On the standard asymptotic confidence ellipsoids of Wald

By L. Le Cam

University of California, Berkeley NSF Grant DMS-8701426

1. Introduction. In his famous paper of 1943 Wald proved asymptotic optimality properties for a variety of tests of simple or composite hypotheses. Wald considers families $\{P_{\theta,n}; \theta \in \Theta\}$ of probability measures indexed by a subset Θ of a Euclidean space. The tests are derived from a recipe that involves estimates T_n of the parameter θ and estimates Γ_n of the inverse covariance matrix of T_n. One forms a chi-square type statistic $(T_n - \theta)' \Gamma_n (T_n - \theta)$ and reject those θ's for which the statistic is too large. For T_n Wald uses the maximum likelihood estimate $\hat{\theta}_n$. For Γ_n he uses the Fisher information matrix J_θ evaluated at the estimate $\hat{\theta}_n$ of θ.

This gives a readily applicable way of constructing tests and confidence ellipsoids.

It has been noted by several authors that Wald's procedure can suffer from some unsatisfactory features. One defect, noted by Hauck and Donner (1977) is that, for fixed θ, a criterion of the type $(\theta - t)' J_t (\theta - t)$ can decrease as $|\theta - t|$ becomes large. This is also noted by Vaeth (1985) who points out in addition that the results of the procedure are not invariant under smooth one to one transformations of the parameter space. Vaeth gives examples of one-dimensional exponential families where, $\hat{\theta}_n$ being the maximum likelihood estimate, $(\theta - \hat{\theta}_n)' J_{\hat{\theta}_n} (\theta - \hat{\theta}_n)$ tends to zero for all fixed θ as $\hat{\theta}_n$ approaches the

boundary of its possible range. An example of this, imitated from Vaeth, will be described in Section 4.

Wald was dealing with a situation where the measures $P_{\theta,n}$ were product measures, distributions of n independent identically distributed observations. He made on these distributions a number of relatively severe assumptions. The consequences of Wald's assumptions will be reviewed below in Section 3. They suggest that instead of Wald's quadratic expressions one could use tests based on the expression $q_n^2(s,t) = -8 \log \int \{dP_{s,n}\, dP_{t,n}\}^{1/2}$. One would use confidence sets of the type $\{\theta : q_n^2(T_n, \theta) \le c_n(\theta)\}$ for suitably selected estimates T_n.

A proposal to use such sets was made by K. Matusita in 1955. This was mostly for multinomial situations. The proposal was later extended to some problems involving heteroschedastic Gaussian families of measures. See Matusita (1967).

We show that the use of the function q_n^2 does indeed mitigate some of the difficulties encountered by Wald's procedure. Unfortunately the procedure has also defects of its own. Briefly, q_n^2 is a monotone function of the Hellinger distance h defined by $h_n^2(s,t) = \frac{1}{2} \int [(dP_{s,n})^{1/2} - (dP_{t,n})^{1/2}]^2$. As such it remains invariant under all one-to-one transformations of the parameter space. It is, of course, somewhat more difficult to compute than Wald's criterion. However, leaving this aside, its main inconveniences are as follows:

1) To compute $q_n^2(T_n, \theta)$ the value of T_n must lie in the parameter space, or the parameter space must be extended to cover the possible range of T_n.

2) q_n^2 is a monotone function of a distance. As such $q_n^2(s,t)$ is a symmetric function of s and t. There are many situations, for example the standard binomial, where the use of sets of the type $\{\theta : q_n^2(T_n, \theta) \le c_n\}$ with c_n independent

of θ cannot capture all the relevant features of the problem.

One could wonder why bother with such a proposal? Why not just use the likelihood ratio method of Neyman and Pearson (1928)? The reason is that, as far as we know, the local asymptotic justification for the likelihood ratio are all based on the article of Wilks (1938) or on the 1943 paper of Wald. They proceed by showing that, under severe conditions, the likelihood ratio method is asymptotically equivalent to Wald's procedure. For a proof of asymptotic optimality for separated hypotheses see Bahadur (1967). Besides, there is no lack of examples where the likelihood ratio method suffers from major difficulties. See for instance Lehmann (1986) page 342 or Le Cam (1979).

Otherwise, the paper is organized as follows. Section 2 reviews definitions and notation. Section 3 gives an account of consequences that can be derived from Wald's assumptions. It does not review the assumptions themselves, only consequences.

Section 4 gives details of an example analogous to one considered by Vaeth (1985). The behavior of a criterion based on our q_n^2 appears satisfactory.

Section 5 is suggested by the heteroschedastic Gaussian approximations that occur naturally in the framework used by Wald. It shows that variations on the definitions of chi-square type criteria can lead to very different answers.

Section 6 touches upon a number of different matters: The effect of the lack of uniformity in the local convergence to Gaussian shift experiments, the need to use estimates that take values outside the parameter sets Θ_n and some possibilities for the extension of the domain of definition of q_n to cover such eventualities.

Section 7 is an aside on covariance stabilizing transformations. An appendix gives a derivation of the formula for $q^2(s,t)$ in the heteroschedastic Gaussian case.

2. Gaussian experiments and distances.

In this section we recall a few facts about Gaussian experiments and approximations by them. The facts are well known but presented here in manner that emphasizes the role of chi-square type expressions and of the function q^2 defined in the Introduction.

Let Θ be an arbitrary set. An experiment $\mathbf{G} = \{G_\theta : \theta \in \Theta\}$ is called Gaussian shift, or simply Gaussian if no confusion ensues, if it satisfies the following two conditions:

1) The measures G_θ; $\theta \in \Theta$ are mutually absolutely continuous,

2) Let $\Lambda(t, s) = \log dG_t/dG_s$. Then the stochastic process $t \rightsquigarrow \Lambda(t, s)$, $t \in \Theta$ is a Gaussian process for the distribution induced by the measure G_s. (The choice of point s does not matter).

Note that the definition does not refer to any algebraic or vector property of Θ. The set Θ was not assumed to have any such structure. However, if one is given a Gaussian shift experiment $\mathbf{G}$ indexed by Θ one is automatically given a map of Θ into a Hilbert space. To see this, consider the process $X(t) = \Lambda(t, s) - E_s \Lambda(t, s)$ under G_s. Let $\mathbf{M}_0(\Theta)$ be the space of all finite signed measures μ with finite support on Θ that are such that $\mu(\Theta) = 0$. Let $\|\mu\|^2$ be the variance of the random variable $\int X(t)\mu(dt)$. The norm or pseudo-norm so defined on $\mathbf{M}_0(\Theta)$ is Hilbertian or pre-Hilbertian. One can identify to zero those μ such that $\|\mu\| = 0$ and then complete to get a Hilbert space. One maps Θ into it by associating to θ the difference $\delta_\theta - \delta_s$ of the Dirac masses carried by θ and s. The Gaussian shift experiment $\{G_\theta : \theta \in \Theta\}$ can be extended to all of $\mathbf{M}_0(\Theta)$ by taking

$$dG_\mu = \exp\{\int X(t)\mu(dt) - \frac{1}{2}\|\mu\|^2\} dG_s .$$

From there it extends to the Hilbert space by continuity.

The square distance between θ on t becomes

$$\| \delta_\theta - \delta_t \|^2 = E_s | X(\theta) - X(t) |^2$$

$$= -8 \log \int [dG_s \, dG_t]^{1/2} .$$

If the measures G_θ had been given by a standard normal density with respect to Lebesgue measure having the form

$$\frac{|\det \Gamma|^{1/2}}{(2\pi)^{k/2}} \exp \{ -\frac{1}{2} (x - \theta)' \Gamma (x - \theta) \}$$

with Γ fixed, independent of θ, one would have $\| \delta_\theta - \delta_t \|^2 = (\theta - t)' \Gamma (\theta - t)$.

Note that the formulas written previously did not use any link between the linear structure of Θ, if any, and the distances or log likelihood ratios. The above formula, with $(x - \theta)$ in the exponent or $(\theta - t)' \Gamma (\theta - t)$ in the square distance assumes such a link. For asymptotic problems where one may want to transform the parameter space a choice of appropriate linkage, or of transformation, may be important.

There are also heteroschedastic Gaussian experiments. In the standard Euclidean case they are given by densities.

$$\frac{|\det \Gamma(\theta)|^{1/2}}{(2\pi)^{k/2}} \exp \{ -\frac{1}{2} (x - \theta)' \Gamma(\theta) (x - \theta) \}$$

where now Γ is a function of the parameter θ.

For such heteroschedastic Gaussian experiments our function q^2 takes a different form. Let P be a Gaussian distribution with center θ and covariance matrix Γ^{-1} on $\mathbb{R}^k$. Let Q be another Gaussian measure with center t and covariance matrix K^{-1} on the same $\mathbb{R}^k$. The value of $q^2 = -8 \log \int [dP dQ]^{1/2}$ is easily seen to be

$$-2 \log \det [\, I - (M^{-1} \Delta)^2 \,] + (t - \theta)' [\, M - \Delta M^{-1} \Delta \,] (t - \theta)$$

where I is the identity matrix and $M = \frac{1}{2}(\Gamma + K)$ while $\Delta = \frac{1}{2}(\Gamma - K)$. See Kraft (1955) or Matusita (1967). Note that the above formula consist of a sum of two terms, one that involves the difference between the inverse of the covariance matrices and one that has a structure of the same type as for Gaussian shift experiments, involving the differences between expectations in a quadratic form.

As we shall see below, the theory originated by Wald in 1943 relies on local approximations by Gaussian shift experiments but global approximation by heteroschedastic Gaussian experiments. To define such approximations properly one can use a distance defined by Le Cam (1964).

Let $\mathbf{E} = \{P_\theta : \theta \in \Theta\}$ and $\mathbf{F} = \{Q_\theta : \theta \in \Theta\}$ be two experiments indexed by the same set Θ. The distance $\Delta(\mathbf{E}, \mathbf{F})$ introduced by Le Cam (1964) is essentially as follows: Except for technicalities, to say that $\Delta(\mathbf{E}, \mathbf{F}) \leq \varepsilon$ is to say that, as long as one uses only loss functions W bounded by zero and unity, any risk function available on one of the experiments can be matched within ε by a risk function available on the other experiment.

Thus to say that a sequence $\{\mathbf{E}_n\}$ of experiments $\mathbf{E}_n = \{P_{\theta,n} : \theta \in V_n\}$ is asymptotically Gaussian shift is to say that there are Gaussian shift experiments $\mathbf{G}_n = \{G_{\theta,n} : \theta \in V_n\}$ such that $\Delta(\mathbf{E}_n, \mathbf{G}_n)$ tends to zero as n tends to infinity.

Note that we have called the index set V_n instead of Θ. This is because this kind of approximation is usually possible only in small neighborhoods V_n that shrink as $n \to \infty$.

For a theory of such approximations, see Strasser (1985), Le Cam (1985) and (1986). The next section will elaborate on such approximations in a context derived from Wald's assumptions.

3. Some consequences of Wald's assumptions.

Wald's paper of 1943 refers to a situation where the parameter space Θ is a fixed subset of a Euclidean space $\mathbb{R}^k$ and where the measures $P_{\theta,n}$ are the joint distributions of n independent identically distributed observations. We shall not recall these assumptions here but will look at some of the conclusions derived from them.

