The Annals of Statistics
1979, Vol. 7, No. 6, 1162–1178

EXPONENTIAL MODELS FOR DIRECTIONAL DATA[1]

BY RUDOLF BERAN

University of California, Berkeley

A rotationally invariant exponential model, which includes the Fisher-von Mises and Bingham distributions as special cases, is proposed for directional data in $R^p (p \geqslant 2)$. A new regression estimator for the model parameters is developed as a competitor to the maximum likelihood estimator. Both the new estimator and the MLE are asymptotically efficient at the postulated model and are robust under small departures from that model. Computationally, the regression estimator is much simpler since it requires no iterations or numerical integrations. Goodness-of-fit can be assessed by fitting nested special cases of the general model to the data.

1. Introduction. Two parametric models, the Fisher-von Mises and the Bingham distributions, play an important role in the statistical analysis of directional data. Let S_p denote the set of all unit (column) vectors in R^p, $p \geqslant 2$, and let μ be the rotation invariant measure on S_p, normalized so that $\mu(S_p) = 1$. The probability elements of the Fisher-von Mises and Bingham distributions are, respectively,

$$a_p(\kappa)\exp(\kappa\nu^T x)\mu(dx), \qquad x \in S_p \tag{1.1}$$

and

$$b_p(D)\exp(\operatorname{tr}[DR^T xx^T R])\mu(dx), \qquad x \in S_p. \tag{1.2}$$

Some restrictions must be imposed upon the possible values of the parameters so as to obtain a one-to-one parametrization of the densities. The usual conventions are: κ a nonnegative scalar, ν a unit vector in S_p, D a $p \times p$ diagonal matrix whose (p, p) element (say) vanishes, R a $p \times p$ orthogonal matrix. Because of its invariance under the mapping $x \to -x$, the Bingham distribution (1.2) is also a possible model for axial data. Both the Fisher-von Mises and Bingham distributions can be derived as conditional distributions, given $X \in S_p$, of suitable multivariate normal distributions in R^p (see Mardia (1972), (1975) for details and a survey of the literature on directional statistics).

Geometrically, the Fisher-von Mises density is unimodal at $x = \nu$ and rotationally symmetric about the modal direction, unless $\kappa = 0$, in which event the distribution is uniform. The Bingham density is an antipodally symmetric density whose possible shapes, when $p = 3$, include a bipolar distribution, symmetric and asymmetric girdle distributions, and the uniform distribution. Thus, the distribu-

Received October 1977; revised August 1978.

[1] This research was supported in part by National Science Foundation Grant MCS 75-10376 and in part by a grant from the Science Research Council at the University of Leeds.

AMS 1970 *subject classifications.* Primary 62E10; secondary 62F10.

Key words and phrases. Directional data, axial data, rotational invariance, exponential family, asymptotically efficient estimator, robust estimator, density estimator, goodness-of-fit tests.

tions (1.1) or (1.2) often provide plausible parametric models for small sets of directional or axial data.

As sample size increases, so does the need to consider alternative models and to assess goodness-of-fit of the various models. A lack of tractable parametric models other than (1.1) or (1.2) has hindered practical efforts to satisfy these obligations. Prompted by these considerations, this paper proposes a general, rotationally invariant, exponential model for directional data and develops robust, asymptotically efficient estimators for the model parameters. Both the Fisher-von Mises and Bingham distributions are special cases of the model. Goodness-of-fit can be assessed by fitting nested special cases of the model to the data and carrying out a likelihood ratio or asymptotically equivalent test.

Let $C(S_p)$ denote the set of all real-valued, continuous functions whose domain is S_p. Let M be any subspace of $C(S_p)$ which is invariant in the following sense: $h(\cdot) \in M$ entails $h(g \cdot) \in M$ for every rotation g in R^p. As a general model for directional data, we propose the distribution whose density with respect to μ is

$$(1.3) \qquad f_h(x) = \exp[h(x) - d(h)]; \qquad h \in M, x \in S_p,$$

where $d(h)$ is chosen to make f_h integrate over S_p to one. The parameter space of this model is the subspace M. Invariance of M ensures rotational invariance of the model in the following sense: if X has a density belonging to the family (1.3), so does gX for every rotation g. This property is desirable since the choice of coordinate system for directional data is often arbitrary, selected mainly for observational convenience.

