Z. Wahrscheinlichkeitstheorie verw. Gebiete
65, 121–134 (1983)

Zeitschrift für
Wahrscheinlichkeitstheorie
und verwandte Gebiete

Factorizing the Information Contained in an Experiment, Conditionally on the Observed Value of a Statistic*

W. Ehm and D.W. Müller

University of Frankfurt, Fachbereich Mathematik, Robert Mayer Str. 10, D-6000 Frankfurt a.M., Federal Republic of Germany

0. Summary. No matter which value t of a statistic T_n has been observed the loss of information, in comparison with the original data, will asymptotically (as $n \to \infty$) always be the same: this statement is interpreted and proved in the framework of "comparison of experiments", under assumptions commonly accepted in asymptotic statistics. The loss of information is described by the conditional experiments $\{\mathscr{L}_\theta\,(\text{data} \mid T_n = t)\colon \theta \in \Theta\}$. These are shown to be all of the same "type", as $n \to \infty$.

1. Introduction and Exposition of Results

This paper extends results which have been used to compare statistics S_n, T_n with respect to the operating characteristics of related tests, estimation procedures etc. In the case where the statistics under consideration are already estimators the most useful idea has been to compare their asymptotic variances. This requires regularity conditions to guarantee the existence of a normal approximation; even then the situation is much obscured by the occasional appearance of "superefficiency". The rather attractive idea which excludes this nuisance has been developed by Hájek (1970), Inagaki (1970): it consists in showing that quite generally the limiting distributions of "regular" sequences of estimators can be represented as convolutions $\mathscr{N}(0, \Gamma) * H$ where Γ is the inverse Fisher information matrix and H is some probability measure. Intuitively, the spread of H reflects the loss of information in comparison with asymptotically optimal estimators. As has been shown by LeCam (1972), the existence of such representations follows easily in the common situation where the distributions of the estimators asymptotically form translation families: if T_n is asymptotically less informative than S_n in the sense of LeCam (1964) its distribution family can be obtained approximately from that of S_n by a randomization; for translation families the randomizing kernel can be chosen to be

* Research supported by the Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 123)

invariant, and thus as a convolution kernel. This approach is both very powerful - it requires few assumptions, and very general - it is not restricted to the case of asymptotic normality. On the other hand, it does not relate the concrete form of the convolution kernel to decision-theoretic concepts.

The approach of the present paper goes back to the idea behind the classical definition of sufficiency, which, in the language of statistical decision theory, can be formulated thus: the conditional experiments given the value of a sufficient statistic do not contain any information (i.e. they consist of identical measures). In principle, this is the idea of a two-stage experiment: first one observes $T_n = t$; secondly, given $T_n = t$, one samples from the conditional distributions $\mathscr{L}_\theta$ (data$|T_n = t$). Similar two-stage decompositions make sense also in certain cases when T_n is far from being sufficient. The conditional experiments corresponding to the second stage then embody the information lost by using T_n. The purpose of this paper is to show that under assumptions commonly accepted in asymptotic statistics, the limiting forms of the conditional experiments are, with high probability, all of the same "type", i.e. contain the same residual information. The assumptions essentially say that the basic experiments can be approximated in norm by linearly indexed exponential families of distributions, as proposed by LeCam (1960); in addition the family of distributions of T_n is assumed to behave asymptotically as a linearly indexed exponential family. Thus, no use is made of asymptotic normality, and the structure behind the Hájek-Inagaki theorem is revealed.

In more detail, the contents of this paper are as follows.

We consider experiments (that is, parametrized families of probability measures) on measurable spaces $(\mathscr{X}_n, \mathscr{A}_n)$ which are such that all conditional distributions appearing below exist. This can be achieved by assumptions about $\mathscr{A}_n$ - for instance by choosing $\mathscr{X}_n$ to be a Polish space (or, more generally, a Borel subset of some compact metric space), and $\mathscr{A}_n$ the family of its Borel subsets. It could also be attained by using a generalized concept of "conditional distribution"; in the sequel, however, we refrain from such abstractions, but rather refer to the Polish case. The sequence of experiments under study will be denoted by $\mathscr{E}_n = (\mathscr{X}_n, \mathscr{A}_n; P_{\theta,n}: \theta \in \Theta)$ or, more simply, $\mathscr{E}_n = (P_{\theta,n}: \theta \in \Theta)$ where $P_{\theta,n}$ are probability measures on $\mathscr{A}_n$ indexed by elements θ of the parameter set Θ, which is an open convex subset of $\mathbb{R}^k$ containing the origin. The statistics of interest are denoted by T_n; they are measurable maps from $\mathscr{X}_n$ to some Polish space $\mathscr{T}$ (which will usually be Euclidean). In standard applications the experiments $\mathscr{E}_n$ consist of n-fold product measures with the parameter θ and T_n already renormalized. There is no loss of generality if one replaces $\mathscr{E}_n$ by the equivalent experiment $(\mathscr{X}_n, \mathscr{A}_n; P_{\theta,n} \times T_n: \theta \in \Theta)$ consisting of the joint distributions of the data X_n and the statistics T_n. This replacement will be assumed henceforth, even if not explicitly mentioned. By a disintegration one can represent $P_{\theta,n} \times T_n$ as $B^{\cdot}_{\theta,n} \times T_n P_{\theta,n}$, with Markov transition kernels $B^{\cdot}_{\theta,n}$ from $\mathscr{T}$ to $\mathscr{X}_n$, which are simply the conditional distributions (under $P_{\theta,n}$) of X_n given T_n. Thus $\mathscr{E}_n$ is decomposed into a two-stage experiment: at first, $\mathscr{G}_n = (T_n P_{\theta,n}: \theta \in \Theta)$ is performed and the value t of T_n is observed; then one samples from the conditional distribution $B^t_{\theta,n}$ of X_n given $T_n = t$; i.e. one performs the experiment $\mathscr{B}^t_n = (B^t_{\theta,n}: \theta \in \Theta)$. Although this decomposition appears somewhat artificial, the conditional experiments are an appropriate device for describing the residual information not yet exhausted by T_n.