Some of Wald's conclusions are global, valid uniformly on Θ. Some are ''local'', valid only in sets that are small enough and that shrink as $n \to \infty$. In the sequel, we shall use the word ''local'' in the following manner. Let q_n^2 be the function defined by $q_n^2(s,t) = -8 \log \int [\, dP_{s,n}\, dP_{t,n}\,]^{1/2}$ for s and t in Θ. A ''local'' property is one that is valid on certain specified sets of the form

$$V_n(\tau_n, b) \;=\; \{\theta : q_n^2(\theta, \tau_n) \le b\}$$

for specified sequences $\{\tau_n\}$ and for arbitrarily fixed values of b.

Although Wald's assertions are for maximum likelihood estimates in the i.i.d. case, it is better to forget about such restrictions and use only conclusions that may hold more generally.

Wald considers estimates T_n with values in $\mathbb{R}^k$ and nonrandom matrices $M_{\tau,n}$; $\tau \in \Theta$ with the following properties.

(A) Let $\{\tau_n\}$ be an arbitrary sequence with $\tau_n \in \Theta$. If $q_n(\theta_n, \tau_n)$ stays bounded then the distributions $\mathbf{L}\{M_{\tau_n,n}(T_n - \theta_n) \,|\, \theta_n\}$ tend to the standard k-dimensional normal $\mathbf{N}(0, I)$

(B) The T_n are asymptotically sufficient in the following sense: There are other families of probability measures $\{Q_{\theta,n} : \theta \in \Theta\}$ defined on the same σ-fields as the $P_{\theta,n}$ and such that:

(i) For $\{Q_{\theta,n}: \theta \in \Theta\}$ the statistics T_n are sufficient (exactly).

(ii) $\sup\{\| P_{\theta,n} - Q_{\theta,n} \|; \theta \in \Theta\}$ tends to zero as n tends to infinity, the norm being the total variation norm.

(C) Let $F_{\theta,n}$ be the distribution of T_n under $P_{\theta,n}$. There are Gaussian distributions $G_{\theta,n}$ centered at θ and Markov kernels K_n' and K_n'' such that $\sup_\theta \{\| F_{\theta,n} - K_n' G_{\theta,n} \|; \theta \in \Theta\}$ and $\sup_\theta \{\| G_{\theta,n} - K_n'' F_{\theta,n} \|; \theta \in \Theta\}$ tend to zero as $n \to \infty$. Here again $\| \cdot \|$ is the total variation norm.

Another assertion is that the Markov kernels K_n' and K_n'' represent small distributions in the following sense: Take an arbitrary sequence $\{\tau_n\}$, $\tau_n \in \Theta$ and a fixed $b < \infty$. Let $\Gamma_{\theta,n}$ be the inverse covariance matrix of $G_{\theta,n}$, assumed to exist. Let $B_n(t, a)$ be the ball $B_n(t,a) = \{x: (x - t)' \Gamma_{\tau_n,n} (x - t) < a\}$.

(D) For every sequence $\{\tau_n\}$ $b < \infty$ and every $\varepsilon > 0$

$$\sup_t \{K_n' [B_n^c(t, \varepsilon) | t]: t \in B_n(\tau_n, b)\}$$

tends to zero as $n \to \infty$. The same condition holds for K_n''. Another condition is as follows

(E) The T_n take their values in Θ

Finally, for ease of reference, we shall also use a condition (F) as follows

(F) For any given $\varepsilon > 0$ there are finite numbers $b = b(\varepsilon)$ and $N = N(\varepsilon)$ such that $n \geq N$ implies

$$\sup_{\theta \in \Theta} P_{\theta,n} \{q_n^2(T_n, \theta) > b\} < \varepsilon.$$

It will be seen below that (F) is already a consequence of (A) to (E).

It should be mentioned that the above (A) to (E) are not exactly Wald's statements. The wording is closer to that of Le Cam (1956). The reasons for this will appear below. The roles of (B) and (C) are simple. The condition (C)

says that the distance between the experiments $\mathbf{F}_n = \{F_{\theta,n} : \theta \in \Theta\}$ and $\mathbf{G}_n = \{G_{\theta,n} : \theta \in \Theta\}$ tends to zero. The sufficiency relation (B) implies in addition that the distance between $\mathbf{F}_n$ and $\mathbf{E}_n = \{P_{\theta,n} : \theta \in \Theta\}$ tends to zero.

The condition (A) is an asymptotic normality condition of the usual type for convergence of distributions. Condition (D) is meant to allow a link between the inverse covariance matrices $\Gamma_{\theta,n}$ of the $G_{\theta,n}$ and the matrices $M_{\theta,n}' M_{\theta,n}$ of (A). Then (A) will imply that, locally, the heteroschedastic Gaussian $\mathbf{G}_n$ can be approximated by Gaussian shift experiments.

Wald does not use the function q_n at all. Instead, he assumes the existence of an underlying Euclidean norm that has special properties and can be used to define what is ‘‘local’’. He uses arguments that are very close to the assertions (C) and (D) about Markov kernels. Instead of our sufficiency condition (B) Wald uses the existence of set transformations. He assigns to each measurable set in the observation space a set in the range of T_n that has, uniformly in θ, almost the same probability. This seems asking for too much. One can readily have a sufficient sub-field $\mathbf{B}$ of a σ-field $\mathbf{A}$ where the conditional expectation of an indicator I_A, $A \in G$ is some fairly arbitrary $\mathbf{B}$-measurable function ϕ bounded by zero and unity. The possibility of replacing such a function by the indicator I_B of a set $B \in \mathbf{B}$ so that $\int |I_B - \phi| \, dP_{\theta,n} < \varepsilon$ uniformly in θ depends on properties of the sufficient statistic that are not readily visible.

Consider for instance i.i.d. observations $X_1, X_2, \ldots, X_n$ from a univariate estimate is $T_n = \overline{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j$. Consider also the sum $S_n = \sum_{j=1}^{n} (X_j - \overline{X})^2$. It is independent of $\overline{X}_n$ with a chi-square distribution. Take a set A_n of the form $A_n = \{X_1, X_2, \ldots, X_n ; S_n \le c_n\}$ where c_n is selected so that $\Pr[S_n \le c_n] = 1/2$. Its conditional expectation ϕ is identically $1/2$. One can find sets $B_n \subset \mathbb{R}$ such that $|P_{\theta,n}[\overline{X}_n \in B_n] - 1/2| < \varepsilon$ for all θ and n but they do not

have a very appealing appearance and to construct them one needs to use fairly fine knowledge of the distributions of $\overline{X}_n$.

Our condition (B) is simpler. It seems preferable. Similar considerations apply to our conditions (C) and (D) and the set transformations of Wald's Lemma 2.

The reader should note that the conjunction of the properties (A) to (E) is very restrictive indeed. This is partly due to the insistence that all convergences occurs uniformly on the entire set Θ. This is reflected directly in conditions (B) and (C). It is implied, for instance in condition (A), by the use of sequences $\{\tau_n\}$ that are entirely arbitrary.

Another aspect of the restrictive nature of the conditions arises from the innocuous looking condition (E). It is not stated as such in Wald's paper but is implied by the fact that Wald takes for T_n the maximum likelihood estimate for the family $\{P_{\theta,n} : \theta \in \Theta\}$. It is very natural in the context of the present paper if we want to replace Wald's quadratic expressions by $q_n^2(T_n, \theta)$. That means in particular that $q_n^2(s, \theta)$ must be defined for $\theta \in \Theta$ and s in the range of T_n. (In a first draft of the present paper, we had not paid enough attention to that requirement. It was pointed out by Yu Lin Chang who deserves my thanks).

Now call boundary point t of Θ an ''ordinary boundary point'' if the contingent of Θ at t is not the entire space $\mathbb{R}^k$. This means simply that there is a unit vector u such that u is not the limit of any sequence $(s_n - t)/|s - t|$ with $s_n \in \Theta$ tending to t.

The combination of (A) and (E) implies that Θ cannot have any ordinary boundary points. There are such sets Θ different from $\mathbb{R}^k$. An example would be the set of all points with rational coordinates. Another example can be constructed as follows. Let $W(x)$, $x \in (-\infty, +\infty)$ be a standard Wiener process with $W(0) = 0$. Let Θ be the set $\{\theta = (x,y);\ y \le W(x)\}$. Then Θ has almost surely

no ordinary boundary points. However usually occuring parameter sets, if they are not all of $\mathbb{R}^k$, often have many ordinary boundary points.

On this particular point Wald's paper seems to contain a gap. He fails to specify what kind of a set Θ might be. He proceeds as if the maximum likelihood estimates were roots of the maximum likelihood equations except for probabilities that tend to zero uniformly in θ as $n \to \infty$. This cannot be at ordinary boundary points.

Some of these problems are avoided in Le Cam (1956) by two devices. The first is to allow ''estimates'' T_n that can take values out of Θ. The other is to relax the uniformity requirements for the convergence properties. Both devices create other problems as we shall see. If one takes $\Theta = \mathbb{R}^k$ the conditions (A) to (E) can be satisfied in some cases. For instance take joint distributions $P_{\theta,n}$ of n i.i.d. observations from a density $f(x, \theta)$ with respect to some dominating measure μ. If μ is Lebesgue measure on $\mathbb{R}^k$ and $f(x, \theta) = f(x - \theta)$ and if the Fisher information matrix exists, then $P_{\theta,n}$ will satisfy (A) to (E).

For more general families, with densities $f(x, \theta)$, one could assume differentiability in quadratic mean of $[f(x, \theta)]^{1/2}$ uniformly for $\theta \in \Theta$. This and conditions of boundedness and non degenaracy of the covariance matrix of the derivative in quadratic mean will ensure the validity of (A) to (E) for suitable estimates T_n, though not necessarily for the maximum likelihood.

Even though examples do exist, uniform convergence on all of $\mathbb{R}^k$ is a large order. Some palliatives for cases where the convergences are not uniform or Θ is not all of $\mathbb{R}^k$ will be discussed in Section 6.

For the time being let us return to conditions (A) to (E) and their implications for the relations between q_n^2 and Wald's quadratics. A first easy result is as follows.

Proposition 1. *Let conditions (B) and (C) be satisfied. For the heteroschedastic experiment* $\mathbf{G}_n$ *of condition (C), let* $g_n^2(s,t) = -8\log\int [\, dG_{s,n}\, dG_{t,n}\,]^{1/2}$. *Then for every* $\varepsilon > 0$ *there is an* $N(\varepsilon) < \infty$ *such that* $n \geq N(\varepsilon)$ *implies either*

$$|\, q_n^2(s,t) - g_n^2(s,t)\,| \;<\; \varepsilon$$

or

$$\min [\, q_n^2(s,t), g_n^2(s,t)\,] \;> 1/\varepsilon$$

for all pairs (s,t) *of elements of* Θ.

Proof. This is an immediate consequence of the fact that the distance between $\mathbf{E}_n = \{P_{\theta,n}\,;\, \theta \in \Theta\}$ and $\mathbf{G}_n = \{G_{\theta,n}\,;\, \theta \in \Theta\}$ tends to zero. It implies that the difference

$$\int [\, dP_{s,n}\, dP_{t,n}\,]^{1/2} \;-\; \int [\, dG_{s,n}\, dG_{t,n}\,]^{1/2}$$

tends to zero uniformly in (s,t). The rest follows by passage to logarithms.