An invariant subspace is called irreducible if it does not contain any proper nontrivial invariant subspace. Let P_k be the subspace of $C(S_p)$ consisting of all homogeneous polynomials of degree k in the components of $x \in S_p$. Let M_k be the subspace of P_k which consists of all harmonic functions in P_k; i.e., $M_k = \{h \in P_k: \Delta h = 0\}$, Δ being the Laplacian in R^p. Then M_k is invariant, irreducible, and is of dimension $\binom{p+k-3}{k}[2k/(p-2)+1]$ if $p \geqslant 3$; of dimension 2 if $p = 2$ (see Dunkl and Ramirez (1971), page 109). Moreover, the family of disjoint subspaces $\{M_k: k \geqslant 0\}$ constitutes all possible nontrivial irreducible invariant subspaces of $C(S_p)$. If $h \in C(S_p)$ belongs to an invariant subspace M, h can be expressed in the form $h = \Sigma_{k \in I} h_k$, where $h_k \in M_k$ and $I \subset Z_+$, the set of all nonnegative integers. In other words, $M = \Sigma_{k \in I} \oplus M_k$.

Thus, in analyzing directional data, it is natural to consider the sequence of nested models with densities

$$(1.4) \qquad \exp[h(x) - d(h)]; \qquad h \in \Sigma_{k=1}^{r} \oplus M_k, \qquad r \geqslant 1$$

and for axial data, the sequence of antipodally symmetric densities

$$(1.5) \qquad \exp[h(x) - d(h)]; \qquad h \in \Sigma_{k=1}^{r} \oplus M_{2k}, \qquad r \geqslant 1.$$

Since each of the subspaces $\{M_k\}$ has a finite dimensional basis in $C(S_p)$, the models (1.4) and (1.5) can be written in canonical exponential family form:

$$f_\beta(x) = \exp[\beta^T v(x) - c(\beta)]; \qquad \beta \in R^q, x \in S_p, \tag{1.6}$$

where the $\{v_i : 1 \leqslant i \leqslant q\}$ are functions in $C(S_p)$ such that $\{1, v_1(x), \ldots, v_q(x)\}$ are linearly independent, and $v(x)$ is the vector $(v_1(x), v_2(x), \cdots, v_q(x))^T$ while $\beta = (\beta_1, \beta_2, \cdots, \beta_q)^T$. The parametrization (1.6) is useful for fitting the model to the data.

Finding basis functions $\{v_i(x)\}$ is simplified by the following results (Dunkl and Ramirez (1971), page 108): for every integer $k \geqslant 0$,

$$P_k = \Sigma_{j=0}^{[k/2]} \oplus M_{k-2j}. \tag{1.7}$$

Thus,

$$\begin{aligned} \Sigma_{k=1}^r \oplus M_k &= P_r + P_{r-1} - \operatorname{proj}_{M_0}(P_{r-1}) \quad \text{if } r \text{ is odd} \\ &= P_r + P_{r-1} - \operatorname{proj}_{M_0}(P_r) \quad \text{if } r \text{ is even,} \end{aligned} \tag{1.8}$$

and

$$\Sigma_{k=1}^r \oplus M_{2k} = P_{2r} - \operatorname{proj}_{M_0}(P_{2r}), \tag{1.9}$$

where $\operatorname{proj}_{M_0}(\cdot)$ denotes the projection into $M_0 = \operatorname{span}\{1\}$. Since any additive constant in the exponent can be absorbed into the normalizing factor, the densities (1.4) and (1.5) can be expressed in the canonical form (1.6) with an appropriate set of monomials as the basis $\{v_i(x)\}$. Specifically, a basis for model (1.5) consists of all distinct monomials of degree $2r$, excluding x_p^{2r}, say. (The last proviso is a consequence of the constraint $x^Tx = 1$ on every $x \in S_p$.) Similarly, a basis for model (1.4) consists of all distinct monomials of degrees r and $r - 1$, excluding x_p^r if r is even or x_p^{r-1} if r is odd.