The experiments $\mathscr{G}_n$ and $\mathscr{B}_n^t$ can be treated neatly when $\mathscr{E}_n$ is a linearly indexed exponential family, i.e. consists of distributions $Q_{\theta,n}$ such that

$$Q_{\theta,n}=\exp(\theta' S_n - A_n(\theta))\cdot Q_{0,n} \tag{1}$$

where S_n is a $\mathbb{R}^k$-valued (sufficient) statistic and $A_n(\theta)$ a normalizing constant. Therefore, in the present paper, more general situations $\mathscr{E}_n=(\mathscr{X}_n, \mathscr{A}_n;\ P_{\theta,n}\colon\ \theta\in\Theta)$ will be treated through approximation by exponential families

$$(\mathscr{X}_n, \mathscr{A}_n;\ Q_{\theta,n}\colon\ \theta\in\Theta). \tag{2}$$

For these, given the null-distribution of S_n, the joint distribution of S_n, T_n is completely determined by the family (in θ) of marginal distributions of T_n; this is because we have assumed

$$\Theta \text{ is an open convex subset of } \mathbb{R}^k \text{ containing the origin.} \tag{A0}$$

This circumstance is responsible for the asymptotic uniqueness of the above-mentioned disintegrations, which is indispensible for our approach. Thus experiments which can be approximated by exponential families form the natural framework for this study. The idea of such an approximation goes back to LeCam (1960). He has shown that under rather general conditions one can approximate $P_{\theta,n}$ by a linearly indexed exponential family (1) in the sense that (A1) (A2) (A3) hold:

$$\|P_{\theta,n}-Q_{\theta,n}\|\to 0 \qquad (\theta\in\Theta;\ n\to\infty); \tag{A1}$$

$$S_n P_{0,n}\to F_0 \ \text{ weakly}, \tag{A2}$$

where F_0 is some probability distribution on $\mathbb{R}^k$. Of course, in (A2), $P_{0,n}$ can be replaced by $Q_{0,n}$ without invalidating this statement. Since $\exp(A_n(\theta))$ is the Laplace transform of S_n, one concludes that $A_n(\theta)$ converges to some function $A(\theta)$ equal to the log Laplace transform of F_0-provided the latter exists; this we postulate:

$$\int \exp(\theta' s)\, F_0(ds)=\exp(A(\theta)) \qquad (\theta\in\Theta). \tag{A3}$$

Conditions (A0) (A1) (A2) (A3) *are assumed throughout this paper.* It is important, and easy to see, that these conditions imply contiguity for the measures $P_{\theta,n}$ as well as for $Q_{\theta,n}$.

As already mentioned, we do not describe "information" by a numerical quantity but rather as the "type" of an experiment. For comparing experiments we shall use the deficiency Δ, which was introduced by LeCam (1964). It is a pseudo-distance between experiments $\mathscr{M}_i=(\mathscr{Y}_i, \mathscr{B}_i;\ M_{\theta,i}\colon\ \theta\in\Theta)$ which, in the dominated case, can be defined as follows (for the general case see LeCam (1964)). Let

$$\delta(\mathscr{M}_1, \mathscr{M}_2)=\inf_L \sup_{\theta\in\Theta} \|L M_{\theta,1}-M_{\theta,2}\| \tag{3}$$

where the inf is taken over all Markov transition kernels L from $\mathscr{Y}_1$ to $\mathscr{Y}_2$ (these spaces are assumed to be Polish and endowed with the sigma-algebras $\mathscr{B}_i$ of Borel sets). From δ one obtains the deficiency Δ by symmetrizing:

$$\Delta(\mathscr{M}_1, \mathscr{M}_2) = \delta(\mathscr{M}_1, \mathscr{M}_2) \vee \delta(\mathscr{M}_2, \mathscr{M}_1). \tag{4}$$

Experiments $\mathscr{M}_1, \mathscr{M}_2$ with $\Delta(\mathscr{M}_1, \mathscr{M}_2)=0$ are called "equivalent"; the equivalence class of an experiment is referred to as its "type". Frequently, the sup in (3) will be restricted to subsets D of Θ; the deficiency distance will then be called Δ_D. Convergence of sequences of experiments in Δ will usually be established by first proving convergence in Δ_D, for all finite subsets D of Θ, and then approximating. The first step is facilitated by considering the Hellinger[1] transform of experiments:

$$\eta(\alpha; \mathscr{M}, D) = \int \prod_{\theta \in D} (dM_\theta)^{\alpha_\theta}, \qquad (\alpha_\theta \geqq 0 \text{ for } \theta \in D, \sum_{\theta \in D} \alpha_\theta = 1), \tag{5}$$

where D is a finite subset of Θ. Pointwise convergence of η is equivalent to convergence of the respective experiments in Δ_D. In case of a linearly indexed exponential family (1) the Hellinger transform takes on the particularly simple form (6):

$$\exp\left(A_n\left(\sum_{\theta \in D} \alpha_\theta \theta\right) - \sum_{\theta \in D} \alpha_\theta A_n(\theta)\right). \tag{6}$$

This applies in particular to the conditional distributions of X_n given $T_n = t$ (under $Q_{\theta,n}$), versions of which can be selected in essentially one way so as to form an exponential family.

Let $\mathscr{F}$ be the exponential family consisting of the distributions $F_\theta = \exp(\theta' s - A(\theta)) \cdot F_0$. According to LeCam's "third lemma" these are the weak limits of $F_{\theta,n} = S_n P_{\theta,n}$ (forming the experiment $\mathscr{F}_n$). By the above arguments one easily shows that $\Delta_D(\mathscr{E}_n, \mathscr{F}) \to 0$ $(n \to \infty)$ for all finite subsets D of Θ; it even follows that $\Delta_K(\mathscr{E}_n, \mathscr{F}) \to 0$ $(n \to \infty)$ for all compact subsets K of Θ if condition (A1) holds uniformly on compacts.

In our more general framework described by the assumptions (A0)...(A3) one obtains the following theorem which has the classical result of Hájek-Inagaki as a corollary. Here $L^1(F_0)$ denotes the band of measures absolutely continuous with respect to F_0.

