ON THE ATTAINMENT OF THE CRAMÉR-RAO BOUND IN $\mathbb{L}_r$-DIFFERENTIABLE FAMILIES OF DISTRIBUTIONS

Ulrich MÜLLER-FUNK, Universität Münster
Friedrich PUKELSHEIM, Cornell University*
Hermann WITTING, Universität Freiburg im Breisgau

BU-941-M† September 1987

ABSTRACT

A rigorous proof is presented that global attainment of the Cramér-Rao bound is possible only when the underlying family of distributions is exponential. The proof is placed in the context of $\mathbb{L}_r(P_\vartheta)$-differentiability. This notion requires strong differentiability in $\mathbb{L}_r(P_\vartheta)$ of the r^{th} root of the likelihood ratio relative to P_ϑ.

* On leave from the Institut für Mathematik der Universität Augsburg. The work of this author is partially supported by the Mathematical Sciences Research Institute, Cornell University.

† This is paper BU-941-M in the Biometrics Unit series, Cornell University.

AMS 1980 subject classification. Primary 62F10.
Key words and phrases. Parametric families, regular experiments.

1. Introduction

It counts among the folklore results of parametric statistics that global equality in the Cramér-Rao inequality obtains only when the underlying family is exponential. A rigorous proof of this result depends on which concept of differentiability is adopted. Wijsman (1973) employs the logarithmic derivatives of the density functions, and solves the associated differential equation including a detailed discussion of the resulting measurability problems. Fabian and Hannan (1977) assume weak $\mathbb{L}_2$-differentiability of the likelihood ratio, see also Barankin (1949, Section 6).

We here place our derivation in the context of $\mathbb{L}_r$-differentiable families, that is, strong $\mathbb{L}_r$-differentiability of the r^{th} root of the likelihood ratio, thus evading any extra integrability assumptions. A detailed exposition of $\mathbb{L}_r$-differentiable families of distributions is given in the textbook Witting (1985). Ibragimov and Has'minskii (1981) work with regular experiments which essentially coincide with the continuous $\mathbb{L}_2$-differentiable families as introduced below. The notion of $\mathbb{L}_2$-differentiability is due to Hàjek (1962, p. 1124) and Le Cam (1966, Section 4).

Depending on the parameter $r \geq 1$ there evolves a hierarchy of differential smoothness that is statistical meaningful: $\mathbb{L}_1$-differentiability is appropriate for deriving locally optimal tests, see Witting (1985, Section 1.8.1), while $\mathbb{L}_2$-differentiablity applies to estimation problems, see Witting (1985, Section 2.7.2) or Ibragimov and Has'minskii (1981, Section I.7.2), and local asymptotic normality, see Ibragimov and Has'minskii (1981, Chapter II). Finally $\mathbb{L}_r$-differentiability, for all $r \geq 1$, holds in exponential families, reflecting their well appreciated smoothness properties.

In Section 2 we recall the Cramér-Rao inequality as it holds in $\mathbb{L}_2$-differentiable families. Global attainment leads to a differential equation whose coefficients are continuous once continuous $\mathbb{L}_r$-differentiability is assumed, as discussed in Section 3. The solution of this differential equation leads to exponential families, as detailed in Theorem 2.

2. The Cramér-Rao inequality

The notion of $\mathbb{L}_2$-differentiability is briefly reviewed since it is central to the version of the Cramér-Rao inequality that is presented in Theorem 1. Let $\mathcal{P} = \{ P_\vartheta : \vartheta \in \Theta \}$ be a family with parameter $\vartheta \in \Theta \subseteq \mathbb{R}^k$, on some fixed sample space $\mathcal{X}$ with sigma-algebra $\mathcal{B}$. The likelihood ratio of a member P_ϑ relative to another member P_{ϑ_0} is denoted by $L_{\vartheta/\vartheta_0}$, that is,

$$L_{\vartheta/\vartheta_0}(x) = p_\vartheta(x)/p_{\vartheta_0}(x) \in [\,0, \infty\,]$$

for $(P_\vartheta + P_{\vartheta_0})$-almost all x, whenever p_ϑ and p_{ϑ_0} are the respective densities of P_ϑ and P_{ϑ_0} with respect to some common dominating measure.

