The Annals of Statistics
1988, Vol. 16, No. 1, 342–355

ON THE CHOICE OF A MODEL TO FIT DATA FROM AN EXPONENTIAL FAMILY

By Dominique M. A. Haughton

Temple University and University of Lowell

Let $X_1, \dots, X_n$ be iid observations coming from an exponential family. The problem of interest is this: Given a finite number of models m_j (smoothly curved manifolds in $\mathbb{R}^k$), choose the best model to fit the observations, with some penalty for choosing models with dimensions which are too large. A result of Schwarz is made more specific and is extended to the case where the models are curved manifolds. If $S(Y, n, j)$ is—up to a constant $C(n)$ independent of the model—the log of the posterior probability of the jth model, where the sample mean $Y_n = (1/n)\Sigma_{i=1}^n X_i$ has been replaced by Y, Schwarz suggested an asymptotic expansion of $S(Y, n, j)$ whose leading terms are $\gamma(Y, n, j) = n \sup_{\psi \in m_j \cap \Theta}(Y\psi - b(\psi)) - \frac{1}{2}k_j \log n$, in the case where the models are affine subspaces of $\mathbb{R}^k$. We establish a similar asymptotic expansion, including the next term, with uniform bounds for Y in a compact neighborhood of $\nabla b(\theta)$, where θ is the true value of the parameter. We suggest a criterion for the choice of the best model that consists of maximizing the three leading terms in the expansion $S(Y, n, j)$. We show that the criterion gives the correct model with probabilities $P_\theta^n \to 1$ as $n \to +\infty$.

0. Introduction. This article is concerned with the problem of choosing, among a finite number of possibly curved models (manifolds in $\mathbb{R}^k$), the "best" model to fit iid observations X_i, $i = 1, 2, \dots$, whose law belongs to an exponential family.

In Section 1 (Proposition 1.2) we show that maximizing the quantities

$$\gamma(n, j) := \log M_j(X_1, \dots, X_n) - \tfrac{1}{2}k_j \log n$$

leads to a correct choice of a model with probabilities $P_\theta^n \to 1$ as $n \to +\infty$, where P_θ is the true law of the observations in the exponential family; here $M_j(X_1, \dots, X_n)$ is the maximum of the likelihood function of the n first observations on the jth model and k_j is the dimension of the jth model. It follows in particular from Proposition 1.2 that this procedure is consistent [see, for example, Woodroofe (1982)].

The main conclusions of this paper, which completes and extends a result of Schwarz (1978), are as follows:

1. For any model m_j where the true parameter θ is in $\operatorname{int}(m_j \cap \Theta)$, the quantities $S(n, j)$, where $S(n, j)$ is the log of the posterior probability of the jth model plus a constant $C(n)$ independent of the model, have an asymptotic expansion whose leading three terms $\Gamma(n, j)$ are given by Theorem 2.3.

Received March 1984; revised May 1986.
AMS 1980 *subject classifications.* Primary 62F12; secondary 62H99.
Key words and phrases. Curved models, Schwarz's criterion, exponential families.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

2. When choosing between two models m_1 and m_2, if θ belongs to m_1 and not to m_2, then the procedure that consists of maximizing the $\Gamma(n, j)$, $j = 1, 2$, will lead to the correct choice of a model (i.e., m_1 rather than m_2) with $P_\theta^n \to 1$ as $n \to +\infty$, even if $k_1 \geq k_2$ (Proposition 1.2 and Remark 2.3).
3. When choosing between two models m_1 and m_2 where the true value θ is in $m_1 \cap m_2$ and $k_1 \neq k_2$, with $P_\theta^n \to 1$ as $n \to +\infty$, the procedure that consists of maximizing the $\Gamma(n, j)$, $j = 1, 2$, will lead to the choice of the model with the smallest dimension and in this case, for any suitably smooth prior, a Bayes procedure will also lead to the same choice with $P_\theta^n \to 1$ as $n \to +\infty$ (Proposition 1.2, Remark 2.3 and Corollary 2.3).
4. When choosing between two models m_1 and m_2 with $k_1 = k_2$ and where the true value θ of the parameter belongs to $m_1 \cap m_2$, then by assumption $(**)$ of Section 1 we really have a choice of three models, m_1, m_2 and $m_1 \cap m_2$: say $m_1 \cap m_2 = m_{i_0}$ for some index i_0; then, the procedure based on the $\Gamma(n, j)$, $j = 1, 2, i_0$ will pick $m_1 \cap m_2$ and coincide with the Bayes procedure with $P_\theta^n \to 1$ as $n \to +\infty$.

Section 2.1 deals with the problem of existence and unicity of MLE's on a curved model [note that Amari (1982) gives a geometrical interpretation of maximum likelihood estimation but is not concerned with the existence problem]. Sections 2.2 and 2.3 are devoted to obtaining an asymptotic expansion for the log of the posterior probability of the jth model; this posterior probability is a function $P(Y_n)$ of the sample mean Y_n of the n first observations. We first obtain an asymptotic expansion for $P(Y)$ as a function of Y (Proposition 2.2) *which is uniform in* Y, *for* Y *in some compact set.* The expansion has a precision of $n^{-(k_j+1)/2}$. In Section 2.3 we apply Proposition 2.2 to the case where Y is replaced by Y_n—this is where we need the uniformity in Y in Proposition 2.2—and obtain the desired asymptotic expansion for the log of the posterior probability of the jth model with a precision of $n^{-1/2}$ in probability.

Section 3 is concerned with the choice of a degree for a polynomial regression. One major issue is that the observations are not iid. We show that, by assuming that *both variables* in the regression are random variables, we can still apply the results of Sections 1 and 2. We describe the different models, which are curved.