Proposition 2. *Let the conditions (A) to (D) be satisfied and let* $K_{\tau,n} = M_{\tau,n}'\, M_{\tau,n}$. *For arbitrary* $\{\tau_n\}$, $\tau_n \in \Theta$ *take sequences* $\{s_n\}$ *and* $\{t_n\}$ *such that* $q_n^2(s_n, \tau_n)$ *and* $q_n^2(t_n, \tau_n)$ *stay bounded. Then the differences*

$$q_n^2(s_n, t_n) - (s_n - t_n)'\, K_{\tau_n,n}\, (s_n - t_n)$$

tend to zero as $n \to \infty$.

Proof. Let the triplets (s_n, t_n, τ_n) be as stated. Consider the binary experiments

$$\mathbf{B}_n = \{P_{s_n,n}, P_{t_n,n}\}, \;\; \mathbf{B}_n' = \{F_{s_n,n}, F_{t_n,n}\} \;\text{ and }\; \mathbf{B}_n'' = \{G_{s_n,n}, G_{t_n,n}\}.$$

We claim that the distances between these three experiments tend to zero. For the pair $(\mathbf{B}_n, \mathbf{B}_n')$ this follows from condition (B). For the pair $(\mathbf{B}_n', \mathbf{B}_n'')$ this is exactly the statement of condition (C) restricted to the pairs (s_n, t_n).

Now introduce the half-space distance $|F - F'|_h$ between two measures on $\mathbb{R}^k$ by

$$|F - F'|_h = \sup_H |F(H) - F'(H)|$$

where the supremum is taken over all the half spaces H of $\mathbb{R}^k$. This distance is invariant under all affine transformations of $\mathbb{R}^k$.

One can readily check that the convergence in (A) if taken according to any of the usual definitions for weak convergence of measures, such as Lévy distance, Prokhorov distance, or pointwise convergence of cumulatives, will imply that the convergence to $\mathbf{N}(0, I)$ takes place in the half-space distance. Indeed, let Z be a $\mathbf{N}(0, I)$ vector. Convergence of $M_{\tau_n,n}(T_n - s_n)$ to $\mathbf{N}(0, I)$ in the half-space distance is equivalent to the statement that for any arbitrary sequence of vectors $v_n \in \mathbb{R}^k$ the Kolmogorov-Smirnov distance between $\mathbf{L}[v_n' M_{\tau_n,n}(T_n - s_n)]$ and $\mathbf{L}[v_n' Z]$ tends to zero. Since the Kolmogorov-Smirnov distance is scale invariant, it is enough to check this for sequences $\{v_n\}$ such that $\|v_n\| = 1$. For these one may suppose that the v_n have a limit v and the result is clear.

On the other hand condition (D) also implies that $|F_{s_n,n} - G_{s_n,n}|_h$ tends to zero. Indeed, since the distance is an affine invariant, one could change coordinates to replace $\Gamma_{s_n,n}$ by the identity matrix I. The balls $B_n(t, a)$ of condition (D) become then ordinary balls in $\mathbb{R}^k$ and the result follows by Slutzky's theorem.

Let then $G^*_{s_n,n}$ be $G_{s_n,n}$ recentered at zero. The preceding argument shows that the half-space distance between $\mathbf{L}[T_n - s_n | s_n]$ and $G^*_{s_n,n}$ tends to zero. Also, by (A), the half-space distance between $\mathbf{L}[T_n - s_n | s_n]$ and a Gaussian distribution with expectation zero and inverse covariance matrix $K_{\tau_n,n}$ will tend

to zero. This means that expressions such as $\Gamma_{s_n,n}^{-1} K_{\tau_n,n}$ or $K_{\tau_n,n}^{-1} \Gamma_{s_n,n}$ will tend to the identity I.

Now, let $G_{s_n,n}'$ be a normal distribution centered at s_n with inverse covariance matrix $K_{\tau_n,n}$. According to the above $|G_{s_n,n} - G_{s_n,n}'|_h \to 0$. This implies also that the total variation norm $\|G_{s_n,n} - G_{s_n,n}'\|$ tends to zero, since we are dealing with Gaussian measures.

The same applies to the pairs (τ_n, t_n). Thus the experiments $\mathbf{B}_n'' = \{G_{s_n,n}, G_{t_n,n}\}$ are asymptotically equivalent to the experiments $\mathbf{B}_n''' = \{G_{s_n,n}', G_{t_n,n}'\}$. It follows that the differences between affinities such as $\rho_n = \int [dP_{s_n,n}\, dP_{t_n,n}]^{1/2}$, $\rho_n' = \int [dF_{s_n,n}\, dF_{t_n,n}]^{1/2}$ and so forth up to ρ_n''' all tend to zero. In a triplet of homoschedastic normal distributions $(G_{s_n,n}', G_{t_n,n}', G_{\tau_n,n}')$ the function q_n is a distance. Thus since, by assumption, $q_n(\tau_n, s_n)$ and $q_n(\tau_n, t_n)$ remain bounded, so will $q_n(s_n, t_n)$. If so the differences between $\log \rho_n$, $\log \rho_n'$ and so forth up to $\log \rho_n'''$ must also tend to zero. Hence the result. □

Corollary. *Let the conditions (A) to (E) be satisfied and let* $q_n(\theta_n, \tau_n)$ *and* $q_n(s_n, \tau_n)$ *stay bounded. Then:*

i) The differences

$$q_n^2(T_n, s_n) \; - \; (T_n - s_n)' K_{\tau_n,n} (T_n - s_n)$$

tend to zero in $P_{\theta_n,n}$ *probability.*

ii) Condition (F) is satisfied.

This is clear.

Although the proof of Proposition 2 may appear devious, the result is hardly surprising. Note however that one can give examples where statistics T_n would satisfy (A) and (B) and where the experiments $\{F_{\theta,n}\,;\, \theta \in \Theta\}$ are approximable

by Gaussian shift experiments $\{G_{\theta,n}\,;\,\theta \in \Theta\}$ but where the conclusion of Proposition 2 does not hold. Such an example occurs in Le Cam and Yang (1988) as part of the discussion of the method of moments, page 515. Here we did make use of condition (D). Considering this, it is perhaps surprising that Proposition 2 admits a partial converse as follows.

Proposition 3. *Assume that conditions (A), (E) and (F) hold. Assume also, that with the notation of Proposition 1*

$$q_n^2(s_n, \tau_n) - (s_n - \tau_n)' K_{\tau_n,n}(s_n - \tau_n)$$

tends to zero whenever $q_n^2(s_n, \tau_n)$ *stays bounded. Then conditions (B) (C) and (D) also hold.*

Proof. We shall only give a brief sketch. For sets of the form $V_n(\tau_n, b) = \{\theta : q_n^2(\theta, \tau_n) \le b\}$ an argument of Le Cam (1977), repeated in Le Cam (1986) page 183, will show that conditions analogous to (B) and (C) will hold but with the entire set Θ replaced by the subsets $V_n(\tau_n, b)$. The argument cited uses an affinity number that is different from the Hellinger affinities used here, but the result is valid either way.

This does not use at all the conditions (E) and (F), but the result is only local. Using (E) and (F) one can carry out a patchwork argument as in Le Cam (1986) Chapter 11, Theorem 3. This will yield (B). To obtain (C) one can carry out a similar patchwork argument, as explained in Le Cam (1986), Chapter 5, Proposition 8, page 78. Another procedure is to prove both (C) and (D) at the same time using a method analogous to the one discussed in Le Cam (1986), Chapter 11, Section 8. Details will be left to the reader. □

It follows from the above propositions that, under the conditions (A) to (F), one can asymptotically treat $q_n^2(T_n, \theta_n)$ as if it was, under $P_{\theta_n,n}$, a central chi-square, just as would be the case for $(T_n - \theta_n)' K_{\theta_n,n}(T_n - \theta_n)$.

If the distributions are induced by $P_{\tau_n,n}$ instead and if $q_n^2(\theta_n, \tau_n)$ remains bounded then $q_n^2(T_n, \theta_n)$ will behave as a non-central chi-square, as would $(T_n - \theta_n)' K_{\tau_n,n}(T_n - \theta_n)$.

If on the contrary $q_n(\theta_n, \tau_n)$ tends to infinity, condition (F) implies that $q_n^2(T_n, \theta_n)$ tends to infinity in $P_{\tau_n,n}$ probability. Since, under such conditions, it would still be possible for $(T_n - \theta_n)' K_{\theta_n,n}(T_n - \theta_n)$ to tend to zero, this seems to imply that, under (A) to (F), confidence sets based on $q_n^2(T_n, \theta)$ may be somewhat better than the confidence ellipsoids of Wald. The function q_n is invariant under any and all one-to-one transformations of the parameter space. It takes into account differences in expectations and differences in covariances for the estimates T_n. Thus one would hope that tests or confidence sets based on them would avoid the difficulties pointed out by Hauck and Donner (1977) and by Vaeth (1985).

There are however some difficulties, one of which is that q_n^2 is more difficult to compute than Wald's quadratics. This suggests using instead of q_n^2 its analogue g_n^2 computed on an approximating heteroschedastic Gaussian experiment. According to Lemma 1, this may be possible. However g_n^2 is not invariant by reparametrization. Thus, care should be exerted.

There is another feature that deserves attention. The chi-square formulas suggest the use of confidence sets $\{\theta : q_n^2(T_n, \theta) \leq c_n\}$ where c_n is some constant, independent of θ. This is not in agreement with Neyman's classical derivation of confidence sets. One should take a $c_n(\theta)$ such that

$$P_{\theta,n}\{q_n^2(T_n, \theta) \leq c_n(\theta)\} \geq 1 - \alpha$$

for the selected significance level α. That such a selection of $c_n(\theta)$ may make a noticeable difference can be seen on the ordinary binomial distribution with probabilities $\binom{n}{x} p^x (1-p)^{n-x}$, $p \in [0,1]$. Take for instance $n = 200$. If $p = 1/200$ the probability that $T_n = \frac{x}{n}$ be equal to zero is roughly .36. Thus in a formula of the type $\{p : q_n^2(T_n, p) \le c_n\}$ the coefficient c_n should be such that $c_n \ge q_n^2(0, \frac{1}{200})$. But then if T_n would take value $T_n = 1/200$ the value $p = 0$ would be deemed acceptable. This is absurd. Sets of the form $\{p : q_n^2(T_n, p) \le c_n(p)\}$ can avoid such difficulties. Unfortunately, except under conditions such as (A) to (F), nothing much is known about the distribution of $q_n^2(T_n, \theta)$. Note that in the present binomial example an observed value $T_n = 1/200$ is hardly compatible with large p, say $p \ge .05$. The Gaussian approximations are of doubtful value.

The same recourse to the original theory of confidence sets shows that, instead of Wald's ellipsoids, one should have used sets such as $\{\theta : (T_n - \theta)' \Gamma_{\theta,n} (T_n - \theta) \le c_n\}$. Except under conditions where the $\Gamma_{\theta,n}$ vary little, there is no reason to hope that such sets would behave any better than Wald's ellipsoids.

We have said several times that our conditions (A) to (F) are too severe. It is often possible to get away with much less, according to the following simple observation: Suppose that you have evidence that the model $\{P_{\theta,n}; \theta \in \Theta\}$ can fit adequately. Suppose also that you have some auxiliary estimate θ_n^* with known variability that says that a certain subset $A_n \subset \Theta$ has a very high probability of covering the true θ. Then the validity of (A) to (F) on the entire Θ is of little relevance. What may matter is that the conditions be satisfied for $\theta \in A_n$, with the added possibility that condition (E) be modified to allow T_n to take values in Θ and not merely in A_n itself.