When $r = 1$, the p monomials $\{x_i : 1 \leqslant i \leqslant p\}$ are a basis for model (1.4) while the $(p-1)(1+p/2)$ monomials $\{x_ix_j : 1 \leqslant i \leqslant j \leqslant p\} - \{x_p^2\}$ form a basis for model (1.5). Thus, the Fisher-von Mises and Bingham distributions are the simplest special cases of the general model (1.3), corresponding to the restrictions $h \in M_1$ and $h \in M_2$, respectively.

The rest of this paper is concerned with estimation of the parameter β in the general model (1.6) and use of the estimates to assess goodness-of-fit. Section 2 reviews relevant results on maximum likelihood estimates in canonical exponential families. Computation of the MLE $\hat\beta_{M,n}$ requires an iterative algorithm. For the discrete exponential family known as the log-linear model, there exist empirically weighted least squares estimates of the parameters which are asymptotically equivalent to the MLE, but are much simpler to calculate (see Grizzle, Starmer, Koch (1969) and Haberman (1974) for details). A heuristic application of the idea to the continuous model (1.6) leads to the regression estimator

$$\begin{aligned} \hat\beta_{R,n} = &\left[\Sigma_{i=1}^n(v(X_i) - \bar v_n)(v(X_i) - \bar v_n)^T\right]^{-1}\Sigma_{i=1}^n(v(X_i) - \bar v_n) \\ &\times(\log(\hat f_n(X_i)) - \bar m_n), \end{aligned} \tag{1.10}$$

where the $\{X_i\}$ are i.i.d. observations, $\bar{v}_n = n^{-1}\Sigma_{i=1}^n v(X_i)$, $\hat{f}_n$ is a suitable nonparametric estimator of the density of X_i and $\bar{m}_n = n^{-1}\Sigma_{i=1}^n \log(\hat{f}_n(X_i))$. Section 3 develops asymptotic distribution theory for $\hat{\beta}_{R,n}$, showing in particular that $n^{\frac{1}{2}}(\hat{\beta}_{M,n} - \hat{\beta}_{R,n}) = o_P(1)$ when the $\{X_i\}$ are independently distributed according to model (1.6). Robustness of both the MLE and the regression estimator of β under small perturbations of the parametric model (1.6) is analyzed in Section 4. A brief discussion of goodness-of-fit tests in Section 5 completes the paper. Though motivated by the directional models (1.4) and (1.5), the asymptotic results (other than those in Section 3.2) remain valid for canonical exponential families with continuous basis v on a compact space.

The results in this paper can be extended to multivariate directional distributions on product spaces of the form $\Pi_{j=1}^t S_{p_j}$. Other questions also arise. Associated with the Fisher-von Mises and Bingham distributions are some interesting estimation and hypothesis testing problems which are suggested by the geometry of the model. For instance: testing whether the modal directions of several independent Fisher-von Mises samples are coplanar and identifying the common plane (Watson (1960)); or testing for rotational symmetry within the Bingham model and identifying the axis of symmetry (Bingham (1974)). We hope to treat elsewhere the generalization of these questions to the directional model (1.3).

2. Maximum likelihood estimates.

Let the observed random variables $\{X_i : 1 \leqslant i \leqslant n\}$ be independent, identically distributed according to some density g on S_p. The log-likelihood function corresponding to the density f_β defined in (1.6) is

$$L_n(\beta) = \beta^T\Sigma_{i=1}^n v(X_i) - nc(\beta). \tag{2.1}$$

From the literature on canonical exponential families, we obtain the following information (see Barndorff-Nielsen (1973), Berk (1972), Crain (1976a, 1976b), Lehmann (1959)):

(i) $L_n(\beta)$ is strictly concave in β;

(ii) $c(\beta)$ is analytic in each component of β and $\nabla c(\beta) = E_\beta(v(X))$, $\nabla^2 c(\beta) = \mathrm{Cov}_\beta(v(X))$, the expectation and covariance matrix being evaluated under the model (1.6);

(iii) $\nabla^2 c(\beta)$ is positive definite;

(iv) If the MLE $\hat{\beta}_{M,n}$ exists, it is unique;

(v) With probability one, there exists an integer $n_0 = n_0(X_1, X_2, \cdots)$ such that the MLE $\hat{\beta}_{M,n}$ exists for every $n \geqslant n_0$;

(vi) A necessary and sufficient condition for existence of the MLE is that $n^{-1}\Sigma_{i=1}^n v(X_i) \in \text{int conv }(K)$, where $K = \text{range }\{v(x); x \in S_p\} \subset R^q$ and conv (K) is the convex hull of K;

(vii) The MLE $\hat{\beta}_{M,n}$ exists iff the equations $E_\beta(v(X)) = n^{-1}\Sigma_{i=1}^n v(X_i)$ have a solution; when a solution exists, it is unique and is the MLE.