To conclude this section, we note that the quantities $\gamma(n, j)$ introduced by Schwarz (1978) arise very naturally as the leading terms of the asymptotic expansion in Section 2.3 and that maximizing the $\gamma(n, j)$ [commonly called the Bayesian information criterion (BIC)] is a consistent procedure. Another well-known procedure is the Akaike information criterion (AIC) [Akaike (1974)]. It has been shown that the AIC is not consistent as $n \to +\infty$ [e.g., Woodroofe (1982) and Hannan (1980)]. The point has been made, though, that inconsistency may not be of great consequence from the point of view of prediction [Geisser and Eddy (1979)]. The AIC seems to have optimality properties in cases such as the selection of the order of the model for estimating parameters of a linear process, the key assumption being that the dimension of the models is allowed to increase with sample size [Shibata (1980, 1981)].

1. Criterion for the correct choice of a model when the models are C^∞ manifolds in $\mathbb{R}^k$.

1.1. *The Bayes procedure. Statement of the problem.* Let X_i, $i = 1, 2, \dots$, be iid observations from an exponential family in standard form with densities $f(X, \phi) = \exp(X\phi - b(\phi))$, with respect to a finite measure on $\mathbb{R}^k$, with $\phi \in \Theta$ the natural parameter space [Lehmann (1959), page 51]. We assume that we have a finite number of competing models $m_j \cap \Theta$ where m_j is a C^∞ k_j-dimensional connected manifold embedded in $\mathbb{R}^k$ [for terminology and basic facts in differential geometry we refer to Spivak (1979)]. An important special case is the case where m_j is a k_j-dimensional affine space in $\mathbb{R}^k$ [Schwarz (1978)]. We assume that

$(*)$ for each $i \neq j$, if a point in the closure of m_i is in $m_j \cap \operatorname{int} \Theta$, then it is in m_i.

We will also assume that

$(**)$ if $k_i = k_j$ for some pair (i, j), $i \neq j$, and if $m_i \cap m_j \neq \varnothing$, then $m_i \cap m_j$ is also an available model m_r and is of lower dimension.

On each of these m_j a natural analogue λ_j of Lebesgue measure is defined. If m_j is a k_j-dimensional affine space in $\mathbb{R}^k$, the defined measure will reduce to the Lebesgue measure. If m_j is a one-dimensional curve (or a two-dimensional surface) in $\mathbb{R}^k$, this measure will be the usual arc length (or the usual surface element). Note that the standard inner product of $\mathbb{R}^k$ induces a natural C^∞ Riemannian structure on m_j. Let g_{ij} be the coefficients of this Riemannian metric on a coordinate neighborhood U of a point p of m_j.

We define *the standard volume element* dV on the Riemannian manifold m_j by

$$dV = (\det g_{ij})^{1/2} |dX^1(p) \wedge \cdots \wedge dX^{k_j}(p)|$$

in a coordinate system (X, U) about the point p [see Spivak (1979), pages 417–418 for details]. This definition does not depend on the coordinate system [Spivak (1979), page 281]. The set function $\lambda_j(A) = \int_A dV$ is then defined for any Borel subset A of m_j and is countably additive.

We will assume that the conditional prior distribution μ_j of the parameter ϕ given the jth model has a density f_j with respect to λ_j which is a nowhere zero C^∞ function on $m_j \cap \Theta$ (this assumption will also hold for measures obtained from volume elements of Riemannian metrics smoothly related to the original one). Let α_j be the prior probability of the jth model. The prior distribution of ϕ is then $\mu = \sum \alpha_j \mu_j$. Note that μ is concentrated on $\cup m_j$ and that the μ_j are mutually orthogonal, clearly if two m_j are of different dimensions, and by assumption $(**)$ if they have the same dimension.

Let $P(n, j)$ be the posterior probability of the jth model given the prior and the n first observations $X_1, \dots, X_n$. We have

$$P(n, j)$$
$$= \alpha_j \int_{m_j \cap \Theta} \exp\left(\phi \sum_{i=1}^{n} X_i - nb(\phi)\right) d\mu_j(\phi)\bigg/ \int_{\Theta} \exp\left(\phi \sum_{i=1}^{n} X_i - nb(\phi)\right) d\mu(\phi).$$

Then a Bayes choice of a model is a choice that maximizes $P(n, j)$.

Let $Y_n = (1/n)\sum_{i=1}^{n} X_i$ and

$$S(n, j) = \log \alpha_j + \log \int_{m_j \cap \Theta} \exp n(\phi Y_n - b(\phi))\, d\mu_j(\phi)$$
$$= \log P(n, j) + C(n)$$

for some $C(n)$. Let θ be the true value of the parameter. We assume throughout that $\theta \in m_j$ for some j (this is no loss of generality since we can always adjoin $\mathbb{R}^k$ as a model).

DEFINITION. The correct choice between models is the model of lowest dimension which contains θ.

1.2. *Criterion for the correct choice of a model.* Using the notation of Section 1.1, let

$$\gamma(n, j) = n \sup_{\phi \in m_j \cap \Theta} (Y_n \phi - b(\phi)) - \tfrac{1}{2} k_j \log n$$

[Schwarz (1978)]. The following proposition proves the consistency of Schwarz's criterion.

PROPOSITION 1.2. *Assume $\theta \in \operatorname{int} \Theta$ and let m_1 and m_2 be two different models. If $\theta \in m_2 \setminus m_1$, or if $\theta \in m_1 \cap m_2$ with $k_2 < k_1$, then*

$$\lim_{n \to +\infty} P_\theta^n(\gamma(n,1) < \gamma(n,2)) = 1.$$

PROOF. Let $f(\phi) = \nabla b(\theta)\phi - b(\phi)$ for $\phi \in \Theta$. The function f attains its unique maximum at θ [Barndorff-Nielsen (1978), Theorems 9.13 and 9.1 and (1), page 141]. Let $\theta \in m_2 \setminus m_1$. Since $\theta \notin \overline{m}_1$ by assumption, let us pick $\varepsilon > 0$ and a neighborhood N of θ such that

$$N \cap m_1 = \varnothing$$

and, for $\phi \notin N$,

$$\nabla b(\theta)\phi - b(\phi) + \varepsilon \le \nabla b(\theta)\theta - b(\theta).$$

We have

$$(*) \qquad \sup_{\phi \in m_1 \cap \Theta} \nabla b(\theta)\phi - b(\phi) + \varepsilon \le \nabla b(\theta)\theta - b(\theta).$$