For an interior point ϑ_0 in Θ the family $\mathcal{P}$ is called $\mathbb{L}_2(P_{\vartheta_0})$-*differentiable* when there exists a k-dimensional statistic $\dot{L}_{\vartheta_0}$ with components which are r-fold integrable under P_{ϑ_0} such that for $\vartheta \to \vartheta_0$ one has

$$\| 2(L^{1/2}_{\vartheta/\vartheta_0} - 1) - (\vartheta - \vartheta_0)^\top \dot{L}_{\vartheta_0} \|_2 = o(|\vartheta - \vartheta_0|), \tag{1}$$

$$P_\vartheta(\{ L_{\vartheta/\vartheta_0} = \infty \}) = o(|\vartheta - \vartheta_0|^2), \tag{2}$$

where $\|T\|_2 = (\int T^2\, dP_{\vartheta_0})^{1/2}$ is the $\mathbb{L}_2(P_{\vartheta_0})$ norm while $|\vartheta| = (\vartheta^\top \vartheta)^{1/2} = (\sum_{i=1}^k \vartheta_i^2)^{1/2}$ is the Euclidean norm. If (1) and (2) are satisfied then the statistic $\dot{L}_{\vartheta_0}$ is P_{ϑ_0}-almost surely unique and is called *the* $\mathbb{L}_2$-*derivative of* $\mathcal{P}$ *at* ϑ_0, or for short, *the* $\mathbb{L}_2(P_{\vartheta_0})$-*derivative*. The covariance matrix

$$\mathcal{I}(\vartheta_0) = \mathcal{C}ov_{\vartheta_0}[\dot{L}_{\vartheta_0}] \tag{3}$$

is the *information matrix* of $\mathcal{P}$ at ϑ_0.

Property (2) means that the singular parts vanish of the right order. When the distributions are pairwise equivalent there are no singular parts, and property (2) is trivially satisfied. Property (1) pertains to the square root of the likelihood ratio rescaled by the factor 2. This rescaling is convenient in that the statistic $\dot{L}_{\vartheta_0}$ then also appears as the $\mathbb{L}_r(\vartheta_0)$-derivative for all $r \leq 2$. The notion of $\mathbb{L}_r$-differentiability makes sense for all $r \geq 1$ and simply replaces 2 by r in (1) and (2).

Theorem 1. *Suppose the family* $\mathcal{P} = \{ P_\vartheta : \vartheta \in \Theta \}$ *is* $\mathbb{L}_2(P_{\vartheta_0})$-*differentiable at an interior point* ϑ_0 *of* $\Theta \subseteq \mathbb{R}^k$. *For some dimension* $s \geq 1$ *let* T *be an* s-*dimensional statistic whose components have a finite variance in a neighborhood of* ϑ_0,

$$\limsup_{\vartheta \to \vartheta_0} \mathcal{V}ar_\vartheta[T_i] < \infty \qquad \text{for all } i = 1, \ldots, s. \tag{4}$$

Then the mean-value function $\gamma(\vartheta) = E_\vartheta[T]$ *is differentiable at* ϑ_0 *with Jacobian matrix* $\mathcal{G}(\vartheta_0)$, *say, and the covariance matrix obeys the* Cramér-Rao inequality

$$\mathcal{C}ov_{\vartheta_0}[T] \geq \mathcal{G}(\vartheta_0)\mathcal{I}(\vartheta_0)^- \mathcal{G}(\vartheta_0)^\top. \tag{5}$$

Moreover, equality holds in (5) if and only if

$$T(x) - \gamma(\vartheta_0) = \mathcal{G}(\vartheta_0)\mathcal{I}(\vartheta_0)^{-}\dot{L}_{\vartheta_0}(x) \tag{6}$$

for P_{ϑ_0}-almost all x.

Proof. For a proof see Witting (1985, Satz 2.133) or Ibragimov and Has'minskii (1981, Theorem I.7.3). The essential step is to establish the identity $\mathcal{G}(\vartheta_0) = E_{\vartheta_0}[T\dot{L}_{\vartheta_0}^{\top}]$ using assumption (4), see Witting (1985, Satz 2.136) or Ibragimov and Has'minskii (1981, Lemma 7.2). Once this identity is established one has

$$\mathcal{C}ov_{\vartheta_0}[\big(T - \gamma(\vartheta_0)\big) - \mathcal{G}(\vartheta_0)\mathcal{I}(\vartheta_0)^{-}\dot{L}_{\vartheta_0}] = \mathcal{C}ov_{\vartheta_0}[T] - \mathcal{G}(\vartheta_0)\mathcal{I}(\vartheta_0)^{-}\mathcal{G}(\vartheta_0)^{\top},$$

so that the inequality as well as the equality condition become evident. It is straightforward to verify that the expressions in (5) and (6) are invariant to the choice of the generalized inverse for $\mathcal{I}(\vartheta_0)$. ◊

There are other versions of the Cramér-Rao inequality some of which present the inequality as a joint property of the underlying family of distributions *and* the estimator under investigation, for an example see Joshi (1976). As pointed out by Pitman (1979, p. 39) a result as in Theorem 1 is appealing in that it essentially applies to all estimators, excepting only those that have an unbounded variance in the neighborhood of ϑ_0.