Let g be any density S_p with respect to the invariant measure μ. Define an R^q-valued functional β_M implicitly through the equation

$$\int v(x) f_{\beta_M(g)}(x)\mu(dx) = \int v(x) g(x)\mu(dx). \tag{2.2}$$

It can be shown that $\beta_M(g)$ exists and is unique (Crain (1974)) if $\inf_{x \in S_p} g(x) > 0$. The asymptotic behavior of $\hat{\beta}_{M,n}$ is described by the following theorem (proved like Crain (1976b); Huber (1967) has related results for MLE's in general.)

THEOREM 1. *Suppose the random variables* $\{X_i : 1 \leqslant i \leqslant n\}$ *are* i.i.d. *with density* g *in* S_p *such that* $\inf_{x \in S_p} g(x) > 0$. *Then*

$$\begin{aligned} &n^{1/2}(\hat{\beta}_{M,n} - \beta_M(g)) \\ &\qquad = [\nabla^2 c(\beta(g))]^{-1} n^{-1/2} \Sigma_{i=1}^n [v(X_i) - E_g(v(X))] + o_p(1). \end{aligned} \tag{2.3}$$

Thus the limiting distribution of $n^{1/2}(\hat{\beta}_{M,n} - \beta_M(g))$ *as* $n \to \infty$ *is* $N(0, \Sigma(g))$ *with*

$$\Sigma(g) = [\nabla^2 c(\beta_M(g))]^{-1} \mathrm{Cov}_g(v(X))[\nabla^2 c(\beta_M(g))]^{-1}. \tag{2.4}$$

When $g = f_\beta$, $\beta_M(g)$ reduces to β and the covariance matrix (2.4) becomes $[\nabla^2 c(\beta)]^{-1}$, which is the inverse of the Fisher information matrix for the density f_β. From the corresponding specialization of (2.3), it can be shown that the MLE $\hat{\beta}_{M,n}$ is asymptotically least dispersed among all regular estimators of β (Hájek (1970)).

Computation of the MLE for the models of Section 1 must be done by iteration. A modified Newton-Raphson algorithm, which converges to the MLE whenever it exists, is defined by the following iterative step:

$$\hat{\beta}_{M,n}^{(t+1)} = \hat{\beta}_{M,n}^{(t)} + \alpha^{(t)} s_n^{(t)}, \qquad t \geqslant 0 \tag{2.5}$$

where

$$s_n^{(t)} = \left[\mathrm{Cov}_{\hat{\beta}_{M,n}^{(t)}}(v(X))\right]^{-1}\left[n^{-1}\Sigma_{i=1}^n v(X_i) - E_{\hat{\beta}_{M,n}^{(t)}}(v(X))\right] \tag{2.6}$$

and $\alpha^{(t)}$ is a positive real number chosen to satisfy the inequality

$$L_n(\hat{\beta}_{M,n}^{(t+1)}) - L_n(\hat{\beta}_{M,n}^{(t)}) \geqslant b\alpha^{(t)} |[\nabla L_n(\hat{\beta}_{M,n}^{(t+1)})]^T s_n^{(t)}|, \tag{2.7}$$

where b is a constant in $(1, \infty)$ which is fixed over all iterations. Numerical integration will usually be necessary to evaluate the moments and $c(\beta)$. The value of $\alpha^{(t)}$ which maximizes $L_n(\hat{\beta}_{M,n}^{(t)} + \alpha^{(t)} s_n^{(t)})$ over all real $\alpha^{(t)}$ necessarily satisfies (2.7). Moreover, for every choice of starting value $\hat{\beta}_{M,n}^{(0)}$, there exists a t_0 such that for every $t \geqslant t_0$, $\alpha^{(t)} = 1$ will satisfy (2.7). A proof of the assertions made in this paragraph may be found in Haberman (1974), Chapter 3 and Appendix C. As a good choice of starting value $\hat{\beta}_{M,n}^{(0)}$ for the algorithm, we recommend the regression estimator $\hat{\beta}_{R,n}$, since it is asymptotically equivalent to the MLE under model (1.6).