Since, by the strong law of large numbers, $Y_n \to \nabla b(\theta)$ with $P_\theta^\infty = 1$ as $n \to +\infty$

(note that $E_\theta X_1 = \nabla b(\theta)$ [Barndorff-Nielsen (1978), page 114]),

$$\sup_{\phi \in m_i \cap \Theta} (Y_n\phi - b(\phi)) \to \sup_{\phi \in m_i \cap \Theta} \nabla b(\theta)\phi - b(\phi) \quad \text{with } P_\theta^\infty = 1$$

as $n \to +\infty$, by continuity of the function $Y \to \sup_{\phi \in m_i \cap \Theta} Y\phi - b(\phi)$ which follows from its convexity, $i = 1,2$. (Note that

$$\nabla b(\theta) \in \operatorname{int}\left\{ Y \in \mathbb{R}^k \mid \sup_{\phi \in \Theta} Y\phi - b(\phi) < +\infty \right\}$$

[Barndorff-Nielsen (1978), page 151].) So with probabilities $P_\theta^n \to 1$ as $n \to +\infty$, we have

$$(**) \qquad \left| \sup_{\phi \in m_i \cap \Theta} (Y_n\phi - b(\phi)) - \sup_{\phi \in m_i \cap \Theta} (\nabla b(\theta)\phi - b(\phi)) \right| < \varepsilon/4, \qquad i = 1,2.$$

Using $(*)$ and $(**)$, with $P_\theta^n \to 1$ as $n \to +\infty$, we have

$$\sup_{\phi \in m_1 \cap \Theta} (Y_n\phi - b(\phi)) + \varepsilon/2 < \sup_{\phi \in m_2 \cap \Theta} (Y_n\phi - b(\phi)),$$

which completes the proof of the first part of the proposition.

Let $\theta \in m_1 \cap m_2$ and $k_2 < k_1$. We put $S_{n,i} = \sup_{\phi \in m_i \cap \Theta} Y_n\phi - b(\phi)$, $i = 1,2$. To prove the proposition, it is enough to show that $|S_{n,1} - S_{n,2}| = O_p(1/n)$. Since $\nabla b(\operatorname{int}\Theta)$ is open, with probabilities $P_\theta^n \to 1$ as $n \to +\infty$, there exists a unique MLE $\hat{\theta}_n$ that satisfies

$$\sup_{\phi \in \Theta} Y_n\phi - b(\phi) = Y_n\hat{\theta}_n - b(\hat{\theta}_n)$$

and $Y_n = \nabla b(\hat{\theta}_n)$ [Barndorff-Nielsen (1978), Theorem 9.13, page 151]. Let $U_n = Y_n\theta - b(\theta)$ and $\hat{U}_n = Y_n\hat{\theta}_n - b(\hat{\theta}_n)$. Since $\theta \in m_1 \cap m_2$, $0 \le S_{n,i} - U_n \le \hat{U}_n - U_n$, $i = 1,2$. Now $\hat{U}_n - U_n = Y_n(\hat{\theta}_n - \theta) + b(\theta) - b(\hat{\theta}_n)$ and $b(\hat{\theta}_n) - b(\theta) = (\hat{\theta}_n - \theta)\nabla b(\theta) + O_p(1/n)$ as obtained by writing a Taylor formula for b about θ and from the efficiency of the MLE $\hat{\theta}_n$ [Huber (1967)]. So $\hat{U}_n - U_n = (Y_n - \nabla b(\theta))(\hat{\theta}_n - \theta) + O_p(1/n)$. By the CLT, $Y_n - \nabla b(\theta) = O_p(1/\sqrt{n})$, and by efficiency of $\hat{\theta}_n$, $\|\hat{\theta}_n - \theta\| = O_p(1/\sqrt{n})$, so $\hat{U}_n - U_n = O_p(1/n)$. □

REMARK 1.2. Proposition 1.2 still holds for any sequence a_n of positive real numbers in place of $\log n$ such that $a_n/n \to 0$ as $n \to +\infty$ and $a_n \to +\infty$ as $n \to +\infty$.

The aim of the following section is to establish an asymptotic expansion for the $S(n, j)$ and show the role of $\log n$.

2. Asymptotic expansion of the $S(n,j)$.

2.1. *Study of the map $\phi \to Y\phi - b(\phi)$ on $m_j \cap \Theta$ when Y is in a neighborhood of $\nabla b(0)$ and $0 \in m_j \cap \operatorname{int}\Theta$.* We note that, by a translation of the parameter, we can assume in this section that $\theta = 0$.

Proposition 2.1. *There exists a neighborhood W of $\nabla b(0)$ in $\mathbb{R}^k$ such that, if $Y \in W$, the map $\phi \to Y\phi - b(\phi)$ attains its maximum on $m_j \cap \Theta$ at a unique point $\bar{\theta}_Y$.*

The idea of the proof is to write the function $\phi \to Y\phi - b(\phi)$ in local coordinates near 0 and apply the implicit function theorem.

Proof. We consider for $\varepsilon > 0$ the following neighborhoods of 0 in Θ:

$$N_{\varepsilon, Y} = \{\phi \in \Theta / Y\phi - b(\phi) > -b(0) - \varepsilon\}, \qquad M_\varepsilon = N_{\varepsilon, \nabla b(0)}.$$

Then

$$\sup_{\phi \in \overline{N}_{\varepsilon, Y} \cap m_j \cap \Theta} (Y\phi - b(\phi)) = \sup_{\phi \in m_j \cap \Theta} (Y\phi - b(\phi)).$$

The function $\phi \to \nabla b(0)\phi - b(\phi)$ attains its unique maximum on int Θ at 0, and if $Y \in \nabla b(\text{int}\, \Theta)$, $\phi \to Y\phi - b(\phi)$ attains its unique maximum at some $\hat{\theta}$ in int Θ [Barndorff-Nielsen (1978), Theorem 9.13].

Lemma 2.1.1 follows easily from remarks on level sets [Barndorff-Nielsen (1978), page 150].