Theorem 1. *Suppose the marginal distributions $G_{\theta,n} = T_n P_{\theta,n}$ converge weakly, say to G_θ $(\theta \in \Theta)$. Then the joint distributions $\mathscr{L}(S_n, T_n | P_{\theta,n})$ converge weakly. The limit distributions $\mathscr{L}(S, T | \theta)$ are of the form $F_\theta \times R$ where R is a Markov kernel from $\mathbb{R}^k$ to $\mathscr{T}$ which does not depend on the parameter θ. The kernel R, as an operator on $L^1(F_0)$, and hence the joint distributions of S and T, are uniquely determined by the relations $G_\theta = R F_\theta$ $(\theta \in \Theta)$.*

[1] Ernst Hellinger was professor of mathematics at the university of Frankfurt from 1914 through 1935

In the terminology of LeCam (1972) this means that the S_n are "distinguished" statistics; moreover the theorem asserts the uniqueness of the randomizing kernel.

When applied to situations where, for some positive definite symmetric linear operator Γ on $\mathbb{R}^k$, the distributions $F'_\theta := \Gamma F_\theta$ and G_θ both form translation families theorem 1 reproduces the classical result of Hájek-Inagaki. (Such situations occur in estimation problems under the assumptions of asymptotic normality of S_n and "regularity" of T_n). In this case the transformed kernel $R' := R\Gamma^{-1}$ satisfies $(R'F'_\tau)*\delta_\theta = R'(F'_\tau * \delta_\theta)$ $(\theta, \tau \in \Theta)$, and R' can be selected in such a way that it commutes with shifts; it therefore reduces to a convolution kernel: $R'F' \equiv F' * H$ for some probability measure H. Despite its general formulation this argument is confined to the case of asymptotic normality within the present framework. For the family $(F'_\theta: \theta \in \Theta)$ is automatically Gaussian since it is a linearly indexed exponential family and translation-invariant at the same time. More precisely, we have $F'_\theta = \mathcal{N}(\theta, \Gamma)$ as follows from the definition of F'_θ. Thus one arrives at the familiar form of the convolution theorem $G_\theta = \mathcal{N}(\theta, \Gamma) * H$.

Let $\mathscr{G}$ be the experiment $(G_\theta: \theta \in \Theta)$; being less informative than $\mathscr{F}$ it consists of mutually absolutely continuous distributions. The densities dG_θ/dG_0 will be denoted by g_θ. Clearly, $\mathscr{L}(S, T|\theta)$ has the correct marginals: $\mathscr{L}(S|\theta) = F_\theta$, $\mathscr{L}(T|\theta) = G_\theta$. Theorem 1 holds in particular if $P_{\theta,n}$ is replaced throughout by $Q_{\theta,n}$, the limit remaining unchanged. Let $\mathscr{H}_n$ be the experiment consisting of $H_{\theta,n} = T_n Q_{\theta,n}$. The joint distribution of X_n, T_n under $Q_{\theta,n}$ will be disintegrated into $C^{\bullet}_{\theta,n} \times H_{\theta,n}$ with Markov kernels $C^{\bullet}_{\theta,n}$ from $\mathscr{T}$ to $\mathscr{X}_n$ representing the conditional distributions of X_n given T_n, under $Q_{\theta,n}$. Given $t \in \mathscr{T}$, let $\mathscr{C}^t_n$ denote the conditional experiment $(C^t_{\theta,n}: \theta \in \Theta)$, which, as an exponential family, is well-defined with $H_{0,n}$-probability one. In the limit, a similar disintegration leads to

$$\mathscr{L}(S, T|\theta) = C^{\bullet}_\theta \times G_\theta \qquad (\theta \in \Theta); \tag{7}$$

the corresponding experiments $(C^t_\theta: \theta \in \Theta)$ will be denoted by $\mathscr{C}^t$.

The two main results can now be stated. The first one (Theorem 2), briefly, asserts that, once the total information converges, the information contained in a statistic converges if and only if the residual information converges.

Theorem 2. *Assume that condition* (A1) *holds uniformly on compact subsets of* Θ. *Let* $G_{\theta,n}$ *converge weakly to* G_θ *with continuous densities* g_θ $(\theta \in \Theta)$. *Then the following statements* (i) (ii) (iii) *are all equivalent.*

(i) $\Delta_K(\mathscr{G}_n, \mathscr{G}) \to 0$ *for every compact subset* K *of* Θ;

(ii) $\sup_{\theta \in K} E_{\theta,n} \Delta_D(\mathscr{B}^{T_n}_n, \mathscr{C}^{T_n}) \to 0$ (D *finite,* K *compact subset of* Θ);

(iii) $\sup_{\theta \in K} E_{\theta,n} \Delta_{K'}(\mathscr{C}^{T_n}_n, \mathscr{C}^{T_n}) \to 0$ (K *and* K' *compact subsets of* Θ).

The remarkable feature here is that conclusions about conditional, and hence joint, behavior of S_n and T_n can be drawn from a statement about marginal distributions only. This is related to the phenomenon discussed above, following (2). The unpleasant necessity of distinguishing between $\mathscr{B}^{\bullet}_n$ and

$\mathscr{C}_n^{\bullet}$ cannot be avoided when dealing with uniformity in θ. The reason is that all that is known about their difference comes from the assertion that

$$\int \|B_{\theta,n}^{T_n} - C_{\theta,n}^{T_n}\| \, dP_{0,n} \to 0. \tag{8}$$

This obviously cannot exclude the possibility of a locally ill-behaved experiment $\theta \rightsquigarrow B_{\theta,n}^{T_n}$, no matter how regular $\theta \rightsquigarrow C_{\theta,n}^{T_n}$ may be.

An important special case, when the limit experiment $\mathscr{G}$ is a linearly indexed exponential family, is treated by Theorem 3, our second main result. It shows that, no matter which value t of T_n has been observed, the residual information is always the same: it can be described by the type of a single experiment $\mathscr{C}$ replacing the bundle $\mathscr{C}^{\bullet}$ of Theorem 2. Convergence in distribution of T_n is not required here; hence, in this situation, the symbol $\mathscr{G}$ does not refer to the above weak limit.