3. Attainment of the Cramér-Rao bound

Global attainment of the Cramér-Rao bound is discussed assuming that the dimensionality of the statistic T and the parameter ϑ coincide, $s = k$, and that the parameter domain Θ is open. From (5) we then obtain the equation

$$\mathcal{C}ov_\vartheta[T] = \mathcal{G}(\vartheta)\mathcal{I}(\vartheta)^-\mathcal{G}(\vartheta)^\top \qquad \text{for all } \vartheta \in \Theta.$$

When the covariance matrix of T is nonsingular then the matrices $\mathcal{G}(\vartheta)$ and $\mathcal{I}(\vartheta)$ are nonsingular as well, and equation (6) yields $\dot{L}_\vartheta = \mathcal{A}(\vartheta)^\top T - b(\vartheta)$ with $\mathcal{A}(\vartheta)^\top = \mathcal{I}(\vartheta)\mathcal{G}(\vartheta)^{-1}$ and $b(\vartheta) = \mathcal{I}(\vartheta)\mathcal{G}(\vartheta)^{-1}\gamma(\vartheta)$. In other words, the derivative $\dot{L}_\vartheta$ is an affine transformation of a statistic T where the coefficients $\mathcal{A}(\vartheta)$ and $b(\vartheta)$ depend on ϑ, but the statistic T does not. To solve this differential equation it is helpful to have the coefficients depend continuously on ϑ.

To this end we introduce continuous $\mathbb{L}_r$-differentiability. For our purposes we may assume that the family $\mathcal{P}$ consists of pairwise equivalent distributions, thus relieving us of the study of singular parts as in (2). The $\mathbb{L}_r(P_\vartheta)$-derivative $\dot{L}_\vartheta$ is a member of the space $\mathbb{L}_r^k(P_\vartheta)$, that is, it is a k-dimensional statistic whose components are r-fold integrable under P_ϑ. Multiplication with $L_{\vartheta/\vartheta_0}^{1/r}$ yields a member of the space $\mathbb{L}_r^k(P_{\vartheta_0})$. Therefore we say that *continuous* $\mathbb{L}_r(P_{\vartheta_0})$*-differentiability* holds when for $\vartheta \to \vartheta_0$ one has (1) with r in place of 2, and

$$\|\dot{L}_\vartheta L_{\vartheta/\vartheta_0}^{1/r} - \dot{L}_{\vartheta_0}\|_r = o(1), \tag{7}$$

where $\|S\|_r = \sum_{i=1}^k (\int S_i^r \, dP_{\vartheta_0})^{1/r}$ is the $\mathbb{L}_r^k(P_{\vartheta_0})$ norm.

It is straightforward to show that continuous $\mathbb{L}_2$-differentiability of $\mathcal{P}$ on Θ implies that the information matrix $\mathcal{I}(\vartheta)$ of (3) and the Jacobian matrix $\mathcal{G}(\vartheta)$ appearing in (5) depend continuously on ϑ. These continuity properties are automatic when regular experiments in the sense of Ibragimov and Has'minskii (1981, Section I.7.1) are assumed.

The only additional assumption not mentioned so far is that the parameter domain Θ ought to be connected so that any two points ϑ_0 and ϑ can be joined by a continuous path ϑ_s, $0 \le s \le 1$. The following theorem summarizes the discussion, leaving only the implication (c) $\Rightarrow$ (a) to be proved.

Theorem 2. *Suppose the family $\mathcal{P} = \{ P_\vartheta : \vartheta \in \Theta \}$ consists of pairwise equivalent distributions, with a parameter domain $\Theta \subseteq \mathbb{R}^k$ that is open and connected. Let T be a k-dimensional statistic whose distributions under $\mathcal{P}$ do not concentrate on a proper affine subspace of $\mathbb{R}^k$; when it exists the Jacobian matrix of its mean-value function $\gamma(\vartheta) = E_\vartheta[T]$ is denoted by $\mathcal{G}(\vartheta)$. Then the following three statements are equivalent:*