Lemma 2.1.1. *There exist a compact set $K \subset \mathbb{R}^k$ and a constant C such that, if $\|Y - \nabla b(0)\| \le C$, $M_\varepsilon \subset K$ and $N_{\varepsilon, Y} \subset K$ for $0 < \varepsilon \le 1$.*

We put $M = m_j$ and $m = k_j = \dim m_j$, and we choose coordinates in $\mathbb{R}^k$ and a coordinate neighborhood $M \cap V$ of 0 on M such that $M \cap V = \{X, y_{m+1}(X), \dots, y_k(X);\ X = (x_1, \dots, x_m) \in U\}$ for some neighborhood U of 0 in R^m with $|y_{m+l}(X)| \le D\|X\|^2$, $\forall\ X \in U$, $l = 1, \dots, k - m$ for some D [cf. Guillemin and Pollack (1974), page 19]. It is easy to prove that we can pick $\varepsilon \le 1$ small enough so that $M \cap M_{2\varepsilon} \subset M \cap V$ and $\delta_\varepsilon > 0$ such that if $\|Y - \nabla b(0)\| < \delta_\varepsilon$, $N_{\varepsilon, Y} \subset M_{2\varepsilon}$ (the existence of such a δ_ε follows easily from Lemma 2.1.1). In our choice of coordinates near 0, 0 has coordinates 0 in R^m and any $\phi \in M_{2\varepsilon} \cap M$, thus any $\phi \in N_{\varepsilon, Y} \cap M$ with $\|Y - \nabla b(0)\| < \delta_\varepsilon$ can be written $\phi(X) = X + O(\|X\|^2)$, as $X \to 0$, where $X = (x_1, \dots, x_m, 0, \dots, 0)$.

We would like to evaluate the function $F(X) = Y\phi(X) - b(\phi(X))$ in a neighborhood of 0. We will need the following lemma, which follows from a Taylor formula with integral remainder.

Lemma 2.1.2. *Let k be a positive integer and $f(X) = O(\|X\|^k)$ denote a C^∞ function of X in a neighborhood U of 0 in R^d such that $f(X)/\|X\|^k$ is bounded in $U \setminus \{0\}$. Then*

$$\frac{\partial f}{\partial x_l}(X) = O(\|X\|^{k-1}), \qquad l = 1, \dots, d, \text{ as } X \to 0.$$

Put

$$Q_{ij} = \frac{1}{2} \frac{\partial^2 b}{\partial \theta_i\, \partial \theta_j}(0)$$

and

$$Q(\phi) = \sum_{i, j=1}^{k} Q_{ij}\phi_i\phi_j.$$

Then

$$F(X) = YX + YO(\|X\|^2) - b(0) - \nabla b(0)(X + O(\|X\|^2))$$
$$-Q(X + O(\|X\|^2)) + O(\|X + O(\|X\|^2)\|^3)$$

near $X = 0$.

Using Lemma 2.1.2 we get, for $i = 1, \ldots, m$,

$$\frac{\partial F}{\partial x_i}(X) = -2\sum_{j=1}^{m} Q_{ij}x_j + (Y - \nabla b(0))O^i(\|X\|) + \left(Y_i - \frac{\partial b}{\partial \theta_i}(0)\right) + O(\|X\|^2),$$

where $O^i(\|X\|) = \partial/\partial x_i(O(\|X\|^2))$ and is an $O(\|X\|)$ by Lemma 2.1.2. Therefore,

$$\frac{\partial^2 F}{\partial x_i\, \partial x_j}(0) = -2Q_{ij} + (Y - \nabla b(0))O^j(1) + O(\|X\|).$$

An application of the implicit function theorem [e.g., Dieudonné (1972), page 277] then shows that, for Y in a neighborhood of $\nabla b(0)$, the equation $\nabla F(X) = 0$ has a unique C^∞ solution ξ_Y in a neighborhood of 0 (note that Q is positive definite). Note also that $\sup_{\phi \in \bar{N}_{\varepsilon, Y} \cap M \cap \Theta} Y\phi - b(\phi)$ is attained at a point $\bar{\theta}_Y$ in $N_{\varepsilon, Y} \cap M \cap \Theta$. So for ε small enough, $\bar{\theta}_Y$ must satisfy $\nabla F(\phi^{-1}(\bar{\theta}_Y)) = 0$, so by the preceding, for ε small enough and $\|Y - \nabla b(0)\| < \delta_\varepsilon$, it is unique. This completes the proof of Proposition 2.1. □

2.2. *Asymptotic behavior of the integrals* $J_n = \int_{m_j \cap \Theta} \exp(n(Y\phi - b(\phi)))\, d\mu_j(\phi)$, *as* $n \to +\infty$ *uniformly in* Y *for* $\|Y - \nabla b(0)\| \le \sigma$, *for some* $\sigma > 0$. This calculation will be an example of "Laplace's method" for multidimensional integrals [see Hsu (1948, 1951) and Skinner (1980)]. We will use the notations of Section 2.1. As in Section 2.1 we assume that $\theta = 0$. We have the following proposition:

PROPOSITION 2.2. *Assume that* $0 \in m_j \cap \operatorname{int} \Theta$ *and that the density* f_j *of* μ_j *on* $m_j \cap \Theta$ *is* C^∞ *and nowhere vanishing on* $m_j \cap \Theta$. *Then there exists a positive number* σ *such that on the compact set* $\{\|Y - \nabla b(0)\| \le \sigma\}$, *uniformly in* Y,

$$J_n = e^{n(Y\bar{\theta}_Y - b(\bar{\theta}_Y))}\left\{\left(\frac{2\pi}{n}\right)^{k_j/2} \frac{f_j(\bar{\theta}_Y)(\det g_{ij}(\xi_Y))^{1/2}}{\{\det(-\partial^2 F/\partial x_i\, \partial x_j(\xi_Y))\}^{1/2}} + O(n^{-(k_j+1)/2})\right\}.$$

The idea of the proof is to write the integral J_n in terms of the μ_j-measure of a neighborhood $N'_{\varepsilon, Y}$ of $\bar{\theta}_Y$, and then to estimate $\mu_j(N'_{\varepsilon, Y})$ by noticing that $N'_{\varepsilon, Y}$ lies between two ellipsoids, and estimating the volume of these ellipsoids.