Theorem 3. *Assume that condition* (A1) *holds uniformly on compact subsets of* Θ. *Then the following statements* (i) (ii) (iii) *are all equivalent.*

(i) *There is a linearly indexed exponential family* $\mathscr{G}$ *such that* $\Delta_K(\mathscr{G}_n, \mathscr{G}) \to 0$, *for every compact subset* K of Θ;

(ii) *there is an experiment* $\mathscr{C} = (C_\theta\colon \theta \in \Theta)$ *such that* $\sup_{\theta \in K} E_{\theta,n} \Delta_D(\mathscr{B}_n^{T_n}, \mathscr{C}) \to 0$ (D *finite*, K *compact subset of* Θ);

(iii) *there is an experiment* $\mathscr{C} = (C_\theta\colon \theta \in \Theta)$ *such that* $\sup_{\theta \in K} E_{\theta,n} \Delta_{K'}(\mathscr{C}_n^{T_n}, \mathscr{C}) \to 0$ (K *and* K' *compact subsets of* Θ).

The type of the experiment $\mathscr{C}$, a linearly indexed exponential family, is uniquely determined by $\mathscr{F}$ and $\mathscr{G}$ through the expression

$$\Delta(\mathscr{F}, \mathscr{C} \otimes \mathscr{G}) = 0; \tag{9}$$

this means that $\mathscr{F}$ can be factored into a product of the limit experiment of T_n and an experiment $\mathscr{C}$ describing the information deficit. Although $\mathscr{C}$ is not unique as an experiment, it can be selected in a natural way; with this choice of $\mathscr{C}$, the cases admitting a factorization (9) can be concisely characterized, moreover the relationship between $\mathscr{C}$ and the bundle $\mathscr{C}^{\bullet}$ of Theorem 2 will become apparent.

Proceeding to the construction of $\mathscr{C}$, let us start with an experiment $\mathscr{G}$, assuming only that it is less informative than $\mathscr{F}$, i.e. $\delta(\mathscr{F}, \mathscr{G}) = 0$. The Markov kernel R transforming F_θ into $G_\theta = RF_\theta$ is unique in the sense that any kernel with this property will define the same family of joint distributions $\mu_\theta := F_\theta \times R \equiv \mathscr{L}(S, T \mid \theta)$ on $\mathbb{R}^k \times \mathscr{T}$ (compare Theorem 1). Having now a representation of $\mathscr{F}$ and $\mathscr{G}$ by joint distributions, as in Theorem 2, one can characterize the situtation of Theorem 3 as follows:

$\mathscr{G}$ is a linearly indexed exponential family if and only if for some function ψ *the statistics* $S - \psi(T)$ *and* T *are independent under* μ_0 *(equivalently: under* μ_θ*). In this case* $\psi(T)$ *differs from* $E_0(S \mid T)$ *by at most an additive constant.*

Our choice of $\mathscr{C} = (C_\theta\colon \theta \in \Theta)$ will be

$$C_\theta = \mathscr{L}(S - \psi(T) \mid \theta). \tag{10}$$

The conditional experiments $\mathscr{C}^t$ appearing in Theorem 2 can then be obtained from $\mathscr{C}$ by simply adding a constant $\psi(t)$ to each observation (thus preserving the type of $\mathscr{C}$):

$$C_\theta^t = \mathscr{L}_\theta(S \mid T=t) = \mathscr{L}(S-\psi(T) \mid \theta) * \delta_{\psi(t)} = C_\theta * \delta_{\psi(t)}. \tag{11}$$

To prove the "only if"-part of the above criterion one writes down two different expressions for the densities $dG_\theta/dG_0(t)$. One is the assumed exponential family form of the densities $\exp(\theta'\sigma(t)-B(\theta))$ where σ is an $\mathbb{R}^k$-valued function on $\mathscr{T}$ and $B(\theta)$ are norming quantities; the other one is the marginal density representation $E_0(d\mu_\theta/d\mu_0 \mid T=t) = E_0(\exp(\theta' S - A(\theta)) \mid T=t)$. Hence

$$E_0(\exp\theta'(S-\sigma(T)) \mid T) = \exp(A(\theta)-B(\theta)), \tag{12}$$

and the independence of $S-\sigma(T)$ and T follows. It remains to show that $E_0(S \mid T)-\sigma(T)$ is constant, which can be done by differentiating (12) with respect to θ at $\theta=0$.

The "if"-part of the criterion can be proved by reversing this argument.

In standard situations, the criterion reduces to familiar independence relations. As an example, consider independent variables U_i having Gamma densities $u^{\theta_i-1}\exp(-u)/\Gamma(\theta_i)$ respectively $(i=1,2)$. Let $S := (S_1, S_2)' \equiv (\log U_1, \log U_2)'$ and $T := U_1+U_2$. The distributions of T under $\theta=(\theta_1,\theta_2)'$ form an exponential family with densities proportional to $\exp(\theta'\sigma(T))$ where $\sigma(t)=(\log t, \log t)'$. Here our criterion asserts the well-known independence of $S-\sigma(T) = (\log U_1/(U_1+U_2), \log U_2/(U_1+U_2))'$ and $T=U_1+U_2$. Hence in this situation our experiment $\mathscr{G}$ corresponds to the Gamma family with parameter $\theta_1+\theta_2$, and our $\mathscr{C}$ is equivalent to the Beta family with parameter (θ_1,θ_2).

The situation becomes particularly simple in the case of asymptotic normality: any nondegenerate factorization of a Gaussian experiment $\mathscr{F} = (\mathscr{N}(\theta,\Gamma): \theta\in\Theta)$ (Γ nonsingular) consists of $(\mathscr{N}(\theta,\Gamma_1): \theta\in\Theta)$ and $(\mathscr{N}(\theta,\Gamma_2): \theta\in\Theta)$, up to equivalence, where $\Gamma_1^{-1}+\Gamma_2^{-1}=\Gamma^{-1}$. This is because the canonical sufficient statistics $S-E_0(S \mid T)$ and $E_0(S \mid T)$ must be normally distributed as independent variables whose sum is Gaussian.