Further elaboration on such matters will be found in Section 6.

4. An example of M. Vaeth.

This section refers to the paper by M. Vaeth (1985) and in particular to the example discussed pages 205-206. Actually we shall not use the exact formulation of Vaeth but a simpler one that exhibits the same phenomenon but in terms of ‘‘exponential integrals’’ instead of Bessel functions.

For a fixed k let $f_k(x,\theta)$ be the density

$$f_k(x,\theta) \;=\; \frac{e^{-\theta x}}{F_k(\theta)}\,\frac{1}{x^k}\,; \quad x \ge 1,\;\; \theta > 0,$$

with respect to the Lebesgue measure on $[1,\infty)$. Here $F_k(\theta) = \int_1^\infty e^{-\theta x}\,\frac{1}{x^k}\,dx$ is the ‘‘exponential integral of order k’’ usually denoted $E_k(\theta)$. We shall use F_k instead of E_k to avoid possible confusion with expectations.

For such a family the following relations hold:

1) $F_{k+1}(\theta) = \frac{1}{k}\,[\,e^{-\theta} - \theta\,F_k(\theta)\,]$

2) $E_\theta X = \frac{F_{k-1}(\theta)}{F_k(\theta)}$

3) $E_\theta X^2 = \frac{F_{k-2}(\theta)}{F_k(\theta)}$

The maximum likelihood estimate $\hat\theta$ is the solution of the equation

$$X \;=\; \frac{F_{k-1}(\hat\theta)}{F_k(\hat\theta)} \;=\; E_{\hat\theta} X,$$

at least for $k \le 3$. For $k > 3$ the range of $E_\theta X$ is limited. One has $E_\theta X \le (k-1)/(k-2)$. Hence, for $X > (k-1)/(k-2)$ the m.l.e. $\hat\theta$ is equal to zero. Otherwise, if $k < 3$, the m.l.e. coincides with the estimate obtained by the

method of moments.

The phenomenon discussed by Vaeth is as follows. Consider the parametrization by $\beta(\theta) = E_\theta X$ so that X is the m.l.e. of $\beta(\theta)$.

To test the hypothesis that $\theta = \theta_1$ or to build confidence intervals, Wald suggests the use of the expression $[X - \beta(\theta_1)]/\hat{\sigma}a$ where $\hat{\sigma}$ is the m.l.e. of the standard deviation of X. For values of k such that $1 \le k \le 2$ this expression tends to zero as X tends to infinity. Thus large values of X, which tend to indicate values of θ close to zero, are held compatible with any value of θ. For $k > 1$, $1 < k \le 2$ this is not too disturbing since the sequences $\{f_k(\cdot, \theta)\}$ and $\{f_k(\cdot, \theta_1)\}$ are contiguous as $\theta \to 0$. In fact $f_k(x, \theta)$ tends to $f_k(x, 0) = (k-1)\, x^{-k}$, $x \ge 1$. For $k < 1$ the phenomenon in question does not occur: the coefficient of variation of X stays finite as $\theta \to 0$.

This can be easily checked by using the classical expansions of $F_k(\theta)$ for θ near zero. They can be found, for instance, in Abramowitz and Stegun (1964) or can be derived directly.

For $k = 1$ the m.l.e. of $\beta(\theta)$ is X itself. It has a variance

$$\operatorname{Var} X = \frac{F_{-1}(\theta)}{F_1(\theta)} - \left[\frac{F_0(\theta)}{F_1(\theta)}\right]^2.$$

Now $F_0(\theta) = e^{-\theta}/\theta$ and for θ tending to zero $F_1(\theta)$ behaves like $-\log\theta - \gamma$ where γ is Euler's constant $\gamma = -\int_0^\infty e^{-y} \log y \, dy \sim .57$. Thus for small θ the variance of X will behave like

$$\frac{F_{-1}(\theta)}{F_1(\theta)} = \left\{\frac{1}{\theta^2} e^{-\theta} + \frac{1}{\theta} e^{-\theta}\right\} \frac{1}{F_1(\theta)}$$

$$\sim \frac{1}{\theta^2 |\log\theta|}.$$

The maximum likelihood equation shows that, for large X, the m.l.e. $\frac{1}{\hat{\theta}}$ behaves like $X \log X$ so that the estimated standard deviation of X is of the order

$$\frac{1}{\hat{\theta}} \frac{1}{\sqrt{|\log \hat{\theta}|}} \sim X \sqrt{\log X},$$

hence the behavior of the criterion $[X - \beta(\theta_1)]/\hat{\sigma}$.

There is nothing particularly surprising about this fact. As $\theta \to 0$ the distribution of X is far from normal. Its expectation and standard deviation are poor indications of location and spread. For instance the median of X behaves like $1/\sqrt{\theta}$ while $E_\theta X$ behaves like $[\theta |\log \theta|]^{-1}$. The α^{th} quantiles behave like $\theta^{-\alpha}$. The distribution of X cannot be ''stabilized'' by a change of location and scale. The observed misbehavior of Wald's criterion extends to some other expressions. For instance if one uses an estimate $\bar{\theta}$ obtained by putting X equal to its median and then estimate the spread of the distribution by an interquartile range computed at $\bar{\theta}$ the resulting ratio will also tend to zero as $X \to \infty$.

All the arguments used above in this section use only one observation. If one has n independent identically distributed observations $X_1, X_2, \ldots, X_n$ their average $\bar{X}_n$ will still be the maximum likelihood estimate of $E_\theta X$. As explained by Vaeth (1985) the misbehavior noted for one observation persists for every value of n. Now let us see how the functions $q_n^2(T_n, \theta)$ of Section 3 can behave. Here $q_n^2(s,t) = nq^2(s,t)$ where q is the function computed for one observation only.

The argument of Section 3 depend on finding estimates T_n that are well behaved and in particular satisfy the condition (F) of Section 3. One can easily show that, here, the m.l.e. $\hat{\theta}_n$ satisfies condition (F) even though, as we shall see, it does not satisfy the other conditions of Section 3 uniformly on

$\Theta = (0, \infty)$.

Proposition 4. *Let* $\{p_\theta ; \theta \in (0, \infty)\}$ *be an exponential family of rank one in an arbitrary parametrization. Then for* n *independent identically distributed observations and for the m.l.e.* $\hat{\theta}_n$ *one has*

$$P_{\theta, n} \{ q_n^2 (\hat{\theta}_n, \theta) \geq 8z \} \leq 2 e^{-z}$$

for all $z > 0$.

Proof. An exponential family in its natural parametrization has the form

$$p_\theta (dx) = \exp \{ \theta x - A(\theta) \} \mu (dx)$$

for some measure μ. Thus

$$q^2 (\theta, t) = 8 \{ \frac{1}{2} [A(\theta) + A(t)] - A [\frac{\theta + t}{2}] \}.$$

Since A is a convex function, for θ fixed, $q^2 (\theta, t)$ increases as $|\theta - t|$ increases. Consider any particular $t > \theta$ and the test based on n observations that minimizes the sum of probabilities of error for θ and t. This sum of probabilities of error is $\| P_{\theta, n} \wedge P_{t,n} \| \leq \exp \{ -\frac{1}{8} q_n^2 (\theta, t) \}$. However, by concavity of the logarithm of likelihood ratios, if the test in question rejects t, the test of θ against $t' > t$ will also reject t'. Thus, except for probability at most $\exp \{ -\frac{1}{8} q_n^2 (\theta, t) \}$ for $P_{\theta, n}$, one will reject all $t' \geq t$. The same argument applies to values $s < \theta$. Hence the result, since the inequality $q_n^2 (\hat{\theta}_n, \theta) \geq 8z$ is invariant under all one to one reparametrizations. □

Note the 8z in the expression in curly brackets of Proposition 4. If $q_n^2 (\hat{\theta}_n, \theta)$ was actually chi-square one could replace it by 2 z for the same bound on the probabilities. Part of the loss can be attributed to the passage from $\| P_{\theta, n} \wedge P_{t,n} \|$ to Hellinger affinities but part may just be due to the fact that,

here, nothing much is known about the distribution of $\hat{\theta}_n$ or $q_n^2(\hat{\theta}_n, \theta)$.

In the present specific example one can obtain a variety of results about the asymptotic behavior of q_n^2. Of course, if θ is kept fixed, independent of n, the variables $\sqrt{n}[\,\overline{X}_n - E_\theta(X)\,]$ will be asymptotically normal and, $\hat{\theta}_n$ being the m.l.e., $q_n^2(\hat{\theta}_n, \theta)$ will be asymptotically χ_1^2. If, on the contrary, the true θ is a θ_n that depends on n and tends to zero, the behavior of $q_n^2(\hat{\theta}_n, \theta)$ can be very different from chi-square. To investigate what can happen consider two sequences $\{s_n\}$ and $\{t_n\}$ both tending to zero and such that $s_n > t_n$. For n observations the Hellinger transform of the pair $\{P_{s_n, n}, P_{t_n, n}\}$ has a logarithm of the form

$$\begin{aligned}\phi_n(\alpha) &= n \log \int [\,f(x, s_n)\,]^{1-\alpha} [\,f(x, t_n)\,]^{\alpha}\, dx \\ &= n\{\log F[\,(1-\alpha)s_n + \alpha t_n\,] - (1-\alpha)\log F(s_n) - \alpha \log F(t_n)\}\end{aligned}$$

where $F(v)$ is the exponential integral $F(v) = F_1(v) = \int_1^\infty e^{-vx} x^{-1}\, dx$.

For small z it has the expansion

$$F(z) = |\log z| - \gamma + \sum_{j=1}^{\infty} a_j z^j,$$

and its logarithm has the expansion

$$\log F(z) = \log|\log z| + \log\left[1 - \frac{\gamma}{|\log z|} + \frac{1}{|\log z|}\sum a_j z^j\right].$$

Let us first look at the $\log|\log|$ term in the expansion. For $\phi_n(\alpha)$ they give a first term

$$\omega_n(\alpha) = n\{\log|\log[\,(1-\alpha)s_n + \alpha t_n\,]| - (1-\alpha)\log|\log s_n| - \alpha \log|\log t_n|\}.$$

To investigate the behavior of this we shall assume $t_n < s_n$ and let $t_n = (1-\xi_n)s_n$, $0 < \xi_n < 1$. Then $(1-\alpha)s_n + \alpha t_n = s_n[\,1 - \alpha\xi_n\,]$ and $|\log[\,(1-\alpha)s_n + \alpha t_n\,]| = |\log s_n| - \log(1 - \alpha\xi_n)$.