(a) $\mathcal{P}$ *is an exponential family in* $\alpha(\vartheta)$ *and* T *and of order* k*, for some continuously differentiable mapping* $\alpha : \Theta \to \mathbb{R}^k$ *whose Jacobian matrices* $\mathcal{A}(\vartheta)$ *have full rank* k*.*
(b) $\mathcal{P}$ *is continuously* $\mathbb{L}_r$*-differentiable on* Θ *for all* $r \geq 1$*, and the covariance matrices* $\mathcal{C}ov_\vartheta[T]$ *are continuous on* Θ*, of full rank* k*, and attain the Cramér-Rao bound* $\mathcal{C}ov_\vartheta[T] = \mathcal{G}(\vartheta)\mathcal{I}(\vartheta)^{-1}\mathcal{G}(\vartheta)^\top$ *for all* $\vartheta \in \Theta$*.*
(c) $\mathcal{P}$ *is continuously* $\mathbb{L}_1$*-differentiable on* Θ*, and the derivatives* $\dot{L}_\vartheta$ *admit a representation* $\dot{L}_\vartheta = \mathcal{A}(\vartheta)^\top T - b(\vartheta)$ *for all* $\vartheta \in \Theta$*, for some continuous mappings* $\mathcal{A} : \Theta \to \mathrm{GL}(k)$ *and* $b : \Theta \to \mathbb{R}^k$*.*

Proof. Fix $\vartheta_0, \vartheta \in \Theta$ and choose a continuously differentiable path ϑ_s, $0 \leq s \leq 1$, from ϑ_0 to ϑ, its derivative is denoted by $\dot{\vartheta}_s$. For $x \in \mathcal{X}$ define

$$\begin{aligned} g_{\mathcal{A}}(s) &= \mathcal{A}(\vartheta_s)\dot{\vartheta}_s, & \alpha(\vartheta) &= \textstyle\int_0^1 g_{\mathcal{A}}(s)\,ds, \\ g_b(s) &= \dot{\vartheta}_s^\top b(\vartheta_s), & \kappa(\vartheta) &= \textstyle\int_0^1 g_b(s)\,ds, \\ f(x) &= \exp\bigl(\textstyle\int_0^1 \dot{\vartheta}_s^\top \dot{L}_{\vartheta_s}(x)\,ds\bigr) = \exp\bigl(\alpha(\vartheta)^\top T(x) - \kappa(\vartheta)\bigr). \end{aligned}$$

Due to the continuity assumptions these quantities are well defined, and f is measurable.

We claim that f is a P_{ϑ_0}-density of P_ϑ. Then neither f nor $\kappa(\vartheta)$ will depend on the path ϑ_s that enters into the definition, and the same will be true for $\alpha(\vartheta)$ since the distributions of T do not concentrate on a proper affine subspace. In order to establish our claim we must verify

$$\int_B f\,dP_{\vartheta_0} = P_\vartheta(B) \tag{8}$$

for all $B \in \mathcal{B}$. But for every $\epsilon > 0$ there exists a partitioning of $\mathbb{R}^k$ into measurable rectangles $R_1, R_2, \ldots$ of diameter less than ϵ. For a fixed set B define $B_i = B \cap T^{-1}(R_i)$. If B_i is a $\mathcal{P}$-nullset then (8) holds for B_i.

Now consider the case that B_i is not a $\mathcal{P}$-nullset. Since $\vartheta \to P_\vartheta(B_i)$ is continuously differentiable, see Witting (1985, Satz 1.179), so is the function $H(B_i, s) = \log P_{\vartheta_s}(B_i)$ and $H(B_i, \cdot)$ is the integral of its derivative $h(B_i, \cdot)$. Thus we have

$$\exp\Bigl(\int_0^1 h(B_i, s)\,ds\Bigr) = \frac{P_\vartheta(B_i)}{P_{\vartheta_0}(B_i)}, \qquad h(B_i, s) = \dot{\vartheta}_s^\top m(B_i, s) - g_b(s),$$

where the integral $m(B_i, s) = \int_{B_i} \mathcal{A}(\vartheta_s)^\top T\,dP_{\vartheta_s}/P_{\vartheta_s}(B_i)$ exists in view of $\mathcal{A}(\vartheta_s)^\top T = \dot{L}_{\vartheta_s} + b(\vartheta_s)$. With this notation we obtain

$$\int_{B_i} f\,dP_{\vartheta_0} = \int_{B_i} \exp\left(\int_0^1 \dot{\vartheta}_s^\top \bigl(\mathcal{A}(\vartheta_s)^\top T - m(B_i, s)\bigr)\,ds\right) dP_{\vartheta_0}\,\frac{P_\vartheta(B_i)}{P_{\vartheta_0}(B_i)}.$$