Theorem 3 thus covers, in particular, the important case of asymptotic normality and statistics T_n which are approximately linear functions of S_n. It is noteworthy, however, that there are situations occurring frequently in practice which admit no such simple answer. For instance, when T_n is of quadratic type the experiment may consist of a noncentral chi-square-family, which is not an exponential family.

2. Proof of Theorem 1

The first step consists in showing that the sequence of distributions $\mathscr{L}(S_n, T_n \mid Q_{0,n})$ converges weakly. Since this sequence is tight (note that its marginal distributions converge by assumption), it suffices to show that it has at most one limit point. Let $\mu_0 = \mathscr{L}(S, T \mid 0)$ be a limit point. Then, for any bounded continuous function u, one gets

$$\int u\, dG_\theta = E_0 u(T)\exp(\theta' S - A(\theta)) \qquad (\theta\in\Theta) \tag{13}$$

where the expectation on the right is taken under μ_0. In deriving (13) contiguity is used. The parameter θ being arbitrary this determines μ_0 uniquely.

In the second step one obtains the existence and the particular form of the limits $\mathscr{L}(S, T|\theta)$ by LeCam's "third lemma" and by sufficiency. By LeCam's lemma, for every θ there will be convergence of $\mathscr{L}(S_n, T_n|Q_{\theta,n})$, say to $\mathscr{L}(S, T|\theta)$; moreover, (14) holds:

$$\frac{d\mathscr{L}(S, T|\theta)}{d\mathscr{L}(S, T|0)} = \frac{d\mathscr{L}(S|\theta)}{d\mathscr{L}(S|0)}. \tag{14}$$

This means that S is sufficient for the family $\mathscr{L}(S, T|\theta)$, implying the existence of a universal conditional distribution R of T given S. Hence the desired disintegration holds. The uniqueness of R follows from the Laplace transform argument used in the first step.

3. Auxiliary Technical Results

In our context exponential families prove useful because of the special form of their densities and Hellinger transforms, and because of their regular analytic behavior. These properties are inherited by the conditional experiments $\mathscr{C}^{\cdot}$, provided these have been chosen properly. One possible construction of $\mathscr{C}^{\cdot}$ is described below.

First note that $H_{0,n}$-almost surely one has

$$h_{\theta,n}(t) \equiv \frac{dH_{\theta,n}}{dH_{0,n}}(t) = E_{0,n}(\exp(\theta' S_n - A_n(\theta))|T_n = t) \tag{15}$$

where $E_{0,n}$ means expectation under $Q_{0,n}$. Up to a set of $H_{0,n}$-probability zero the right hand side of (15) can be represented as

$$\int \exp(\theta' S_n - A_n(\theta))\, dC_{0,n}^t \tag{16}$$

with a Markov kernel $C_{0,n}^{\cdot}$ from $\mathscr{T}$ to $\mathscr{X}_n$ satisfying $C_{0,n}^{\cdot} \times H_{0,n} = \mathscr{L}(X_n, T_n|Q_{0,n})$. In the sequel we always select the version of $h_{\theta,n}(t)$ given by (16). For each t, the experiment $(C_{\theta,n}^t\colon \theta \in \Theta)$, up to technicalities, ought to be an exponential family; therefore, to avoid measure-theoretic difficulties, we start by *defining*

$$dC_{\theta,n}^t = \frac{\exp(\theta' S_n)\, dC_{0,n}^t}{\int \exp(\theta' S_n)\, dC_{0,n}^t}, \tag{17}$$

which, after (16), is nothing but

$$\frac{\exp(\theta' S_n - A_n(\theta))}{h_{\theta,n}(t)}\, dC_{0,n}^t. \tag{18}$$

This definition, however, makes sense only for those t for which $h_{\theta,n}(t) < +\infty$ *simultaneously* for all $\theta \in \Theta$. But this happens with $H_{0,n}$-probability one, due to the convexity of the log Laplace transform. Hence we take (18) as the definition of $C_{\theta,n}^t$ for those t which satisfy $h_{\theta,n}(t) < +\infty$ $(\theta \in \Theta)$, whereas, for the remaining t, the measure $C^t_{\theta,n}$ may be defined as the point mass at zero. This definition makes $\mathscr{C}^t_n=(C^t_{\theta,n}\colon \theta\in\Theta)$ an exponential family ($t\in\mathscr{T}$) whose members are versions of the conditional distributions in question. Moreover, $\mathscr{C}^t_n$ is essentially unique in the sense that $\tilde{C}^t_{\theta,n}=C^t_{\theta,n}$ ($\theta\in\Theta$) for $H_{0,n}$-almost all t whenever $(\tilde{C}^t_{\theta,n})$ is an exponential family with $\tilde{C}^{\bullet}_{\theta,n}\times H_{\theta,n}=\mathscr{L}(X_n, T_n|Q_{\theta,n})$ ($\theta\in\Theta$). In this sense the conditional experiments $\mathscr{C}^t_n$ are uniquely defined, with $H_{0,n}$-probability one. The same applies to the limiting experiments $\mathscr{C}^t$. Clearly, no technical problems arise when the parameter set is finite, as is the case in the statements (ii) of Theorem 2 and 3. Thus the original conditional experiments $\mathscr{B}^t_n$ can be used there.