Similarly $\quad |\log t_n| = |\log s_n| - \log(1 - \xi_n) \quad = |\log s_n| \left\{ 1 - \dfrac{\log(1 - \xi_n)}{|\log s_n|} \right\}.$

This yields

$$\omega_n(\alpha) \;=\; n \left\{ \log[\,1 - \frac{1}{|\log s_n|} \log(1 - \alpha\,\xi_n)\,] - \alpha \log[\,1 - \frac{\log(1 - \xi_n)}{|\log s_n|}\,] \right\}.$$

To study this it is convenient to introduce the notation

$$\delta_n \;=\; \frac{1}{|\log s_n|}$$

so that

$$\omega_n(\alpha) \;=\; n\,\{\log[\,1 - \delta_n \log(1 - \alpha\,\xi_n)\,] - \alpha \log[\,1 - \delta_n \log(1 - \xi_n)\,]\}.$$

We shall distinguish three cases:

Case A, $n\,\delta_n \to \infty$. Then, for $\omega_n(\alpha)$ to stay bounded, ξ_n must tend to zero. In such a case one has $-\log(1 - \alpha\,\xi_n) \sim \alpha\,\xi_n + \frac{1}{2}\,\alpha^2\,\xi_n^2$ and $-\log(1 - \xi_n) \sim \xi_n + \frac{1}{2}\,\xi_n^2$ and $\omega_n(\alpha)$ behaves like

$$n \left\{ \log[\,1 + \delta_n(\alpha\,\xi_n + \frac{1}{2}\,\alpha^2\,\xi_n^2)\,] - \alpha \log[\,1 + \delta_n(\xi_n + \frac{1}{2}\,\xi_n^2)\,] \right\}.$$

Expanding the logarithms once more, one sees that the terms in $\delta_n\,\xi_n$ cancel. The expression remains bounded if $n\,\delta_n\,\xi_n^2$ remains bounded. If $n\,\delta_n\,\xi_n^2 \to \sigma^2$ the term $\omega_n(\alpha)$ tends to $\frac{1}{2}\,\sigma^2\,[\,\alpha^2 - \alpha\,]$. This is the logarithm of the Hellinger transform for a Gaussian experiment.

This suggests looking at a family $\mathbf{F}_n = \{Q_{\lambda,n}\}$ where $Q_{\lambda,n}$ is $P_{\theta,n}$ with a θ taken equal to $s_n + \lambda\,s_n\,[\log s_n / n\,]^{1/2}$ with λ restricted so that $\theta > 0$. It can be shown that the experiments $\mathbf{F}_n$ converge to a Gaussian shift experiment linearly indexed by λ. Thus, the corresponding $q_n^2(\hat{\theta}_n, \theta)$ will still behave asymptotically

as chi-square, with one degree of freedom.

Case B, $n\delta_n \to b$, finite, positive. Then $\omega_n(\alpha)$ can stay bounded for values ξ_n that stay away from zero and unity. If $\xi_n \to \xi$, $0 < \xi < 1$ then $\omega_n(\alpha)$ tends to $-b[\log(1 - \alpha\xi) - \alpha\log(1 - \xi)]$. This shows that, under $P_{s_n,n}$, the distribution of $\log dP_{t_n,n}/dP_{s_n,n}$ tends to a shifted gamma distribution. The sequences are contiguous.

Case C, $n\delta_n \to 0$. In this case it is possible to let ξ_n tend to unity in such a way that $n\delta_n \log(1 - \xi_n)$ stays bounded. If $-n\delta_n \log(1 - \xi_n)$ tends to a limit b then $\omega_n(\alpha) \to -b\alpha$. This is the log Hellinger transform for a pair (Q_0, Q_1) where the part of Q_1 that is dominated by Q_0 has a constant density equal to e^{-b}. The part of Q_1 that is Q_0 singular has mass $1 - e^{-b}$. This implies that the sequence $\{P_{s_n,n}\}$ is contiguous to $\{P_{t_n,n}\}$ but the reverse is not true. Here $q_n^2(s_n, t_n)$ tends to 4 b and $\| P_{s_n,n} \wedge P_{t_n,n} \|$ tends to e^{-b}.

In the above derivations we have used only the log log term in the expansion of log F. However, it is easy to check that the other terms tend to zero.

In all cases, the logarithm of likelihood ratio $\Lambda_n = \log \dfrac{dP_{t_n,n}}{dP_{s_n,n}}$ has the form $\Lambda_n = a_n \overline{X}_n + b_n$ where a_n and b_n are constants and where $\overline{X}_n = \dfrac{1}{n}\sum_{j=1}^{n} X_j$. Since $\overline{X}_n$ is the maximum likelihood estimate of its expectation $\beta(\theta) = E_\theta \overline{X}_n$ the expression $q_n^2(\hat{\theta}_n, \theta)$ can also be written in terms of $\overline{X}_n$ and β as, say $\overline{q}_n^2(\overline{X}_n, \beta)$. Since, by Proposition 4, $\overline{q}_n^2(\overline{X}_n, \beta_n)$ remain bounded in $P_{\theta,n}$ probability no matter what $\beta_n = \beta(\theta_n)$ does, one can approximate $\overline{q}_n^2(\overline{X}_n, \beta_n)$ by the expressions used above for $-8\,\omega_n(1/2)$. In case B this leads to an approximation of the type

$$-8\,b\log 2 + 8\,b\log\left\{\left[\frac{\overline{X}_n}{\beta_n}\right]^{1/2} + \left[\frac{\beta_n}{\overline{X}_n}\right]^{1/2}\right\}.$$

Since $(\overline{X}_n - c_n)/\beta_n$ is approximately distributed as a gamma variable with exponent b, this does not seem to behave like a chi-square. For case C the variables $\overline{q}_n^2(\overline{X}_n, \beta_n)$ seem to behave in the same manner as $(n/\log\beta_n)\log\overline{X}_n/\beta_n$. By Proposition 4 this must stay bounded in $P_{\theta_n, n}$ probability for $\beta_n = E_{\theta_n} X$. If θ_n is replaced by a $t_n = (1 - \xi_n)\theta_n$ such that $(n/\log\beta_n)\log(1 - \xi_n) \to b$, then $(n/\log\beta_n)\log(\overline{X}_n)/\beta_n$ will have for $P_{t_n, n}$ a distribution with a mass $1 - e^{-b}$ tending to infinity. This should be taken into account in the construction of confidence intervals.

5. Some heteroschedastic gaussian cases.

As seen in Section 3 heteroschedastic gaussian experiments occur routinely in asymptotic theory. In fact the conditions (A) to (E) of Section 3 provide for a situation where the experiments $\mathbf{E}_n = \{P_{\theta, n}\,;\, \theta \in \Theta_n\}$ are such that $\Delta(\mathbf{E}_n, \mathbf{G}_n) \to 0$ for the heteroschedastic experiment $\mathbf{G}_n$ of condition (D). For this reason we shall study here the behavior of some heteroschedastic gaussian experiments. However, for simplicity we shall only use parameters θ that run through an interval of the real line, say $\Theta = [a, \infty)$ where a is a large positive number.

Let X be a normal variable whose distribution depends on a parameter $\theta \in [a, \infty)$. Assume that, given θ, the variable X has expectation θ and a variance $\sigma^2(\theta) = \dfrac{1}{\gamma(\theta)}$.

The family so obtained defines an affinity

$$\rho(s,t) = \left\{\frac{4\gamma(s)\gamma(t)}{[\gamma(s) + \gamma(t)]^2}\right\}^{1/4} \exp\{-\frac{1}{4}\,\frac{\gamma(s)\gamma(t)}{\gamma(s) + \gamma(t)}\,|t - s|^2\}$$

yielding

$$q^2(s,t) = -8\log\rho(s,t)$$
$$= 2\log\left\{\frac{[\gamma(s)+\gamma(t)]^2}{4\gamma(s)\gamma(t)}\right\} + 2\frac{\gamma(s)\gamma(t)}{\gamma(s)+\gamma(t)}|t-s|^2.$$

We shall be interested in situations where γ is a smooth decreasing function that tends to zero rapidly as $\theta \to \infty$. For a first example let us take $\gamma(\theta) = e^{-2\theta}$. Then

$$q^2(s,t) = 4\log\cosh(t-s) + \frac{e^{-(s+t)}}{\cosh(t-s)}(t-s)^2,$$

indicating that, for s and t large, the main contribution to $q^2(s,t)$ will arise from the first term. This is the term that takes into account the difference between the variances at s and t.

The negative of the logarithm of the likelihood function is

$$\frac{1}{2}(X-\theta)^2 e^{-2\theta} + \theta = \frac{1}{2}[X - 1/2\log v]^2 v^{-1} + \frac{1}{2}\log v,$$

where v is the variance $v = e^{2\theta}$. The maximum likelihood equation is

$$\exp\{2\hat{\theta}\} = (X-\hat{\theta}) + (X-\hat{\theta})^2.$$

In terms of the variance v this becomes

$$\hat{v} = (X - \frac{1}{2}\log\hat{v}) + (X - \frac{1}{2}\log\hat{v})^2,$$

showing that, for $|X|$ large, $\hat{v}$ will behave approximately like $X^2 + X$. Approximate solution of the likelihood equation shows that, for $|X|$ large, $\hat{\theta}$ (restricted to (a,∞), a large) behaves approximately like $\log|X|$. Some standard methods of constructing confidence intervals can lead to very different results. The standard ''equal tails'' intervals with probability of coverage near .955 would be given by inverting the inequalities $\theta - 2e^{\theta} \le X \le \theta + 2e^{\theta}$. However, for $X \ge -(1+\log 2)$ the lower barrier is ineffective. The resulting intervals would

be half infinite, of the form $[c(X), \infty)$. A similar phenomenon occurs for X negative but $|X|$ large.

If, on the contrary, one uses intervals of the type $X - 2\hat{\sigma} \le \theta \le X + 2\hat{\sigma}$ where $\hat{\sigma}$ is estimated, then the intervals would take the form $[a, c_1(X)]$ with an ineffective bounding for small values. For instance if one estimate θ by $\log|X|$ for $|X| \ge 1$ one would estimate $\hat{\sigma}$ by $|X|$ and get intervals of the type $X - 2|X| \le \theta \le X + 2|X|$. The lower bound is always negative and therefore ineffective since we assume $\theta \ge a$ with $a > 0$, large. As $|X| \to \infty$ these intervals produce an instance of the Hauck-Donner-Vaeth phenomenon. They accept any finite value of θ.

For confidence intervals based on the function q the situation is different. Let us take some estimate $\tilde{\theta}$. If $\tilde{v} = \sigma^2(\tilde{\theta})$ is large, the main contribution to $q(t, \tilde{\theta})$ will be $4 \log\cosh(t - \tilde{\theta})$. Thus the intervals will be given approximately by an inequality of the type $\{\theta : \cosh(\theta - \tilde{\theta}) \le e^{b/4}\}$, that is $|\theta - \tilde{\theta}| \le \cosh^{-1}(e^{b/4}) = c$.

For simplicity, let us use the crude estimate $\tilde{v} = X^2$ so that $\tilde{\theta} = \frac{1}{2}\log X^2 = \log|X|$ with $|X|$ assumed ≥ 1. Then we have intervals equivalent to $|\theta - \log|X|| \le c$. For θ very large these intervals have a probability of coverage about equal to $P\{\log|\xi| \le c\}$ for a ξ with a $\mathbf{N}(0,1)$ distribution. Thus one can consider intervals obtained from a value c of the order of $\log 2$. Note that these have a fixed length as $|X| \to \infty$. This is in sharp contrast with the intervals obtained from chi-square type formula.