PROOF. By using Proposition 2.1 and its proof we choose $\sigma' > 0$ small enough so that $\bar{\theta}_Y$ exists and equals $\phi(\xi_Y)$ (see Section 2.1) for $\|Y - \nabla b(0)\| \leq \sigma'$. We put $f(\phi) = \exp(Y\phi - b(\phi) - (Y\bar{\theta}_Y - b(\bar{\theta}_Y)))$ and $I_n = \int_{m_j \cap \Theta} f^n \, d\mu_j(\phi) = E_{\mu_j} f^n$. Note that $0 < f \leq 1$. Let $g = 1 - f$; then $0 \leq g < 1$. If $G(t) = \mu_j(g \leq t)$ is the distribution function of g, then, integrating by parts,

$$Ef^n = \int_0^1 (1 - t)^n \, dG(t) = n \int_0^1 (1 - t)^{n-1} G(t) \, dt$$

and $G(t) = \mu_j(f \geq 1 - t) = \mu_j(N'_{\varepsilon, Y})$ where $\varepsilon = -\log(1 - t)$ and $N'_{\varepsilon, Y} = \{\phi | Y\phi - b(\phi) > Y\bar{\theta}_Y - b(\bar{\theta}_Y) - \varepsilon\}$. It is easy to check that for ε small enough, say $\varepsilon \leq \varepsilon_0$, there exists $\theta_\varepsilon > 0$ such that if $\|Y - \nabla b(0)\| \leq \theta_\varepsilon$, $N'_{\varepsilon, Y} \subset M_{4\varepsilon}$ and $M_{4\varepsilon}$ is included in a coordinate neighborhood on m_j near 0 as in Section 2.1.

We wish to estimate the $\mu_j(N'_{\alpha, Y})$ for α small enough and $\|Y - \nabla b(0)\| \leq \min(\theta_\alpha, \sigma')$. By the preceding and Section 2.1, $\phi \in N'_{\alpha, Y}$ can be written $\phi(X) = X + O(\|X\|^2)$. As in Section 2.1, we define $F(X) = F(X, Y) = Y\phi(X) - b(\phi(X))$. Then

$$\mu_j(N'_{\alpha, Y}) = \int_{F(X) - F(\xi_Y) \geq -\alpha} f_j(\phi(X)) \big(\det g_{ij}(X)\big)^{1/2} \, dX,$$

where f_j is the density of μ_j on $m_j \cap \Theta$ and g_{ij} are the coefficients of the Riemannian structure induced on m_j by the Euclidean structure of R^k, expressed in the chosen coordinate system on m_j. If A is the quadratic form defined by

$$A(V) = -\frac{1}{2} \sum_{i, j=1}^{m} \frac{\partial^2 F}{\partial x_i \, \partial x_j}(\xi_Y) V_i V_j,$$

where $m = \dim m_j$, we have $F(X) - F(\xi_Y) = -A(X - \xi_Y) + R(Y, X)$, where $R(Y, X)$ denotes the integral remainder in a Taylor expansion for F about ξ_Y. Note that A is positive definite for $\|Y - \nabla b(0)\| \leq \eta^*$ for some η^*. Let α_i be the positive eigenvalues of A, $i = 1, \dots, m$. Then $\min \alpha_i$ and $\max \alpha_i$ are continuous functions of Y since $\max \alpha_i = \|A\|$ and $\min \alpha_i = \|A^{-1}\|^{-1}$ where $\|A\| = \sup_{\|X\| \leq 1} \|AX\|$.

We define

$$\rho = \inf_{\|Y - \nabla b(0)\| \leq \eta^*} (\min \alpha_i). \tag{1}$$

Note that $\rho > 0$. We will use the following lemma, which is easily proved.

LEMMA 2.2.1. *There exists a constant K independent of Y such that $|R(Y, X)| \leq K \|X - \xi_Y\|^3$ for $X \in \phi^{-1}(M_{4\varepsilon_0})$ and $\|Y - \nabla b(0)\| \leq \eta^*$.*

Note that if Δ_ε is the diameter of $\phi^{-1}(M_{4\varepsilon})$, $\Delta_\varepsilon \to 0$ as $\varepsilon \to 0$. Therefore, we can pick $0 < \varepsilon_1 < \varepsilon_0$ small enough so that $K \Delta_{\varepsilon_1} < \rho/2$, hence $\rho - K \Delta_{\varepsilon_1} > \rho/2$, where K is as in Lemma 2.2.1. We take $\alpha \leq \varepsilon_1$ and $\|Y - \nabla b(0)\| < \min(\theta_{\varepsilon_1}, \eta^*, \sigma')$; then it is straightforward to show that $X \in \phi^{-1}(N'_{\alpha, Y}) \Rightarrow \|X - \xi_Y\| \leq \sqrt{2\alpha/\rho}$.

Clearly

$$\mu_j(N'_{\alpha,Y}) = \int_{A(X-\xi_Y)-R(Y,X)\le\alpha} H_j(X)\,dX,$$

where

$$H_j(X) = f_j(\phi(X))\left(\det g_{ij}(X)\right)^{1/2}.$$

With a few elementary arguments it can be shown that

$$\begin{aligned}(2)\qquad \int_{A(X-\xi_Y)+\partial_\alpha^*\|X-\xi_Y\|^2\le\alpha} H_j(X)\,dX &\le \int_{A(X-\xi_Y)-R(Y,X)\le\alpha} H_j(X)\,dX\\ &\le \int_{A(X-\xi_Y)-\partial_\alpha^*\|X-\xi_Y\|^2\le\alpha} H_j(X)\,dX,\end{aligned}$$

where $\partial_\alpha^* = K\sqrt{2\alpha/\rho}$.