3. Regression estimator.

3.1. *Asymptotic distributions.* The asymptotic behavior of the regression estimator $\hat{\beta}_{R,n}$ defined in (1.10) will be derived from that of the estimator

$$\hat{\alpha}_n = [\Sigma_{i=1}^n t(X_i) t^T(X_i)]^{-1} \Sigma_{i=1}^n t(X_i) \log[\hat{f}_n(X_i)], \tag{3.1}$$

where $t^T(x) = (1, v^T(x))$ and the density estimator $\hat{f}_n$ is described below. Under the canonical exponential family model f_β, $\hat{\alpha}_n$ turns out to be a consistent estimator for $\alpha = (-c(\beta), \beta^T)^T$. The regression estimator $\hat{\beta}_{R,n}$ is simply the subvector of $\hat{\alpha}_n$ obtained by deleting the first component of $\hat{\alpha}_n$. Formula (1.10) can be derived from (3.1) by using a well-known expression for the inverse of a partitioned symmetric matrix (cf. Rao (1965) page 29).

Let P be the probability on S_p whose density with respect to μ is g and let P_n be the empirical probability measure which assigns mass n^{-1} to each of the observations $\{X_i : 1 \leqslant i \leqslant n\}$. The density estimator $\hat{f}_n$ will be required to possess the following properties: if the $\{X_i : 1 \leqslant i \leqslant n\}$ are i.i.d. with density g, then

$$\begin{aligned}(3.4)\qquad &\sup_{x\in S_p}|\hat{f}_n(x) - g(x)| = o_p(1)\lim_{n\to\infty}n^{1/2}E_g\int\big[\hat{f}_n(x) - g(x)\big]^2\mu(dx) = 0\\ &n^{1/2}\int t(x)\big[\hat{f}_n(x) - g(x)\big]\mu(dx) = n^{1/2}\int t(x)d(P_n - P) + o_p(1).\end{aligned}$$

Construction of such an $\hat{f}_n$ is deferred to subsection 3.2. Our immediate goal is to state and prove a theorem for $\hat{\alpha}_n$, and hence $\hat{\beta}_{R,n}$, which is similar to Theorem 1.

THEOREM 2. *Suppose the random variables* $\{X_i : 1 \leqslant i \leqslant n\}$ *are* i.i.d. *with density* g *on* S_p, $\inf_{x\in S_p}g(x) > 0$, *and the density estimator* $\hat{f}_n$ *satisfies* (3.4). *Then*

$$(3.5)\qquad n^{1/2}(\hat{\alpha}_n - \alpha(g)) = V^{-1}(g)n^{-1/2}\Sigma_{i=1}^n\big[u(X_i) - E_g(u(X))\big] + o_p(1)$$

where

$$\begin{aligned}&V(g) = E_g\big[t(X)t^T(X)\big]\\ (3.6)\qquad &\alpha(g) = V^{-1}(g)E_g\big[t(X)\log(g(X))\big]\\ &u(x) = t(x)\big[1 + \log(g(x)) - t^T(x)\alpha(g)\big].\end{aligned}$$

Thus the limiting distribution of $n^{1/2}(\hat{\alpha}_n - \alpha(g))$ *as* $n \to \infty$ *is* $N(0, S(g))$ *with*

$$(3.7)\qquad S(g) = V^{-1}(g)\mathrm{Cov}_g(u(X))V^{-1}(g).$$

The nonsingularity of $V(g)$ follows from the assumed linear independence of the components of $t(x)$. In the special case $g = f_\beta$, $\log(g(x))$ reduces to $\alpha^T t(x)$; hence $\alpha(g) = \alpha$, $u(x) = t(x)$, and (3.5) implies

$$\begin{aligned}(3.8)\qquad &n^{1/2}(\hat{\beta}_{R,n} - \beta)\\ &\qquad = \big[\mathrm{Cov}_\beta(v(X))\big]^{-1}n^{-1/2}\Sigma_{i=1}^n\big[v(Xi) - E_\beta(v(X))\big] + o_p(1).\end{aligned}$$