One could object that, for the normal family $\mathbf{N}(\theta, e^{2\theta})$ used above, there is nothing ''asymptotic''. However, the same kind of analysis will apply for each fixed n to a family of the type $\{\mathbf{N}(\theta, \frac{1}{n}e^{2\theta}) ; \theta \ge 1\}$. This shows that for many densities $f(x, \theta)$ that are sufficiently smooth functions of θ, the analysis will

apply for n i.i.d. observations $X_1, \ldots, X_n$ provided that the Fisher information decreases exponentially fast as $\theta \to \infty$.

The situation described here may seem extreme in that one would not expect to encounter very often observations X whose standard deviation is an exponential function of their expectation. However, the same kind of analysis can be carried out for a variety of other cases. In fact the exponential increase was suggested by the standard binomial if put in its logit form. Let Z be a binomial $\mathbf{B}(n, p)$ variable. Take as new variable $X = \log \frac{Z}{n-Z}$ and as new parameter $\theta = \log \frac{p}{1-p}$. For fixed θ as $n \to \infty$ $\sqrt{n}(X - \theta)$ is asymptotically normal with mean zero and variance $[1 + e^{\theta}]^2 e^{-\theta}$. Thus, if we pretend that X is actually normal, we are in the situation just described with variances equivalent to $e^{|\theta|}$ for $|\theta|$ large. This means that an analysis similar to the above could be applied to this case. However, as we have already seen in Section 3, one need to take further precautions because the asymptotic normality is far from been uniform in θ.

6. The choice of estimates, the domain of q_n and other remarks.

The use of confidence sets of the type $\{\theta : q_n^2(T_n, \theta) \leq c_n(\theta)\}$ require the use of appropriate estimates T_n, especially if one takes function $c_n(\theta)$ that are independent of θ.

In this restricted case the least one can require is that $q_n^2(T_n, \theta)$ be bounded in probability independently of θ.

For independent observations, with individual distributions $p_{\theta,j,n}$, existence of such estimates has been proved by Le Cam (1975), (1986) and Birgé (1983) under a metric dimension restriction on the set Θ. Instead of the function q_n^2, Le Cam used a square metric $H_n^2(s, t) = \sum_j h_{n,j}^2(s, t)$ where

$$h^2_{n,j}(s,t) \;=\; \frac{1}{2}\int [\,((dp_{s,j,n})^{1/2} \;-\; (dp_{t,j,n})^{1/2}\,]^2 .$$

The arguments of Le Cam or Birgé even provide bounds for $\sup_\theta E_\theta H^2_n(T_n,\theta)$ that depend on the behavior of the dimension function of Θ for the metric H_n. Such bounds can be translated into probabilistic bounds for $q^2_n(T_n,\theta)$ whenever the individual $h_{n,j}(s,t)$ are small. In fact they give bounds for $E_\theta\, q^2_n(T_n,\theta)$ for a Poissonized version of the system $\{p_{\theta,j,n}\,;\,\theta \in \Theta, j = 1,2,...\}$. However the arguments used in the constructions of Le Cam or Birgé are too crude to provide approximations to the distributions of $q^2_n(T_n,\theta)$.

The cases covered by the arguments of Le Cam (1975) (1986) or Birgé (1983) are much less restricted than those where our Assumptions (A) to (E) of Section 3 can be satisfied. However, even (A) to (E) can be satisfied in cases where the maximum likelihood estimates behave in a disastrous fashion. There exist many examples. To cite only one, consider densities of the form $f(x-\theta)$ with $f(x) = c\,|x|^{-1}\exp\{-|x|^2/2\}$ where $x \in \mathbb{R}^4$ and $|x|$ is the length of x. In this example, and many others, the LAN based theory of Le Cam (1974), (1986) will provide suitable estimates. Note however that the LAN theory and the conditions (A) to (E) rely on the virtues of special parametrizations. This is clear for (A) to (E). The conditions (B) (C) and (E) or (F) are parametrization free. However Condition (A) is not.

If the smallest eigenvalues of the matrices $M_{\tau_n,n}{}'\, M_{\tau_n,n}$ of condition (A) tend to infinity as $n \to \infty$, one can claim a form of local invariance under smooth transformations of the parameter space. If ϕ is such a smooth transformation, replacing T_n by $\phi(T_n)$ and θ by $\phi(\theta)$ will not matter much locally since for T_n close to θ one has approximately $\phi(T_n) - \phi(\theta) \sim \dot\phi(\theta)(T_n - \theta)$ for a derivative $\dot\phi$. However for large $|T_n - \theta|$ the difference $\phi(T_n) - \phi(\theta)$ may bear no relation to $\dot\phi(\theta)(T_n - \theta)$. The difficulties pointed out by Hauck and Donner and by

Vaeth stem at least in part from the fact that $|T_n - \theta|$ may be large but $|[\dot{\phi}(\theta)]^{-1}[\phi(T_n) - \phi(\theta)]|$ may be small. If so, that can mean that confidence sets from quadratics based on T_n may be all right while those based on $\phi(T_n)$ may be poor. The choice of parametrization leads to complex problems. For instance one can try to make the distribution under $P_{\theta,n}$ of the quadratic $(T_n - \theta)' K_{\theta,n}(T_n - \theta)$ close to that of a chi-square. An alternative endeavor is to try to make the experiments $\{F_{\theta,n}; \theta \in \Theta\}$ as close to Gaussian shift ones as possible. For this latter aim, see Mammen (1987) and Section 7 below.

Still another method is to avoid reparametrizations of Θ altogether. Le Cam (1986), Chapter 11, proposes to ignore the vector structure of Θ and work instead in a space $\mathbf{M}_0(\Theta)$ of finite signed measures on Θ. The estimates T_n are then replaced by ''centering variables'' Z_n with values in $\mathbf{M}_0(\Theta)$ and the problem becomes one of choosing appropriate quadratic forms Γ_n on $\mathbf{M}_0(\Theta)$. It is shown in Le Cam (1986) that, for the asymptotically Gaussian shift case, one can construct quadratic forms with suitable properties. They are chosen locally around auxiliary estimates $\hat{\omega}_n$ of θ with values in Θ and in such a way that $\Gamma_n(Z_n - \delta_\tau)$ becomes large if $P_{\theta,n}$ is the true distribution but $q_n(\tau, \theta)$ becomes large. (Here δ_τ is the Dirac measure that gives mass unity to τ).

This technique also gives a partial answer to possible extensions of the definition of $q_n^2(T_n, \theta)$ if T_n takes values outside of Θ. However the technique is entirely dependent on local approximability of the $\{P_{\theta,n}; \theta \in \Theta\}$ by Gaussian shift experiments and its application may be complex.

This brings us back to the conditions (A) to (E) and the fact that they are unlikely to hold for a parameter set Θ that is not the entire $\mathbb{R}^k$. In very many cases one does have to deal with proper subsets Θ of $\mathbb{R}_k$ on which one could prove that conditions (A) to (D) hold for suitably constructed statistics T_n. However these statistics do not always take their values in $\mathbb{R}^k$.

It is their tempting to enlarge the domain of the function q_n to pairs (s, t) where either s or t is in $\mathbb{R}^k$, but not in Θ. For some cases, such as the binomial $\mathbf{B}(n, p)$, $0 < a \le p \le b < 1$, an extension is immediate and then (A) to (D) will be satisfied. More generally, since q_n^2 is a monotone increasing function of a Hilbertian distance, the problem is analogous to the following problem: One is given a set $\Theta \subset \mathbb{R}^k$ with a Hilbertian distance H. Here H^2 would be $H^2(s, t) = \frac{1}{2} \int [(dP_{s,n})^{1/2} - (dP_{t,n})^{1/2}]^2$. Can one extend the defunction of H to all of $\mathbb{R}^k$ and what would be a reasonable extension?

We do not know what the situation is. Fortunately one can often get around the problem by a simple procedure that also allows dispensing with the severe restriction that conditions (A) to (D) be valid uniformly on all of Θ.

The procedure is simply to first use some trustworthy confidence region A_n with good coverage probability and then verify (A) to (D) on A_n only. One lets the estimate T_n take values arbitrarily in Θ or in any set for which q_n can be defined. this can lead to statements lacking in mathematical aesthetics but not in practical value, since, in practice, n does not tend to infinity.

To give an example, let us consider Vaeth's situation as described in Section 4. That is let us assume that we observe variables $X_1, X_2, \ldots, X_n$ that are all independent and distributed according to the density

$$f(x, \theta) = \frac{e^{-\theta x}}{F_1(\theta)\, x}, \quad x \ge 1, \quad \theta > 0.$$

Let us assume that n is very substantial, say $n = 10^6$. Then one can construct an empirical cumulative distribution function $H_n(x) = \frac{1}{n} \sum_{j=1}^{n} I(X_j \le x)$. Let $H(x, \theta)$ be the cumulative corresponding to $f(\cdot, \theta)$. Define a restricted range R_n for θ by

$$R_n = \{\theta : \sqrt{n} \sup_x |H_n(x) - H(x,\theta)| \leq 4\}.$$

For any θ_0, the probability under $P_{\theta_0,n}$ that R_n does not contain θ_0 is of the order of 10^{-13}. This does not tend to zero as or tends to infinity but it is close enough to zero for many and perhaps most practical purposes.

For $n = 10^6$ this would limit us to a range $R_n = \{\theta ; \sup_x |H_n(x) - H(x,\theta)| \leq 4.10^{-3}\}$.

The first thing to do is to check that R_n is not empty. If R_n is empty, as will happen often, the modelling by the densities $f(x,\theta)$ is not credible and one should rethink the problem anew. If R_n is not empty and if one is sufficiently convinced of the validity of the model, one can argue as follows.

For the densities $f(x,\theta)$ the vertical distance between cumulatives or just 1/2 of the L_1-distance between densities. Since this is larger than the square Hellinger distance, one sees that two values (s,t) that belong to R_n must be such that

$$h^2(s,t) = \frac{1}{2}\int \{[f(x,s)]^{1/2} - [f(x,t)]^{1/2}\}^2 dx \leq \frac{8}{1000}$$

giving for $q^2(s,t) = -8\log[1 - h^2(s,t)]$ an upper bound approximately equal to .065.

Assuming again that the model is correct one can try using the inequality of Proposition 4, Section 4 that says that for the maximum likelihood estimate $\hat{\theta}_n$ one has

$$P_{\theta,n}\{nq^2(\hat{\theta}_n, \theta) \geq 222\} \leq 10^{-12}.$$

Applying this to pairs (s,t) as above, this would yield a bound of the order of $q^2(s,t) \leq (.045)\,10^{-2}$, a sizeable improvement on the preceding bound, leading to a more restricted range R_n^*.

The next step would be to check whether on R_n, or on the more restricted range suggested by $\hat{\theta}_n$, a Gaussian shift approximation seems appropriate. Now one can try Gaussian shift approximations whose random term is linear in θ itself, or in $\beta(\theta) = E_\theta X$ or in any other parametrization that seems suitable. According to Mammen (1987) the parametrization that yields the best fitting local Gaussian shift experiment is one where one uses a function $\theta \rightsquigarrow \xi(\theta)$ such that $|\xi(t) - \xi(\theta)|^2 = nh^2(\theta, t)$, or, approximately, $|\xi(t) - \xi(\theta)|^2 = q_n^2(t, \theta)$.