Let $E = \{X \mid A(X-\xi_Y) + \partial_\alpha^*\|X-\xi_Y\|^2 \le \alpha\}$. We wish to estimate $\int_E H_j(X)\,dX$. For this we will need to estimate the volume of the ellipsoid E. To do this we will make the change of variable $X - \xi_Y = P(Z - \xi_Y)$ in R^m, where P is an orthogonal matrix such that P^tAP is diagonal and A also denotes the matrix of the positive definite form A. The Jacobian of the transformation is $|dX/dZ| = \det P = \pm 1$. We have $A(X-\xi_Y) = \sum_{i=1}^m \alpha_i(Z-\xi_Y)_i^2$ and $\|Z - \xi_Y\|^2 = \|X-\xi_Y\|^2$ since P is orthogonal. Then

$$\int_E dX = \int_{\sum_{i=1}^m(\alpha_i+\partial_\alpha^*)(Z-\xi_Y)_i^2\le\alpha} dZ.$$

Therefore,

$$\int_E dX = (\alpha\pi)^{m/2}\Big/\prod_{i=1}^m(\alpha_i+\partial_\alpha^*)^{1/2}\Gamma\left(\frac{m}{2}+1\right).$$

We note here that

$$\prod_{i=1}^m \alpha_i^{-1/2} = \left\{\det\left(-\frac{1}{2}\left(\frac{\partial^2 F}{\partial x_i\,\partial x_j}(\xi_Y)\right)\right)\right\}^{-1/2} = 2^{m/2}\left\{\det\left(-\frac{\partial^2 F}{\partial x_i\,\partial x_j}(\xi_Y)\right)\right\}^{-1/2}.$$

We expand the function $H_j(X)$ about ξ_Y: $H_j(X) = H_j(\xi_Y) + h_j(X-\xi_Y)$ with $|h_j(X-\xi_Y)| \le M\|X-\xi_Y\|$, for some M independent of Y, $X \in \phi^{-1}(M_{4\varepsilon_0})$ and $\|Y - \nabla b(0)\| \le \min(\theta_{\varepsilon_1}, \eta^*, \sigma')$. We have

$$\int_E H_j(X)\,dX = H_j(\xi_Y)\lambda(E) + \int_E h_j(X-\xi_Y)\,dX,$$

where λ is the Lebesgue measure on R^m, and $|\int_E h_j(X-\xi_Y)\,dX| \le M\int_E\|X - \xi_Y\|\,dX$. It is easy to show that if $A(X-\xi_Y) + \partial_\alpha^*\|X-\xi_Y\|^2 \le \alpha$, then

$$\|X-\xi_Y\| \le \frac{\sqrt{\alpha}}{\sqrt{\rho+\partial_\alpha^*}};$$

it follows that

$$\int_E H_j(X)\,dX = H_j(\xi_Y)(\alpha\pi)^{m/2}\Big/\left(\Gamma\left(\frac{m}{2}+1\right)\prod_{i=1}^{m}(\alpha_i+\partial_\alpha^*)^{1/2}\right)+H(\alpha),$$

with $|H(\alpha)| \le M\pi^{m/2}\alpha^{(m+1)/2}/(\Gamma(m/2+1)(\rho+\partial_\alpha^*)^{(m+1)/2})$. By the same reasoning we get

$$\int_{A(X-\xi_Y)-\partial_\alpha^*\|X-\xi_Y\|^2\le\alpha} H_j(X)\,dX$$

$$= H_j(\xi_Y)(\alpha\pi)^{m/2}\Big/\Gamma\left(\frac{m}{2}+1\right)\prod_{i=1}^{m}(\alpha_i-\partial_\alpha^*)^{1/2}+H^1(\alpha),$$

where

$$|H^1(\alpha)| \le M\pi^{m/2}\alpha^{(m+1)/2}\Big/\left(\Gamma\left(\frac{m}{2}+1\right)(\rho-\partial_\alpha^*)^{(m+1)/2}\right)$$

for α small enough so that $\partial_\alpha^* < \rho$, say for $\alpha \le \alpha_0$, with $\alpha_0 < \varepsilon_1$. Now

$$\prod_{i=1}^{m}(\alpha_i+\partial_\alpha^*)^{-1/2} = \prod_{i=1}^{m}\alpha_i^{-1/2}(1+O(\sqrt{\alpha})) \quad \text{as } \alpha\to 0,$$

uniformly in Y. Also $H(\alpha) = O(\alpha^{(m+1)/2})$ and $H^1(\alpha) = O(\alpha^{(m+1)/2})$ uniformly in Y for $\|Y-\nabla b(0)\| \le \min(\theta_{\alpha_0}, \eta^*, \sigma')$. Using these estimates, and inequality (2), we have

$$\int_{A(X-\xi_Y)-R(Y,X)\le\alpha} H_j(X)\,dX$$

$$= H_j(\xi_Y)(\alpha\pi)^{m/2}\Big/\left(\Gamma\left(\frac{m}{2}+1\right)\prod_{i=1}^{m}\alpha_i^{1/2}\right)+O(\alpha^{(m+1)/2})$$

uniformly in Y for $\|Y-\nabla b(0))\| \le \min(\theta_{\alpha_0}, \eta^*, \sigma')$. We have now shown the following lemma.

LEMMA 2.2.2. *There exist positive numbers σ and α_0 such that if $\|Y-\nabla b(0)\| \le \sigma$ and $\alpha \le \alpha_0$, then $\mu_j(N'_{\alpha,Y}) = C_m(Y)\alpha^{m/2}+\beta(\alpha)$ with*

$$C_m(Y)$$

$$= f_j(\phi(\xi_Y))(\det g_{ij}(\xi_Y))^{1/2}\left\{\det\left(-\frac{\partial^2 F}{\partial x_i\,\partial x_j}(\xi_Y)\right)\right\}^{-1/2}(2\pi)^{m/2}\Big/\Gamma\left(\frac{m}{2}+1\right)$$

and $|\beta(\alpha)| \le \alpha^{(m+1)/2}\beta^1(\alpha)$ where $\beta^1(\alpha)$ is bounded and independent of Y.