The Hellinger transforms $\eta(\alpha;\mathscr{C}^t_n, D)$ have an interesting form relating them to the densities $h_{\theta,n}$:

$$\eta(\alpha;\mathscr{C}^t_n, D)=\frac{h_{\sum\limits_{\theta\in D}\alpha_\theta\theta,\,n}(t)}{\prod\limits_{\theta\in D} h_{\theta,n}(t)^{\alpha_\theta}}\,\eta(\alpha;\mathscr{F}_n, D); \tag{19}$$

(the same applies to the limiting experiment $\mathscr{C}^t$). This can be seen by considering (6) in connection with the Laplace transform of $\mathscr{L}(S_n|C^t_{0,n})$:

$$\lambda^t_n(\theta)\equiv\int \exp(\theta' S_n)\,dC^t_{0,n}=h_{\theta,n}(t)\exp(A_n(\theta)), \tag{20}$$

which, in view of (18), is related to the Hellinger transform as follows:

$$\eta(\alpha;\mathscr{C}^t_n, D)=\lambda^t_n\Big(\sum_{\theta\in D}\alpha_\theta\theta\Big)\Big/\prod_{\theta\in D}\lambda^t_n(\theta)^{\alpha_\theta}. \tag{21}$$

Since Laplace transforms converge uniformly on compacts whenever they converge pointwise, the assumptions (A1) (A2) (A3) will imply that

$$A_n(\theta)\to A(\theta)\quad\text{uniformly on compact subsets of }\Theta. \tag{22}$$

The asymptotic behavior of the conditional experiments therefore depends solely on the densities $h_{\theta,n}$.

In the proofs of the main theorems one establishes: (a) the convergence in probability of the transforms $\eta(\alpha;\mathscr{C}^{T_n}_n, D)$ uniformly in α; and (b) the uniform (norm-) precompactness in probability of the families $(C^{T_n}_{\theta,n}\colon \theta\in K)$, K compact. To this end two auxiliary results are needed which mainly exploit the properties of Laplace transforms.

Lemma 1. *Let $\varepsilon>0$ and D a finite subset of Θ. Then there exist numbers n_0 and b depending on D and ε such that*

$$Q_{0,n}\Big[\sup_{\alpha\neq\alpha'}|\alpha-\alpha'|^{-1}\,|\eta(\alpha;\mathscr{C}^{T_n}_n, D)-\eta(\alpha';\mathscr{C}^{T_n}_n, D)|\leqq b\Big]\geqq 1-\varepsilon$$

as soon as $n\geqq n_0$.

The *proof* relies on a well-known property of families of analytic functions: uniform boundedness implies a uniform Lipschitz condition. The function θ

$\leadsto \lambda_n^t(\theta)$, as a Laplace transform, admits a unique analytic continuation to $\Theta + i\mathbb{R}^k$. By (20), for every finite subset D of Θ and every $\theta^* \in \operatorname{conv}(D)$, the convex hull of D, one obtains

$$|\lambda_n^t(\theta^* + iv)| \leqq \lambda_n^t(\theta^*) \leqq \max_{\theta \in D} \lambda_n^t(\theta) \qquad (v \in \mathbb{R}^k); \tag{23}$$

here, the second inequality is a consequence of the convexity of the log Laplace transform. Another application of (20) then gives the boundedness in $Q_{0,n}$-probability of the sequence $\lambda_n^{T_n}(\theta)$ since by contiguity $h_{\theta,n}(T_n)$ is bounded in probability and $A_n(\theta) \to A(\theta)$ $(\theta \in D)$. One concludes that there is a number b such that for all n

$$Q_{0,n}[\sup_{\theta_1, \theta_2 \in \operatorname{conv}(D)} |\theta_1 - \theta_2|^{-1} |\lambda_n^{T_n}(\theta_1) - \lambda_n^{T_n}(\theta_2)| \leqq b] \geqq 1 - \varepsilon. \tag{24}$$

In view of (21) this implies an analogous bound for the Hellinger transforms if one can show that $\lambda_n^{T_n}(\theta)$, with high probability, is bounded away from zero and infinity $(\theta \in D)$. Boundedness from above has already been shown; boundedness from below follows by a standard contiguity argument.

Similar arguments will yield the uniform precompactness, stated as Lemma 2.

Lemma 2. *Let* $\varepsilon > 0$ *and* K *a compact subset of* Θ. *Then there exist numbers* n_0 *and* b *depending on* K *and* ε *such that*

$$Q_{0,n}[\sup_{\substack{\theta_1, \theta_2 \in K \\ \theta_1 \neq \theta_2}} |\theta_1 - \theta_2|^{-1/2} \| C_{\theta_1,n}^{T_n} - C_{\theta_2,n}^{T_n} \| \leqq b] \geqq 1 - \varepsilon$$

as soon as $n \geqq n_0$.

The *proof* uses a well-known inequality relating the total-variation distance to the Hellinger distance:

$$\| C_{\theta_1,n}^t - C_{\theta_2,n}^t \|^2 \leqq 4(1 - \int \sqrt{d C_{\theta_1,n}^t \, d C_{\theta_2,n}^t}). \tag{25}$$

By (17) and (20) the right-hand side of (25) is equal to

$$4 \left(1 - \lambda_n^t \left(\frac{\theta_1 + \theta_2}{2}\right) / \sqrt{\lambda_n^t(\theta_1) \, \lambda_n^t(\theta_2)}\right). \tag{26}$$

In view of (24) it suffices to show that $\lambda_n^{T_n}(\theta)$ is bounded in probability away from zero and infinity, uniformly in $\theta \in K$. Uniform boundedness from above has already been established in the proof of Lemma 1 for the sufficiently large class of compacts $\operatorname{conv}(D)$, D finite. Every family of distributions having locally uniformly bounded Laplace transforms is tight. Similarly $\mathscr{L}(S_n | C_{0,n}^{T_n})$ is "tight in probability", in the strong sense that it has a universal compact epsilon-support; more precisely, for every $\varepsilon > 0$ there is a compact subset L of $\mathbb{R}^k$ such that

$$Q_{0,n}[C_{0,n}^{T_n}\{S_n \in L\} \geqq 1 - \varepsilon] \geqq 1 - \varepsilon. \tag{27}$$

This in turn implies that $\inf_{\theta \in K} \lambda_n^{T_n}(\theta)$ is bounded away from zero, in probability.