As explained before, the choice of parametrization will not affect at all the distribution of terms such as $q_n^2(\hat{\theta}_n, \theta) = nq^2(\hat{\theta}_n, \theta)$. It will however affect the quadratic expressions used to approximate $q_n^2(\hat{\theta}_n, \theta)$ and the quality of the chi-square approximation to the distribution of the quadratic. For instance, if we use the parametrization by $\beta(\theta)$, one would use a quadratic of the type $[\overline{X}_n - \beta(\theta)]^2 [\mathrm{Var}_{\hat{\theta}} X]^{-1}$. This seems safe as long as $\mathrm{Var}_\theta X$ does not change much in the range R_n^* to which we have limited θ, and as long as the left most point of R_n^* stays sufficiently far from zero. In fact, let l_n be the left most point of the range R_n^*. According to Section 4, one will still be able to use normal approximations even if l_n is small provided that $n |\log l_n|^{-1}$ be large.

In any event this shows that the technique can be used in spite of the fact that the assumptions (A) to (E) do not hold uniformly on Θ.

7. Reduction to the homoschedastic case.

As mentioned previously, the difficulties pointed out by Hauck and Donner (1977) and Vaeth (1985) arise in part from the fact that the estimates T_n may have covariances that vary rapidly with θ. If Θ is a subset of the real line, one can use variance stabilizing transformations. Vaeth (1985) shows that such transformations will prevent gross misbehavior of Wald's criterion, at least for the case of exponential families.

When $\Theta \subset \mathbb{R}^k$, $k > 1$, the situation is more complex. Under the conditions (B) (C) (D) of Section 3 one can approximate the experiments $\mathbf{E}_n = \{P_{\theta,n}\,;\, \theta \in \Theta\}$ by heteroschedastic Gaussian ones, say $\mathbf{G}_n = \{G_{\theta,n}\,;\, \theta \in \Theta\}$. Such an approximation yields for each θ a quadratic form on $\mathbb{R}^k$, using as a matrix the inverse covariance matrix of the Gaussian $G_{\theta,n}$. Thus we obtain on Θ a structure of a Riemannian manifold. Transformations of Θ that would replace these variable inverse covariance matrices $\Gamma_{\theta,n}$ by a fixed matrix amount to an isometric imbedding of the Riemannian manifold into a Euclidean space. There are theorems that indicate the possibility of such embeddings: E. Cartan says that, locally, a manifold of dimension k can be imbedded in R^m for $m = \frac{1}{2}\,k\,(k + 1)$. This is only a local result. There is a global embedding result of Nash that says that the entire manifold, if it is of class C^3 can be isometrically embedded in $\mathbb{R}^m$ for a value of m greatly in excess of $\frac{1}{2}\,k\,(k + 1)$. See for instance J.T. Schwartz (1969), page 43 and J. Dieudonné (1971), volume 4, page 341. A survey of the situation by M.L. Gromov and V.A. Rokhlin (1970) gives a value $m = \frac{1}{2}\,k\,(k + 1) + 3k + 5$.

Unfortunately the dimension m of the embedding space is very much larger than that of the original $\Theta \subset \mathbb{R}^k$. Thus Θ becomes a very thin subset of $\mathbb{R}^m$. Of course, to prevent misbehavior of Wald's criteria it is not necessary to make the inverse covariance matrices $\Gamma_{\theta,n}$ of $G_{\theta,n}$ exactly constant. Taking a particular θ_0 as a reference point, let $\Gamma_n = \Gamma_{\theta_0,n}$ and take positive definite square roots. If the eigenvalues of $\Gamma_n^{-1/2}\ \Gamma_{\theta,n}\ \Gamma_n^{-1/2}$ do stay close to unity as θ varies then Wald's criteria will at least not be subject to gross misbehavior. This suggest the possibility of embedding that are not exactly isometric but only close to isometric. For these we do not know what the situation is.

The problem of selection of a parametrization is also related to the problem of selection of best fitting local Gaussian shift experiments studied by E. Mammen (1987). He shows that for one dimensional exponential families around a point θ_0, the optimal choice is a parametrization in which a the expectation $\xi(t)$ of the normal approximation satisfies the relation

$$H_n^2(t, \theta_0) = |\xi(t) - \xi(\theta_0)|^2$$

for $H_n^2(s,t) = nh^2(s,t)$ as in the beginning of Section 6. According to Mammen (1988), personal communication, the result extends to k-dimensional situations for $k > 1$. The possibility of embedding Θ in a Hilbert space and using global Gaussian shift approximations is also discussed in Le Cam (1986) page 266.

Note however that a reparametrization cannot change the distributions of likelihood ratios. Thus, for any such embedding to lead to good Gaussian shift approximations one needs at least that, locally the experiments $\{P_{\theta,n}; \theta \in A_n\}$ $A_n = \{\theta : q_n^2(\theta, \theta_0) \leq c\}$ admit good Gaussian shift approximations.

Appendix

The affinity for two Gaussian measures

In Section 3 and Section 5 we have used a formula for the Hellinger affinity $\int [\, dG_1\, dG_2\,]^{1/2}$ between two Gaussian measures that may have different means and covariance matrices. The formula goes back at least to C.H. Kraft (1955). However, in the published version of Kraft's thesis it was misprinted and is barely recognizable. Matusita (1967) also gives an expression, but it is different from the form used here. The derivation of the formula is not difficult but its implications are many. Thus we shall present here two derivations. One is intended for finite dimensional situations. The other is meant for arbitrary dimensions.

Let P and Q be two gaussian measures on a finite dimensional vector space $\mathbb{R}^k$. Assume that P has expectation θ and that the inverse of its covariance matrix is Γ so that it has a density

$$\frac{(\det\Gamma)^{1/2}}{(2\pi)^{k/2}} \exp\{-\frac{1}{2}(x-\theta)'\Gamma(x-\theta)\}$$

with respect to the Lebesgue measure of $\mathbb{R}^k$. Similarly, let Q have expectation t and inverse covariance matrix K.

Multiplying the square root of the densities will yield, in the exponent, a quadratic form that, except for the coefficient $(-\frac{1}{4})$, is equal to

$$(x-\theta)'\Gamma(x-\theta) + (x-t)'K(x-t).$$

To reduce this to a tractable form assume that Γ and K are both invertible and introduce a centering v by

$$(\Gamma + K)v = \Gamma\theta + Kt.$$

Then the above quadratic becomes

$$(x-v)'(\Gamma+K)(x-v) + (v-\theta)'\Gamma(v-\theta) + (v-t)'K(v-t).$$

The term $(v-\theta)$ can be expressed as $(v-\theta) = (\Gamma+K)^{-1}K(t-\theta)$. Similarly $(v-t) = (\Gamma+K)^{-1}\Gamma(\theta-t)$.

Let M be the matrix $M = \frac{1}{2}(\Gamma+K)$ and let $\Delta = \frac{1}{2}(\Gamma-K)$ so that $\Gamma = M+\Delta$ and $K = M-\Delta$. Then $(v-\theta)$ takes the form $(v-\theta) = \frac{1}{2}(t-\theta) - \frac{1}{2}M^{-1}\Delta(t-\theta)$ and $v-t$ is obtained by changing Δ to $-\Delta$. Then the sum of the two terms of the quadratic that do not involve $(x-v)$ will yield a quadratic in $(\theta-t)$ with a matrix $\frac{1}{4}\{(I-M^{-1}\Delta)'\Gamma(I-M^{-1}\Delta) + (I+M^{-1}\Delta)K(I+M^{-1}\Delta)\}$. Direct computation shows this to be equal to $\frac{1}{2}[M-\Delta'M^{-1}\Delta]$. Thus the term in the exponent of

the square root of the product of the densities is

$$-\frac{1}{4}(x - v)'(\Gamma + K)(x - v) - \frac{1}{8}(t - \theta)'[M - \Delta' M^{-1}\Delta](t - \theta).$$

Now integrate out the variable x. One will get a term equal to $[\det \frac{(\Gamma + K)}{2}]^{-1/2}$ multiplied by $\exp\{-\frac{1}{8}(t - \theta)'[M - \Delta' M^{-1}\Delta](t - \theta)\}$. Combining this with the determinantal coefficients of the densities yields an affinity equal to

$$\frac{[(\det \Gamma)(\det K)]^{1/4}}{[\det M]^{1/2}} \exp\{-\frac{1}{8}(t - \theta)'[M - \Delta' M^{-1}\Delta](t - \theta)\}$$

and the result quoted in Section 2 follows by using the fact that the determinant of a product of two matrices (square of same order) is the product of the determinants. This shows that $\int \sqrt{dP\, dQ}$ is a product of two terms. One of them involves differences between expectations in the form

$$\exp\{-\frac{1}{8}(t - \theta)'[M - \Delta' M^{-1}\Delta](t - \theta)\}.$$

The other involves only the inverse of the covariance matrices in the form

$$\{\det[I - M^{-1}\Delta)^2]\}^{1/4}.$$

For many uses this determinantal form is not convenient. A form using the covariances themselves can also be used. Let $A = \Gamma^{-1}$ be the covariance matrix of P and let $B = K^{-1}$ be the corresponding matrix for Q. The determinantal term in $\int \sqrt{dP\, dQ}$ can also be written

$$[\det A]^{-1/4}[\det B]^{-1/4}\{\det[\frac{A^{-1} + B^{-1}}{2}]\}^{-1/2}.$$

Its fourth power is

$$\{[\det A][\det B]\det\{[\frac{A^{-1} + B^{-1}}{2}]^2\}\}^{-1} = \frac{\det AB}{\det(\frac{A+B}{2})^2}.$$

Write $S = \frac{1}{2}(A + B)$, $D = \frac{1}{2}(A - B)$. Then $A = S + D$, $B = S - D$ and $AB = S^2 - D^2$. Thus the determinantal term can also be written as fourth root of

$$\det[\, I - (S^{-1} D)^2 \,].$$

This form is particularly convenient for passage to infinite dimensions. To look at such a case, take a vector space V over the real numbers. Let X and Y be two processes $v \rightsquigarrow X(v)$ and $v \rightsquigarrow Y(v)$ indexed by V and linear in v. Assume that $X(v)$ is Gaussian with expectation zero and variance $E\,|X(v)|^2 = \tilde{A}(v)$. Similarly let $E\,Y(v) = 0$ and $E\,|Y(v)|^2 = \tilde{B}(v)$. These are squares of Hilbertian seminorms on V. The processes X and Y yield distributions that can be represented by measures P and Q on the algebraic dual of V.

If V contains sequences $\{v_n\}$ such that $\tilde{A}(v_n) + \tilde{B}(v_n)$ stays bounded away from zero but $\min[\,\tilde{A}(v_n), \tilde{B}(v_n)\,] \to 0$ then P and Q are obviously disjoint. Thus if $\int [\,dP\,dQ\,]^{1/2} = \rho(P, Q) > 0$ the two seminorms $\tilde{A}^{1/2}$ and $\tilde{B}^{1/2}$ must be equivalent in the sense that there exists numbers a, b, $0 < a \le b < \infty$ such that $a\,\tilde{A}(v) \le \tilde{B}(v) \le b\,\tilde{A}(v)$ for all v. This shows that there will be no loss of generality in assuming that if $\tilde{S} = \frac{1}{2}(\tilde{A} + \tilde{B})$ then $\tilde{S}(v) = 0$ only at $v = 0$. Also one can assume V complete for the norm $\tilde{S}^{1/2}$ so that $(V, \tilde{S}^{1/2})$ is a Hilbert space. Let $\mathbf{X}$ be the dual of V for the norm $\tilde{S}^{1/2}$. It is clear that the inner product $[\,\cdot\,|\,\cdot\,]_A$ defined by $\tilde{A}$ on V can be represented as $[\,u\,|\,v\,]_A = \langle u, Av \rangle$ where the bracket $\langle u, x \rangle$ is the evaluation of the linear function $x \in \mathbf{X}$ at $u \in V$ and A is a linear map of V onto $\mathbf{X}$ such that $\langle u, Av \rangle = \langle v, Au \rangle$. Similarly, the inner product corresponding to B can be written $\langle u, Bv \rangle$ and the inner product defined by S is $\langle u, Sv \rangle$ where S is the canonical identification of the Hilbert space $(V, \tilde{S}^{1/2})$ with its dual. The inverse S^{-1} of that identification map sends $\mathbf{X}$ onto V. We

shall also denote the norms of $(V, \tilde{S}^{1/2})$ and of its dual by the symbols $\|v\|$, so that $\|v\|^2 = \tilde{S}(v)$.