Proposition 2.2 now follows easily from Lemma 2.2.2 applied to $G(t) = \mu_j(N'_{\varepsilon,Y})$ with $\varepsilon = -\log(1-t)$, and known facts about Euler's beta and gamma functions [see Haughton (1983) for details]. □

2.3. *Asymptotic expansion of the* $S(n, j)$. The following theorem will show the special role of $a_n = \log n$ in Schwarz's criterion. We will use the notation and assumptions of the previous sections.

THEOREM 2.3. *Let the true* $\theta \in m_j \cap \operatorname{int} \Theta$. *If*

$$S(n, j) = \log \alpha_j + \log \int_{m_j \cap \Theta} \exp(Y_n \phi - b(\phi))\, d\mu_j(\phi)$$

with $Y_n = n^{-1}\sum_{i=1}^{n} X_i$, *if* $\bar{\theta}_n^j$ *is the unique point on* $m_j \cap \Theta$ *where the function* $Y_n \phi - b(\phi)$ *attains its maximum, defined with probabilities converging to* 1 *as* $n \to \infty$, *then*

$$S(n, j) = n \sup_{\phi \in m_j \cap \Theta} (Y_n \phi - b(\phi)) - \frac{1}{2} k_j \log\left(\frac{n}{2\pi}\right) + \log \alpha_j + \log f_j(\bar{\theta}_n^j)$$

$$- \frac{1}{2} \log \det\left(\frac{\partial^2 b}{\partial \phi_r\, \partial \phi_s} (\bar{\theta}_n^j) \right) + O_p(n^{-1/2}).$$

PROOF. We will need a few lemmas. Let $M = m_j$. We can assume that $\theta = 0$.

LEMMA 2.3.1. *If* g_{ij} *are the coefficients of the Riemannian structure induced on* M *by the Euclidean structure of* R^k, *corresponding to the coordinate neighborhood* $M \cap V$ *of Section* 2.1, *then* $g_{ij}(0) = \delta_{ij}$.

PROOF. An easy calculation shows that

$$g_{ij}(X) = \delta_{ij} + \sum_{l=1}^{k-m} \frac{\partial y_{m+l}}{\partial x_i}(X) \frac{\partial y_{m+l}}{\partial x_j}(X). \qquad \square$$

LEMMA 2.3.2. *Let* f *be a* C^∞ *function on a convex neighborhood* U *of* 0 *in* R^k. *Then there exists* $\sigma > 0$ *such that, if* $\|Y - \nabla b(0)\| \le \sigma$, *we have* $f(\xi_Y) = f(0) + R(Y)$, *where* $|R(Y)| \le C\|\xi_Y\|$ *for some constant* C *independent of* Y (*where* ξ_Y *is as defined in Section* 2.1).

PROOF. The proof is similar to the proof of Lemma 2.2.1 and is omitted.

LEMMA 2.3.3. *If* $Y_n = n^{-1}\sum_{i=1}^{n} X_i$, *then* $\|\xi_{Y_n}\| = O_p(n^{-1/2})$.

PROOF. The lemma follows from the central limit theorem.

Theorem 2.3 now follows easily from Proposition 2.2 and Lemmas 2.3.1 and 2.3.2. $\square$

We now give a proposition which will show that when choosing between models m_i and m_j such that $k_i \neq k_j$ and $\theta \in m_i \cap m_j \cap \operatorname{int} \Theta$, with probabilities $P_\theta^n \to 1$ as $n \to +\infty$, the Bayes choice and the choice based on the quantities

$$\Gamma(n, j) = n \sup_{\phi \in m_j \cap \Theta} Y_n \phi - b(\phi) - \frac{1}{2} k_j \log\left(\frac{n}{2\pi}\right)$$

$$+ \log \alpha_j + \log f_j(\bar{\theta}_n^j) - \frac{1}{2} \log\left(\det\left(\frac{\partial^2 b}{\partial \phi_r \, \partial \phi_s}\right)\right)$$

coincide.

COROLLARY 2.3. *If $\theta \in m_{j_1} \cap m_{j_2} \cap \operatorname{int} \Theta$ and $k_{j_1} \neq k_{j_2}$, then $P_\theta^n(S(n, j_1) > S(n, j_2)$ and $\Gamma(n, j_1) \leq \Gamma(n, j_2)) \to 0$ as $n \to +\infty$.*

The proof is straightforward and is left to the reader.

REMARK 2.3. Note that $\Gamma(n, j) = \gamma(n, j) + O_p(1) + O_p(n^{-1/2})$ so $\Gamma(n, j) = \gamma(n, j) + o_p(\log n)$, where $\gamma(n, j)$ was defined in Section 1.2. Proposition 1.2 therefore holds with $\gamma(n, j)$ replaced by $\Gamma(n, j)$.

3. Choice of degree in a polynomial regression. Let (x_i, y_i) be a set of data in $\mathbb{R}^2$ where $y_i = \sum_{j=0}^d a_j x_i^j + \varepsilon_i$ and the ε_i are iid $N(0, \varepsilon^2)$. If we consider x_i not as a random variable but as an "incidental parameter," given the "structural parameters" $(a_0, a_1, \ldots, a_d, d, \varepsilon)$, the law of the y_i is $N(m_i, \varepsilon^2)$ with $m_i = \sum_{j=0}^d a_j x_i^j$. Schwarz's criterion does not apply to the observations y_i since they are not iid. We will show that we can still apply a Schwarz criterion to this problem by considering x_i as a random variable. This will also show the necessity of considering "curved models." We will assume therefore that the ε_i are iid $N(0, \varepsilon^2)$, the x_i are iid $N(m, \tau^2)$ and that all the ε_i are independent of all the x_i. Let $\eta = (d, a_0, \ldots, a_d, \varepsilon^2, m, \tau^2)$. We assume that η has a prior law of the form $\sum \alpha_k \mu_k$ where α_k is the prior probability that $d = k$.