4. Proof of Theorem 2

"(i) implies (ii)"

As a first step, pointwise convergence of the Hellinger transforms $\eta(\cdot;\mathscr{C}_n^{T_n}, D)$ in $Q_{0,n}$-probability will be shown. As in (19), in the limit, the Hellinger transforms are of the form

$$\eta(\alpha;\mathscr{C}^t, D)=\frac{g_{\sum\limits_D \alpha_\theta\theta}(t)}{\prod\limits_D g_\theta(t)^{\alpha_\theta}}\,\eta(\alpha;\mathscr{F}, D), \tag{28}$$

G_0-almost surely. In view of (19) it therefore suffices to prove that

$$\frac{h_{\sum\limits_D \alpha_\theta\theta, n}}{\prod\limits_D h_{\theta,n}^{\alpha_\theta}} \to \frac{g_{\sum\limits_D \alpha_\theta\theta}}{\prod\limits_D g_\theta^{\alpha_\theta}} \tag{29}$$

in $H_{0,n}$-probability. Let $D^*=D\cup\{\theta^*\}$, with $\theta^*=\sum\limits_D \alpha_\theta\theta$. If one denotes the likelihood ratios $h_{\tau,n}/\sum\limits_{D^*} h_{\theta,n}$ and $g_\tau/\sum\limits_{D^*} g_\theta$ by $r_{\tau,n}$ and r_τ respectively, (29) can be replaced by the equivalent assertion (30) that

$$\prod_{\theta\in D} (r_{\theta^*,n}/r_{\theta,n})^{\alpha_\theta} - \prod_{\theta\in D} (r_{\theta^*}/r_\theta)^{\alpha_\theta} \to 0 \tag{30}$$

in $H_{0,n}$-probability, or – by contiguity – in $H_{\tau,n}$-probability ($\tau\in D^*$). Since, by contiguity, the ratios $r_{\tau,n}$ are bounded away from zero and one (with high probability), for proving (30) it suffices to show that

$$r_{\theta,n}\to r_\theta \qquad \text{in } H_{\tau,n}\text{-probability } (\theta,\tau\in D^*). \tag{31}$$

For this, the assumed weak convergence $H_{\theta,n}\to G_\theta$ is not enough, since it only implies convergence of (31) in distribution. If, however, in addition to weak convergence one has convergence in Δ of the experiments ($H_{\theta,n}$: $\theta\in D^*$) to (G_θ: $\theta\in D^*$) and the densities g_θ are continuous, (31) will indeed follow. This is a general proposition implicitly contained in LeCam (1979; Thm. 1, Sect. 3, Chap. 7). Since these conditions are fulfilled in our case, (31) and thus the pointwise convergence of the Hellinger transforms are established.

In a second step, (iii) will be derived by approximation arguments using Lemma 1 and 2. By Lemma 1 we get uniform convergence of the Hellinger transforms, in $H_{0,n}$-probability. This immediately implies that

$$\Delta_D(\mathscr{C}_n^{T_n}, \mathscr{C}^{T_n})\to 0 \qquad \text{in } Q_{0,n}\text{-probability, for finite } D. \tag{32}$$

By Lemma 2 this can be extended to

$$\Delta_K(\mathscr{C}_n^{T_n}, \mathscr{C}^{T_n})\to 0 \qquad \text{in } Q_{0,n}\text{-probability, for compact } K. \tag{33}$$

Finally, this convergence takes place also in $P_{\theta,n}$-probability uniformly on compact subsets of Θ, by the assumed uniform validity of (A1) and by the uniform precompactness of $(Q_{\theta,n}\colon \theta\in\text{compact})$.

"*(iii) implies (i)*"

As in the preceding part of the proof it suffices to show pointwise convergence of the Hellinger transforms $\eta(\cdot\,;\mathscr{H}_n, D)$ to $\eta(\cdot\,;\mathscr{G}, D)$ for every finite subset D of Θ. These can be written in the form

$$\eta(\alpha;\mathscr{H}_n, D)=\int \prod_D h_{\theta,n}^{\alpha_\theta}/h_{\theta^*,n}\, dH_{\theta^*,n}, \tag{34}$$

$$\eta(\alpha;\mathscr{G}, D)=\int \prod_D g_\theta^{\alpha_\theta}/g_{\theta^*}\, dG_{\theta^*} \tag{35}$$

respectively ($\theta^*=\sum_D \alpha_\theta \theta$). The integrands turn out to be the reciprocals of the quantities occurring in (29), the difference of which tends to zero in $H_{\theta^*,n}$-probability, by (iii). By contiguity, the same holds for the integrands of (34) (35). Convergence of the integrals (34) (35) then follows from uniform integrability of the integrands which is a consequence of the inequality

$$\prod_D h_{\theta,n}^{\alpha_\theta}/h_{\theta^*,n} \leqq \max_D h_{\theta,n}/h_{\theta^*,n} \tag{36}$$

and of contiguity.

"*(ii) and (iii) are equivalent*"

This is a consequence of a standard inequality (see, for instance, LeCam (1974)) relating the norm difference between probability measures to the corresponding norm differences of their conditional distributions. When applied to $B^{\cdot}_{\theta,n}\times G_{\theta,n} = P_{\theta,n}\times T_n$ and $C^{\cdot}_{\theta,n}\times H_{\theta,n}=Q_{\theta,n}\times T_n$, this inequality reads as follows:

$$\begin{aligned}\int \|B^t_{\theta,n}-C^t_{\theta,n}\|\,(G_{\theta,n}(dt)+H_{\theta,n}(dt))\\ \leqq 4\,\|P_{\theta,n}\times T_n - Q_{\theta,n}\times T_n\|.\end{aligned} \tag{37}$$

Since the right-hand side of (37) is nothing but $4\,\|P_{\theta,n}-Q_{\theta,n}\|$, this yields the desired result up to an approximation argument based on Lemma 2.

5. Proof of Theorem 3

"*(i) implies (iii)*"

First we show that the (random) Hellinger transforms of the conditional experiments converge weakly in distribution to a constant random variable, for fixed argument α; from this convergence in probability will follow, for each α

separately. This degeneracy is decisive for the appearance of a limit experiment $\mathscr{C}$ not depending on t, as asserted by (iii).