Comparison of (3.8) with the corresponding specialization of (2.3) in Theorem 1 shows that $n^{1/2}(\hat{\beta}_{M,n} - \hat{\beta}_{R,n}) = o_p(1)$ under the sequence of distributions $\{\Pi_{i=1}^n f_\beta(x_i)\}$; both $n^{1/2}(\hat{\beta}_{M,n} - \beta)$ and $n^{1/2}(\hat{\beta}_{R,n} - \beta)$ are asymptotically $N(0, [\mathrm{Cov}_\beta(v(X))]^{-1})$ in this case.

To prove Theorem 2, we begin with

LEMMA 1. *Let* $\{Z_n : n \geqslant 1\}$ *be a sequence of processes with sample paths in* $C(S_p)$ *such that the sequence of maxima* $\{\sup_{x\in S_p}|Z_n(x)|\}$ *is bounded in probability.*

Then

$$n^{1/2}\int Z_n(x)d(P_n - P) = O_p(1). \tag{3.9}$$

PROOF. For every subset $A \subset C(S_p)$ which is bounded in sup norm, an argument by contradiction shows that

$$\sup_{b\in A}|n^{1/2}\int b(x)d(P_n - P)| = O_p(1). \tag{3.10}$$

The assumption on Z_n entails that for every $\varepsilon > 0$, there exists a bounded subset $A_\varepsilon \subset C(S_p)$ such that $P[Z_n \notin A_\varepsilon] < \varepsilon$. Let T_n denote the integral in (3.9). Since

$$P[|T_n| > \delta] \leqslant P[|T_n| > \delta, Z_n \in A_\varepsilon] + P[Z_n \notin A_\varepsilon], \tag{3.11}$$

the lemma follows.

Having established this lemma, we proceed to the

PROOF OF THEOREM 2. The estimator $\hat{\alpha}_n$ can be expressed in the form

$$\hat{\alpha}_n = \left[\int t(x)t^T(x)dP_n\right]^{-1}\Sigma_{i=1}^3 W_{in} \tag{3.12}$$

where

$$\begin{aligned} W_{1n} &= \int t(x)\log(g(x))dP \\ W_{2n} &= \int t(x)\log(g(x))d(P_n - P) \\ W_{3n} &= \int t(x)\left[\log(\hat{f}_n(x)) - \log(g(x))\right]dP_n. \end{aligned} \tag{3.13}$$

For n sufficiently large and $V(g)$ defined by (3.6),

$$\begin{aligned} \left[\int t(x)t^T(x)dP_n\right]^{-1} &= \left[V(g) + \int t(x)t^T(x)d(P_n - P)\right]^{-1} \\ &= V^{-1}(g) - V^{-1}(g)\left[\int t(x)t^T(x)d(P_n - P)\right]V^{-1}(g) + O_p(n^{-1}). \end{aligned} \tag{3.14}$$

Note that the nonsingularity of $V(g)$ ensures w.p. 1 the asymptotic nonsingularity of the random matrix inverted in (3.14). Using (3.14),

$$\begin{aligned} &\left[\int t(x)t^T(x)dP_n\right]^{-1}W_{1n} \\ &\qquad = \alpha(g) - V^{-1}(g)\int t(x)t^T(x)\alpha(g)d(P_n - P) + O_p(n^{-1}). \end{aligned} \tag{3.15}$$

The first assertion in (3.4) and the assumption $\inf_{x\in S_p}g(x) > 0$ imply that

$$\sup_{x\in S_p}|\log(\hat{f}_n(x)) - \log(g(x))| = o_p(1). \tag{3.16}$$

An application of Lemma 1, followed by Taylor expansion of the logarithm to the second derivative term and use of the second and third parts of (3.4) yields

$$\begin{aligned} n^{1/2}W_{3n} &= n^{1/2}\int t(x)\left[\log(\hat{f}_n(x)) - \log(g(x))\right]dP + o_p(1) \\ &= n^{1/2}\int t(x)\left[\hat{f}_n(x) - g(x)\right]\mu(dx) + o_p(1) \\ &= n^{1/2}\int t(x)d(P_n - P) + o_p(1). \end{aligned} \tag{3.17}$$