Now let $z_i = (x_i, y_i)$. The z_i are iid given η and their density is

$$f(x, y) = \frac{1}{2\pi\tau\varepsilon} \exp\left[-\frac{(x - m)^2}{2\tau^2} - \frac{\left(y - \sum_{j=0}^d a_j x^j\right)^2}{2\varepsilon^2}\right], \tag{3}$$

so $f(x, y) = \exp[\sum_{j=1}^{3d+2} \theta_j T_j(x, y) - b(\theta)]$, where the $T_i(x, y)$ are defined by

$$T_1(x, y) = -x^2, \qquad T_2(x, y) = -y^2, \qquad T_3(x, y) = y, \ldots,$$

$$T_{3+j}(x, y) = x^j y, \qquad j = 0, 1, \ldots, d, \qquad T_{4+d}(x, y) = -x, \tag{4}$$

$$T_{5+d}(x, y) = -x^3, \qquad T_{6+d}(x, y) = -x^4, \ldots, T_{3d+2}(x, y) = -x^{2d},$$

the θ_i are defined by

$$\begin{aligned}
&\theta_1 = 1/2\tau^2 + \left(2a_0a_2 + a_1^2\right)/2\varepsilon^2, \qquad \theta_2 = 1/2\varepsilon^2,\\
&\theta_3 = a_0/\varepsilon^2, \qquad \theta_4 = a_1/\varepsilon^2, \dots, \theta_{3+j} = a_j/\varepsilon^2, \qquad j = 0, \dots, d,\\
&\theta_{4+d} = a_0a_1/\varepsilon^2 - m/\tau^2, \qquad \theta_{5+d} = \sum_{j=0}^{3} a_ja_{3-j}/2\varepsilon^2,\\
&\theta_{6+d} = \sum_{j=0}^{4} a_ja_{4-j}/2\varepsilon^2, \dots, \theta_{2d+2} = \sum_{j=0}^{d} a_ja_{d-j}/2\varepsilon^2,\\
&\theta_{2d+3} = \sum_{j=1}^{d} a_ja_{d+1-j}/2\varepsilon^2, \dots, \theta_{3d+1} = \sum_{j=d-1}^{d} a_ja_{2d-1-j}/2\varepsilon^2,\\
&\theta_{3d+2} = a_d^2/2\varepsilon^2,
\end{aligned} \tag{5}$$

and where $b(\theta)$ is defined by normalization. The family of distributions with densities $f(x, y)$ with respect to the Lebesgue measure on $\mathbb{R}^2$ is an exponential family with natural parameter space

$$\Theta = \left\{\theta \in \mathbb{R}^{3d+2} \,|\, \int_{\mathbb{R}^2} \exp\left[\sum_{j=1}^{3d+2} \theta_j T_j(x, y)\right] dx\, dy < \infty\right\}.$$

Let us describe the different models: We define the *maximum model* m_d to be the set of $\theta \in \mathbb{R}^{3d+2}$ defined parametrically by (5) with $(a_0, \dots, a_d) \in \mathbb{R}^{d+1}$, $\varepsilon^2 > 0$, $\tau^2 > 0$. Then m_d is a closed manifold in $\mathbb{R}^{3d+2}$ of dimension $3d + 2 - (2d - 2) = d + 4$ [Spivak (1979), page 65, Proposition 12]. The other models, corresponding to $a_d = 0$, $a_d = a_{d-1} = 0, \dots, a_d = a_{d-1} = \cdots = a_1 = 0$ will be denoted by $m_{d-1}, m_{d-2}, \dots, m_0$. Note that $m_j = \{\theta \in m_d/\theta_{3+j+1} = \cdots = \theta_{3+d} = 0\}$, so m_j is a closed submanifold of m_d of dimension $j + 4$ in m_d. In particular, dim $m_1 = 5$ and dim $m_0 = 4$, as submanifolds of m_d. Note that all the m_j are curved, and that, assuming the appropriate regularity for the density of μ_k with respect to the volume element on m_k, the results of Section 2.3 apply.

Acknowledgment. I would like to thank Richard M. Dudley for help and guidance in preparing this paper.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical identification model. *IEEE Trans. Automat. Control* **19** 716–723.

AMARI, S. I. (1982). Differential geometry of curved exponential families—curvatures and information loss. *Ann. Statist.* **10** 357–385.

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

DIEUDONNÉ, J. (1972). *Eléments d'Analyse* **1**. Gauthier-Villars, Paris.

DOOB, J. L. (1953). *Stochastic Processes.* Wiley, New York.

FORTUS, R. (1979). Approximations to Bayesian sequential tests of composite hypotheses. *Ann. Statist.* **7** 579–591.

GEISSER, S. and EDDY, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** 153–160.

GUILLEMIN, V. and POLLACK, A. (1974). *Differential Topology*. Prentice-Hall, Englewood Cliffs, N.J.

HANNAN, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8** 1071–1081.

HAUGHTON, D. (1983). On the choice of a model to fit data from an exponential family. Ph.D. thesis, Massachusetts Institute of Technology.

HSU, L. C. (1948). A theorem on the asymptotic behavior of a multiple integral *Duke Math. J.* **15** 623–632.

HSU, L. C. (1951). On the asymptotic evaluation of a class of multiple integrals involving a parameter. *Amer. J. Math.* **73** 625–635.

HUBER, P. (1967). The behavior of maximum likelihood estimators under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. Univ. California Press.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.

ROBERTS, A. and VARBERG, D. (1973). *Convex Functions*. Academic, New York.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SCHWARZ, G. (1981). Applying asymptotic shapes to nonexponential families. *Ann. Statist.* **9** 461–464.

SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164.

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

SKINNER, L. A. (1980). Note on the asymptotic behavior of multidimensional Laplace integrals. *SIAM J. Math. Anal.* **11** 911–917.

SPIVAK, M. (1979). *A Comprehensive Introduction to Differential Geometry* **1**, 2nd ed. Publish or Perish, Berkeley, Calif.

WOODROOFE, M. (1982). On model selection and the arc-sine laws. *Ann. Statist.* **10** 1182–1194.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF LOWELL
LOWELL, MASSACHUSETTS 01854