{R}^d \to \mathbb{R}^+$ is a bowl-shaped function such that $\mathbb{E}\left[l(X+c)\right] < \infty$ for all $c \in \mathbb{R}^d$, then*

$$\mathbb{E}\left[l(X+c)\right] \geq \mathbb{E}\left[l(X)\right]$$

for all $c \in \mathbb{R}^d$.

Proof. T.16.17. in Keener TS p.331 or L.10.2. in Ibragimov&Has’minskii SE p.157. This is actually a corollary of Anderson’s lemma (which can be found in L.10.1. in Ibragimov&Has’minskii SE p.155 or S.3.8. in Pfanzagl MS p.59). Applying the corollary, which can be expressed through convolutions, to centered multivariate normals yields the result L.8.5. in vdW AS p.113 (see C.5.7. in Hopfner AS p.135 for a proof). See also "The Brunn-Minkowski Inequality" (2002) by Gardner for broader perspective. □

Corollary 11 (Hajek, 1970))**.** *Suppose the conditions of the convolution theorem hold for $(\mathbb{P}_\theta : \theta \in \Theta)$ and T_n. If $l \colon \mathbb{R}^d \to \mathbb{R}^+$ is bowl-shaped, then*

$$\liminf_{n\to\infty} \mathbb{E}_{\theta_0}\left[l(\sqrt{n}(T_n - \theta_0))\right] \geq \mathbb{E}_{\theta_0}\left[l(Z_{\theta_0})\right]$$

where $Z_{\theta_0} \sim N(0, I^{-1}(\theta_0))$.

Proof. C.1. in Wellner p.110 and "Asymptotic optimality theorem" in Bickel&Wellner p.26. This obtains by combining the convolution theorem with Anderson’s lemma. □

Proposition 12 (**LAM Theorem** (Hajek, 1972))**.** *Let $(\mathbb{P}_\theta : \theta \in \Theta)$ Borel probability measures where $\Theta \subseteq \mathbb{R}^d$ is open. Let T_n be any estimator. If $(\mathbb{P}_\theta : \theta \in \Theta)$ is differentiable* *in quadratic mean at $\theta_0 \in \Theta$ with invertible Fisher information matrix $I(\theta_0)$ and $l\colon \mathbb{R}^d \to \mathbb{R}^+$ is a bowl-shaped function, then for any $\delta > 0$,*

$$\liminf_{n\to\infty} \sup_{\{\theta : \|\theta-\theta_0\|<\delta\}} \mathbb{E}_{\theta}[l(\sqrt{n}(T_n-\theta))] \geq \mathbb{E}\,[l(Z_{\theta_0})]$$

where $Z_{\theta_0} \sim N(0, I^{-1}(\theta_0))$.

Proof. T.12.1. in Ibragimov&Has'minskii SE p.162 for $\varphi(\varepsilon) = n^{-1/2}$. Then R.2.12.2. in Ibragimov&Has'minskii SE p.168 allows to restate the inequality as in T.4.2. in Wellner notes p.110 or "Locally asymptotic minimax theorem" in Bickel&Wellner p.27 or T.16.25. in Keener TS p.340, that is,

$$\lim_{h\to+\infty} \liminf_{n\to\infty} \sup_{\{\theta : \sqrt{n}\|\theta-\theta_0\|\leq h\}} \mathbb{E}_{\theta}[l(\sqrt{n}(T_n-\theta))] \geq \mathbb{E}\,[l(Z_{\theta_0})]$$

Another more refined version is T.8.11. in vdW AS p.117-118, which yields

$$\sup_{I} \liminf_{n\to\infty} \sup_{h\in I} \mathbb{E}_{\theta_0+h/\sqrt{n}}[l(\sqrt{n}(T_n-\theta_0+h/\sqrt{n}))] \geq \mathbb{E}\,[l(Z_{\theta_0})]$$

where the first supremum runs over all finite subsets I of $\mathbb{R}^d$. The result in vdV is proved using the weak topology for experiments. (It is proved in greater generality in 3.11.5. in WCEP vdW&Wellner p.417.) The condition of QMD with invertible Fisher can again be weakened to LAN. The finer result as in vdV can be extended to other limit experiments (see e.g. T.5. in Pollard's thoughts (2000), Lecture 7 in Pollard's Paris lectures (2021), S.7.4. in Torgersen (1991), S.62. in Strasser (1985), or S.5. in vdV (2002)). □

References

Bahadur, R. R. (1964): "On Fisher's bound for asymptotic variances," *The Annals of Mathematical Statistics*, 35(4), 1545–1552.

Bickel, P. J., C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov (1993): *Efficient and adaptive estimation for semiparametric models*, vol. 4. Springer.

Fort, G., M. Lerasle, and E. Moulines (2020): "Statistique et apprentissage," *Notes du cours MAP433 at Polytechnique*.

Gardner, R. (2002): "The brunn-minkowski inequality," *Bulletin of the American mathematical society*, 39(3), 355–405.

Höpfner, R. (2014): *Asymptotic statistics: with a view to stochastic processes*. Walter de Gruyter.

Ibragimov, I. A., and R. Z. Has' Minskii (1981): *Statistical estimation: asymptotic theory*. Springer.

Keener, R. W. (2010): *Theoretical statistics: Topics for a core course*. Springer.

Le Cam, L. (1953): "On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates," *Univ. Calif. Publ. in Statist.*, 1, 277–330.

Lehmann, E. L., and G. Casella (1998): *Theory of point estimation*. Springer.

Loubes, J.-M. (2010): "Cours de statistique asymptotique," *Lecture notes at Toulouse*.

Pfanzagl, J. (2017): *Mathematical statistics: essays on history and methodology*. Springer.

Pollard, D. (2000): "Some thoughts on Le Cam's statistical decision theory," *Preprint*.

——— (2001): "Lecture in Paris on Le Cam's theory," *Lecture notes*.

Shao, J. (2006): *Mathematical statistics*. Springer.

Stein, C. (1956): "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, vol. 3, pp. 197–207. University of California Press.

Strasser, H. (1985): *Mathematical theory of statistics: statistical experiments and asymptotic decision theory*, vol. 7. Walter de Gruyter.

Torgersen, E. (1991): *Comparison of statistical experiments*, vol. 36. Cambridge University Press.

Tsybakov, A. (2014): "Introduction aux methodes statistiques," *Lecture notes for MAP433 at Polytechnique*.

Tsybakov, A. B. (2008): *Introduction to nonparametric estimation*. Springer.

van der Vaart, A. (2002): "The statistical work of lucien le cam," *The Annals of Statistics*, 30(3), 631–682.

van der Vaart, A., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer.

Van der Vaart, A. W. (1998): *Asymptotic statistics*. Cambridge University Press.

Wellner, J. A. (2005): "Advanced theory of statistical inference," *Lecture notes for STAT580 at UWash*.