As a consequence of the approximations (3.14), (3.15) and (3.17), we have

$$(3.18)\quad n^{1/2}(\hat{\alpha}_n - \alpha(g))$$
$$= V^{-1}(g)n^{1/2}\int t(x)\left[1 + \log(g(x)) - t^T(x)\alpha(g)\right]d(P_n - P) + o_p(1),$$

which is equivalent to (3.5). The theorem follows.

3.2. *Density estimators.* It remains to construct a density estimator $\hat{f}_n$ on S_p which possesses the asymptotic properties listed in (3.4). We will explore only one of the possible approaches to the problem—that of window density estimators. Let w: $[0, \infty) \to R$ be any function which satisfies integrability and smoothness assumptions to be determined. For every x, $y \in S_p$, put $w_n(1 - x^Ty) = A_{n,p}^{-1}w[c_n^{-1}(1 - x^Ty)]$, where $\{c_n\}$ is a sequence of positive constants converging to zero at a rate to be chosen later and $A_{n,p} = \int w[c_n^{-1}(1 - x^Ty)]\mu(dy)$. Note that $A_{n,p}$ does not depend upon x because the measure μ is invariant under rotation. Taking polar coordinates for S_p with x as pole yields the expression

$$(3.19)\quad A_{n,p} = B^{-1}(1/2, (p-1)/2)\int_0^{\pi} w\left[c_n^{-1}(1 - \cos(\theta))\right]\sin^{p-2}(\theta)d\theta$$
$$= B^{-1}(1/2, (p-1)/2)c_n^{(p-1)/2}\int_0^{2/c_n} t^{(p-3)/2}(2 - c_nt)^{(p-3)/2}w(t)dt,$$

where $B(\cdot, \cdot)$ is the beta function. Define

$$(3.20)\qquad \hat{f}_n(x) = n^{-1}\Sigma_{i=1}^n w_n(1 - x^TX_i).$$

If $\int_0^\infty t^{(p-3)/2}|w(t)|dt < \infty$ and w is nonnegative, $\hat{f}_n$ is a probability density on S_p with respect to the measure μ.

The density estimator (3.20) is not really new, although its statistical behavior appears not to have been studied. A graphical density estimator used by geologists amounts to the special case $p = 3$ and $w(t) = 1$ in $[0, 1]$, vanishing elsewhere (Watson (1970), page 78). Another special case of (3.20), of greater interest for this paper but apparently untried in practice, arises when $w(t) = \exp(-t)$ on R^+. In this instance,

$$(3.21)\qquad A_{n,p} = (2c_n)^{p/2-1}\Gamma(p/2)\exp(-c_n^{-1})I_{p/2-1}(c_n^{-1})$$

with $I_r(\cdot)$ denoting the modified Bessel function of the first kind and order r (cf. Whittaker and Watson (1927) page 373); moreover, $w_n(1 - x^TX_i)$ regarded as a function of $x \in S_p$ is just the Fisher-von Mises density with modal direction X_i.

The amount of bias $E(\hat{f}_n(x)) - g(x)$ in the density estimator (3.20) depends upon the smoothness of g and the choice of window w. To obtain density estimators $\hat{f}_n$ which satisfy (3.4) when $p = 2$ or 3, the cases of greatest practical importance, we investigate the asymptotic behavior of the estimator (3.20) under two different sets of assumptions on g and w.

The first set of assumptions:

A1. $w : R^+ \to R^+$ is positive, bounded and continuous, with

$$\int_0^\infty t^{(p-1)/2}w(t)dt < \infty \qquad \text{and} \qquad \int_0^\infty t^{(p-3)/2}w(t)dt < \infty.$$

A2. There exists a $p \times 1$ vector function $\nabla g : S_p \to R^p$ such that:

(i) For every $x \in S_p$, $x^T \nabla g(x) = 0$;

(ii) for every $x, y \in S_p$,

$$g(y) = g(x) + (y - x)^T \nabla g(x) + r(x, y)$$

where $|r(x, y)| \leqslant B\|y - x\|^2$ for some universal constant B and euclidean norm $\|\cdot\|$.