Let $\mathscr{G}$ consist of measures G_θ of the form

$$G_\theta = \exp(\theta' U - B(\theta)) \cdot G_0 \qquad (\theta \in \Theta), \tag{38}$$

with some $\mathbb{R}^k$-valued statistic U. Let D be a finite subset of Θ and fix α; denote $\sum_D \alpha_\theta \theta \in \operatorname{conv}(D)$ by θ^*. By (i) and (A1) one has $\Delta_D(\mathscr{H}_n, \mathscr{G}) \to 0$, which can be equivalently expressed by the convergence in distribution of the vectors of likelihood ratios:

$$\mathscr{L}((h_{\theta,n}/h_{\theta^*,n})_{\theta \in D} | H_{\theta^*,n}) \to \mathscr{L}((g_\theta/g_{\theta^*})_{\theta \in D} | G_{\theta^*}). \tag{39}$$

We now consider

$$(\xi_\theta)_{\theta \in D} \rightsquigarrow \Big(\prod_{\theta \in D} \xi_\theta^{\alpha_\theta}\Big)^{-1} \tag{40}$$

as a functional on the vector of likelihood ratios. By contiguity, this is continuous in almost all points, with respect to the limit distribution of (39). Therefore, its distributions converge weakly:

$$\mathscr{L}\Big(h_{\theta^*,n}/\prod_D h_{\theta,n}^{\alpha_\theta} \Big| H_{\theta^*,n}\Big) \to \mathscr{L}\Big(g_{\theta^*}/\prod_D g_\theta^{\alpha_\theta} \Big| G_{\theta^*}\Big). \tag{41}$$

The experiment $\mathscr{G}$ being an exponential family of the form (38) the right-hand side of (41) turns out to be a point measure concentrated at

$$\exp\Big(\sum_{\theta \in D} \alpha_\theta B(\theta) - B(\theta^*)\Big), \tag{42}$$

the reciprocal of the Hellinger transform of $\mathscr{G}$ (see (6)). Hence, by (19),

$$\eta(\alpha; \mathscr{C}_n^{T_n}, D) \to \eta(\alpha; \mathscr{F}, D)/\eta(\alpha; \mathscr{G}, D) \tag{43}$$

in $Q_{\theta^*,n}$-probability, this convergence being in fact uniform in α according to Lemma 1. The limit will therefore be the Hellinger transform of some experiment $\mathscr{C}_D$. Taking the limit $D \uparrow \Theta$ (compare LeCam (1972; p. 251), LeCam (1979; Thm. 2, Chap. 3)) one finds an experiment $\mathscr{C} = (C_\theta\colon \theta \in \Theta)$ having Hellinger transform (43; right-hand side), for any finite subset D of Θ:

$$\eta(\alpha; \mathscr{C}, D) = \eta(\alpha; \mathscr{F}, D)/\eta(\alpha; \mathscr{G}, D). \tag{44}$$

The remaining arguments concerning uniformity and approximation are as above and will be omitted.

"*(iii) implies (i)*"

As a Δ-limit of a sequence of exponential families whose canonical sufficient statistics S_n have Laplace transforms that stay locally bounded, the experiment $\mathscr{C}$ will itself be an exponential family. Following the corresponding part of the proof of Theorem 2 one shows that

$$\eta(\alpha; \mathscr{H}_n, D) = \int \prod_D h_{\theta,n}^{\alpha_\theta}/h_{\theta^*,n} dH_{\theta^*,n} \to \eta(\alpha; \mathscr{F}, D)/\eta(\alpha; \mathscr{C}, D). \tag{45}$$

Therefore there is an experiment $\mathscr{G} = (G_\theta: \theta \in \Theta)$ such that

$$\eta(\alpha; \mathscr{G}, D) = \eta(\alpha; \mathscr{F}, D)/\eta(\alpha; \mathscr{C}, D) \tag{46}$$

(D finite subset of Θ); from this the remark (9) in the introduction follows immediately. By uniform precompactness of $\mathscr{H}_n$ we get $\Delta_K(\mathscr{G}_n, \mathscr{G}) \to 0$ (K compact subset of Θ).

It remains to conclude from (9) that $\mathscr{G}$ is an exponential family. By contiguity, $\mathscr{G}$ is a homogeneous experiment in the sense that all G_θ are mutually absolutely continuous. A homogeneous experiment $\mathscr{G} = (G_\theta: \theta \in \Theta)$ is an exponential family of rank less than or equal to k if and only if the supports of the joint distributions of the log likelihood ratios $(\log dG_\theta/dG_0)_{\theta \in D}$ under G_0 are at most k dimensional (D finite subset of Θ). This is because the process of log likelihood ratios $(\log dG_\theta/dG_0)_{\theta \in \Theta}$ can be represented as a linear, and hence continuous, image of some of its finite dimensional marginal vectors when this condition is satisfied (see LeCam (1974; p. 91)). By this criterion, since $\mathscr{C}$ and $\mathscr{C} \otimes \mathscr{G}$ are exponential families, the same will be true for $\mathscr{G}$.

Acknowledgment. Thanks are due to L. LeCam for careful reading and helpful comments.

References

Hájek, J.: A characterization of limiting distributions of regular estimates. Z. Wahrscheinlichkeitstheorie verw. Gebiete **14**, 323–330 (1970)

Inagaki, N.: On the limiting distribution of a sequence of estimators with uniformity property. Ann. Inst. Statist. Math. **22**, 1–13 (1970)

LeCam, L.: Locally asymptotically normal families of distributions. Univ. of California Publications in Statistics, Univ. of California Press, Berkeley and Los Angeles, Vol. **3**, p. 37–98 (1960)

LeCam, L.: Sufficiency and approximate sufficiency. Ann. Math. Statist. **35**, 1419–1455 (1964)

LeCam, L.: Limits of experiments. Proc. Sixth Berkeley Sympos. Math. Statist. Probab., Univ. of California Press, Berkeley and Los Angeles, Vol. **1**, p. 245–261 (1972)

LeCam, L.: Notes on asymptotic methods in statistical decision theory. Centre de Recherches Mathématiques, Université de Montréal, Canada (Publication CRM-245) (1974)

LeCam, L.: unpublished book manuscript (1979)

Received April 30, 1982; in final form June 3, 1983