For $x \in B_i$ the point $\mathcal{A}(\vartheta_s)^\top T(x)$ lies in the image $\mathcal{A}(\vartheta_s)^\top(R_i)$, and—being an average—so does $m(B_i, s)$. Therefore they have maximal distance $\|\mathcal{A}(\vartheta_s)\|\epsilon \leq \Lambda\epsilon$, say, with Λ being the maximum for $0 \leq s \leq 1$ of the operator norm of $\mathcal{A}(\vartheta_s)$. Hence with $c = \Lambda \int_0^1 |\dot{\vartheta}_s|\, ds$ the inner integral is bounded by $\pm c\epsilon$. Summation over i gives

$$e^{-c\epsilon} P_\vartheta(B) \leq \int_B f\, dP_{\vartheta_0} \leq e^{c\epsilon} P_\vartheta(B).$$

Since ϵ is arbitrary our claim is established. Hence $\mathcal{P}$ is an exponential family.

Fixing ϑ_0 and varying ϑ defines α on all of Θ, with $\alpha(\vartheta_0) = 0$. Next we show that α is differentiable at ϑ_0. For a point ϑ close to ϑ_0 we may choose a straight-line path $\vartheta_s = \vartheta_0 + s(\vartheta - \vartheta_0)$, whence $\alpha(\vartheta) = \int_0^1 \mathcal{A}(\vartheta_s)(\vartheta - \vartheta_0)\, ds$. Then

$$\frac{|\alpha(\vartheta) - \alpha(\vartheta_0) - \mathcal{A}(\vartheta_0)(\vartheta - \vartheta_0)|}{|\vartheta - \vartheta_0|} \leq \int_0^1 \frac{\left|\left(\mathcal{A}(\vartheta_s) - \mathcal{A}(\vartheta_0)\right)(\vartheta - \vartheta_0)\right|}{|\vartheta - \vartheta_0|}\, ds.$$

Since this tends to 0 as ϑ tends to ϑ_0 we have that α is differentiable at ϑ_0, with Jacobian matrix $\mathcal{A}(\vartheta_0)$.

Fixing ϑ and varying ϑ_0 we similarly know that for $\vartheta_1 \neq \vartheta_0$ the distribution P_ϑ has a P_{ϑ_1}-density proportional to $\exp\bigl(\alpha_1(\vartheta)^\top T\bigr)$ where α_1 is differentiable at ϑ_1 and has Jacobian matrix $\mathcal{A}(\vartheta_1)$. The chain rule $dP_\vartheta/dP_{\vartheta_0} = (dP_\vartheta/dP_{\vartheta_1})(dP_{\vartheta_1}/dP_{\vartheta_0})$ leads to $\alpha(\vartheta) = \alpha(\vartheta_1) + \alpha_1(\vartheta)$. Hence since α_1 is differentiable at ϑ_1 so is α, and their common Jacobian matrix is $\mathcal{A}(\vartheta_1)$. Thus α is differentiable on Θ and has nonsingular Jacobian matrices $\mathcal{A}(\vartheta)$.

This also implies that α is an open mapping, whence the image $\alpha(\Theta)$ is an open subset of the canonical parameter domain of the exponential family $\mathcal{P}$. Since T is not concentrated on a proper affine subspace the family $\mathcal{P}$ must then be of order k. $\Diamond$

It is worth mentioning that Theorem 2 provides a particular instance where $\mathbb{L}_1$-differentiability entails $\mathbb{L}_r$-differentiability for $r \geq 1$. The converse is true quite generally, namely that $\mathbb{L}_r$-differentiability implies $\mathbb{L}_s$-differentiability for $s \leq r$, see Witting (1985, Satz 1.190).

References

Barankin, E.W. (1949). Locally best unbiased estimates. *The Annals of Mathematical Statistics* **20** 477–501.

Fabian, V. and Hannan, J. (1977). On the Cramér-Rao inequality. *The Annals of Statistics* **5** 197–205.

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

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

Joshi, V.M. (1976). Ona the attainment of the Cramér-Rao lower bound. *The Annals of Statistics* **4** 998–1002.

Le Cam, L. (1966). Likelihood functions for large numbers of independent observations. pp. 167–187 in: *Research Papers in Statistics. Festschrift for J. Neyman.* F.N. David, Ed. John Wiley & Sons, London.

Pitman, E.J.G. (1979). *Some Basic Theory for Statistical Inference.* Chapman and Hall, London.

Wijsman, R.A. (1973). On the attainment of the Cramér-Rao lower bound. *The Annals of Statistics* **1** 538–542.

Witting, H. (1985). *Mathematische Statistik I.* Teubner-Verlag, Stuttgart.

17. September 1987