Consider then a finite dimensional subspace H of V. For the processes X and Y restricted to H we have distributions P_H and Q_H. It is clear that $\int [dP_H\, dQ_H]^{1/2} \geq \int [dP\, dQ]^{1/2}$.

Let Π be the orthogonal projection of V onto H in the Hilbert space $(V, \tilde{S}^{1/2})$. Let Π^t be the transpose of Π on the dual space $\mathbf{X}$ of V. One can show that $\Pi^t S\Pi = S\Pi$. Indeed, let $H^0 = \{x : \langle v, x\rangle = 0, \text{ all } v \in H\}$ be the polar of H in $\mathbf{X}$. For any $y \in \mathbf{X}$ one has $\langle (1-\Pi)v, y\rangle = \langle v, (1-\Pi)^t y\rangle$. Thus if one has $v \in H$ $\langle v, (I-\Pi)^t y\rangle = 0$ and therefore $(I-\Pi)^t y \in H^0$. The defining relation for H^0 can also be written $\langle \Pi v, x\rangle = 0 = \langle v, \Pi^t x\rangle$ for all $v \in V$. This means that if $x \in H^0$ then $\Pi^t x = 0$. Therefore $H^0 = (I-\Pi)^t \mathbf{X}$.

Now take any $v \in V$ and consider $\langle w, S\Pi v\rangle = \langle w, \Pi^t S\Pi v\rangle + \langle W, (1-\Pi)^t S\Pi v\rangle$. The second term on the right is equal to $\langle (I-\Pi)w, S\Pi v\rangle = [(I-\Pi)w \,|\, \Pi v]$ where $[\cdot|\cdot]$ is the inner product corresponding to $\tilde{S}$ on V. Thus $\langle w, S\Pi v\rangle = \langle w, \Pi^t S\Pi v\rangle$ for all v, implying $S\Pi = \Pi^t S\Pi$ and $SH = \Pi^t \mathbf{X}$. Let $[\cdot|\cdot]_A$ be the inner product defined on V by $\tilde{A}$. The map A from V to $\mathbf{X}$ is such that $\langle u, Av\rangle = [u \,|\, v]_A$ for all pairs (u, v) of elements of V. By the same argument there is also a map A_H from H to the space $H' = SH = \Pi^t \mathbf{X}$ such that $\langle u, A_H v\rangle = [u \,|\, v]_A$ for all pairs (u,v) of elements of H. This gives $\langle u, A_H v\rangle = \langle u, Av\rangle$ for all such pairs. Equivalently $\langle \Pi u, A_H v\rangle = \langle \Pi u, Av\rangle$ for all $u \in V$ and $v \in H$. Therefore $A_H v = \Pi^t Av$. Defining B_H in a similar manner, we get a difference $D_H = \frac{1}{2}(A_H - B_H)$.

(Note that AH need not be in SH = H′. An example is given by the matrices $A = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$ on the plane). Considering P_H and Q_H as measures on $H' = SH$ the formula derived above shows that $[\rho(P_H, Q_H)]^4 = \det[I - (S^{-1}D_H)^2] \geq \rho^4(P, Q) = \beta$, say. Let m be the dimension of H and let $\lambda_1, \lambda_2, \ldots, \lambda_m$ be the eigenvalues of $S^{-1}D_H$. Since $\prod_{j=1}^{m}(1 - \lambda_j^2) \geq \beta$ and since $|\lambda_j| \leq 1$ one must have $0 \leq \lambda_j^2 \leq 1 - \beta$ and $\sum_{j=1}^{m} \lambda_j^2 \leq -\log\beta$. However $\sum_{j=1}^{m} \lambda_j^2 = \sum_{j=1}^{m} \| S^{-1}D_H u_j \|^2$ for any orthogonal base $\{u_j ; j = 1, \ldots, m\}$ of H. since S^{-1} preserves the norms, this also means that $\sum_{j=1}^{m} \| D_H u_j \|^2 \leq -\log\beta$.

This is true for any H and any orthogonal sequence $\{u_1, u_2, \ldots, u_m, \ldots\}$ in V yielding $\sum_{j=1}^{\infty} \| D u_j \|^2 \leq -\log\beta$. Thus, if P and Q are not disjoint, D and $S^{-1}D$ are Hilbert-Schmidt operators and $(S^{-1}D)^2$ is an operator with finite trace.

A consequence of this is that one can find in V a basis $\{u_j ; j \in J\}$ that is orthogonal for the norm $\tilde{S}^{1/2}$ and also for the norms $\tilde{A}^{1/2}$ and $\tilde{B}^{1/2}$. To obtain it, let u_1 be such that $\tilde{A}(u_1)$ is maximized subject to $\| u_1 \| \leq 1$. This is the same problem as maximizing $\frac{1}{2}[\tilde{A}(u) - \tilde{B}(u)]$ subject to $\frac{1}{2}[\tilde{A}(u) + \tilde{B}(u)] \leq 1$. Equivalently again, subject to the same condition, we are to maximize $< u, Du >$. Since D is a compact operator, there does exist a u_1 that achieves the maximum. If $u_1, \ldots, u_n$ have been determined, one selects u_{n+1} to maximize $\tilde{A}(u)$ subject to $\| u \| \leq 1$ among those u's that are orthogonal to $u_1, \ldots, u_n$. For this basis $S^{-1}D$ is represented by a diagonal matrix and there is no difficulty in writing the determinant of $I - (S^{-1}D)^2$ as a product $\prod_j \{(1 - \lambda_j^2) ; j \in J\} = \rho^4(P, Q)$.

A corollary of the above is that if $\rho(P, Q) > 0$ then P and Q are mutually absolutely continuous. From this it is easy to derive the Hájek-Feldman theorem: Two Gaussian measures are either mutually absolutely continuous or disjoint.

References

1. Abramowitz, M. and Stegun, I.A. (1964). *Handbook of Mathematical Functions* National Bureau of Standard Applied Mathematics Series. Vol 55.

2. Bahadur, R.R. (1967). ‘‘An optimal property of the likelihood ratio statistic.’’ *Proc. 5th Berkeley Symp. Math. Stat. Probab.* **1**, pp. 13-26.

3. Birgé, L. (1983). ‘‘Approximation dans les espaces métriques et théorie de l’estimation.’’ *Z. Wahrsch. verw. Gebiete* **65**, p. 181-237.

4. Dieudonné, J. (1971). *Elements d’analyse*, vol 4. Gauthier-Villars, Paris 411 pages.

5. Gromov, M.L. and Rokhlin, V.A. (1970). ‘‘Embeddings and immersions in Riemannian Geometry.’’ *Russian Math. Surveys vol. 25*, #5, pages 1-57.

6. Hauck, W.W. and Donner, A. (1977). ‘‘Wald’s test as applied to hypotheses in logit analysis.’’ *J. Amer. Statistical Assoc.* **72**, p. 851-853. Corrigendum (1980) **75**, p. 482.

7. Kraft, C.H. (1955). ‘‘Some conditions for consistency and uniform consistency of statistical procedures’’. *Univ. California Pub. Statistics* vol. 2 #6 pp. 125-142.

8. Le Cam, L. (1956). ‘‘On the asymptotic theory of estimation and testing hypotheses.’’ *Proc. 3rd Berkeley Symp. Math. Stat. and Prob.* **1**, pp. 129-156.

9. Le Cam, L. (1960). ‘‘Locally asymptotically normal families of distributions.’’ *Univ. of California Publ in Statistics* **3**, pp. 37-98.

10. Le Cam, L. (1964). ‘‘Sufficiency and approximate sufficiency.’’ *Ann. Math. Stat.* **35**, pp.1419-1455.

11. Le Cam, L. (1975). ‘‘On local and global properties in the theory of asymptotic normality of experiments’’. *Stochastic Processes and Related Topics* M.L. Puri editor - Academic Press pp. 13-54.

12. Le Cam, L. (1977). ‘‘On the asymptotic normality of estimates.’’ Proc. of the Symposium to honor J. Neyman. *Pánstw. Wydawn. Nauk.* Warsaw pp. 203-217.

13. Le Cam, L. (1979). Maximum likelihood. An introduction. *Lectures notes #18. Univ. Maryland*.

14. Le Cam, L. (1985). ‘‘Sur l’approximation de familles de mesures par des familles gaussiennes.’’ *Ann. Inst. H. Poincaré* **21**, pp. 225-287.

15. Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag. xxvi + 742 pages.

16. Le Cam, L. and Yang, G.L. (1988). ‘‘On the preservation of local asymptotic normality under information loss.’’ *Ann. Statist.* **16**, No. 2, 483-520.

17. Lehmann, E.L. (1986). *Testing Statistical Hypotheses*. 2nd Edition. Wiley xx + 600 pages.

18. Mammen, E. (1987). ‘‘Local optimal Gaussian approximation of an exponential family’’. *Probab. Theory Related Fields* **76**, #1, pp. 103-119.

19. Mantel N. (1987). ‘‘Understanding Wald’s test for exponential families.’’ *The American Statistician* **41**, pp. 147-149.

20. Matusita, K. (1955). ‘‘Decision rules based on the distance, for problems of fit, two-samples and estimation.’’ *Ann. Math. Stat.* **26**, 613-640.

21. Matusita, K. (1961). ‘‘Interval estimation based on the notion of affinity.’’ *Bull. Internat. Statist. Inst.* **38**, part 4, 241-244.

22. Matusita, K. (1967). ‘‘Classification based on distance in multivariate Gaussian cases.’’ *Proc. 5th. Berkeley Symp. Math. Stat. Probab. vol 1*, 299-304.

23. Neyman, J. and Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* **20** Part I, 175-240 Part II, 263-294.

24. Schwartz, J.T. (1969). *Nonlinear functional analysis*. Gordon and Breach N.Y. 236 pages.

25. Strasser, H. (1985). *Mathematical Theory of Statistics*. Walter de Gruyter xii + 492 pages.

26. Vaeth, (1985). ‘‘On the use of Wald’s test in exponential families.’’ *Inter. Statistical Review.* **53**, pp. 199-214.

27. Wald, A. (1943). ‘‘Tests of statistical hypotheses concerning several parameters when the number of observations is large.’’ *Trans. Amer. Math. Soc.* **54**, pp. 426-482.

28. Wilks, S.S. (1938). ‘‘Shortest average confidence intervals from large samples.’’ *Ann. Math. Stat.* **9**, 166-175.