Part (i) of A2 is a convenient normalization of $\nabla g(x)$, rather than an additional restriction. Indeed, $\nabla g(x)$ can be replaced by $\nabla g(x) + \alpha x$, α an arbitrary scalar, without affecting the validity of the expansion in part (ii).

The second set of assumptions:

B1. $w : R^+ \to R$ is bounded and continuous, with

$$\int_0^\infty t^{(p-1)/2}|w(t)|dt < \infty, \qquad \int_0^{2/c_n} t^{(p-1)/2} w(t)dt = o\big(c_n^{1/2}\big)$$

and

$$\int_0^\infty t^{(p-3)/2}|w(t)|dt < \infty, \int_0^\infty t^{(p-3)/2} w(t)dt \neq 0.$$

B2. There exists on S_p a $p \times 1$ vector function $\nabla g(x)$ and a bounded $p \times p$ matrix function $\nabla^2 g(x)$ such that:

(i) For every $x \in S_p$, $x^T \nabla g(x) = 0$ and $x^T \nabla^2 g(x) x = 0$;

(ii) for every $x, y \in S_p$,

$$g(y) = g(x) + (y - x)^T \nabla g(x) + (y - x)^T \nabla^2 g(x)(y - x) + r(x, y)$$

where $|r(x, y)| < B\|y - x\|^3$ for some universal constant B.

The normalization $x^T \nabla^2 g(x) x = 0$ in $B2$ is possible because $\nabla^2 g(x)$ can be replaced by $\nabla^2 g(x) + \gamma x x^T$, γ an arbitrary scalar without affecting the validity of the expansion in part (ii) of $B2$.

The directional densities (1.4) and (1.5) satisfy both $A2$ and $B2$ because the basis functions $\{v_i : 1 \leqslant i \leqslant q\}$ which appear in the canonical form (1.6) can be monomials.

Corresponding to the two sets of assumptions are the following two lemmas.

LEMMA 2. *Suppose the random variables* $\{X_i : 1 \leqslant i \leqslant n\}$ *are* i.i.d. *with density* g *on* S_p, *assumptions* A1, A2 *are satisfied, and* $\lim_{n\to\infty} c_n = 0$. *Then*

$$\sup_{x\in S_p}|E_g(\hat{f}_n(x)) - g(x)| = O(c_n)$$

$$\sup_{x\in S_p}\operatorname{Var}_g(\hat{f}_n(x)) = O\big(n^{-1}c_n^{-(p-1)/2}\big) \tag{3.22}$$

$$\sup_{x\in S_p}|\hat{f}_n(x) - E_g(\hat{f}_n(x))| = O_p\big(n^{-1/2}c_n^{-(p-1)/2}\big)$$

and

$$n^{1/2}\int t(x)\big[\hat{f}_n(x) - E_g(\hat{f}_n(x))\big]\mu(dx) = n^{1/2}\int t(x)d(P_n - P) + o_p(1). \tag{3.23}$$

PROOF. Because of assumption A2, the difference $E(\hat{f}_n(x)) - g(x)$ can be expressed as the sum of two terms $T_{1n}(x)$, $T_{2n}(x)$ such that

$$(3.24) \qquad T_{1n}(x) = \left[A_{n,p}^{-1} \int w\left[c_n^{-1}(1 - x^T y)\right](y - x)\mu(dy)\right]^T \nabla g(x)$$

$$|T_{2n}(x)| \leqslant B A_{n,p}^{-1} \int w\left[c_n^{-1}(1 - x^T y)\right] \|y - x\|^2 \mu(dy).$$

By switching to polar coordinates for S_p, with x as pole, we find that $T_{1n}(x)$ is a multiple of $x^T \nabla g(x)$, which vanishes, and that

$$(3.25) \qquad |T_{2n}(x)| \leqslant D_{n,p} \int_0^\pi w\left[c_n^{-1}(1 - \cos(\theta)\right](1 - \cos(\theta))\sin^{p-2}(\theta)\,d\theta$$

$$= D_{n,p} c_n^{(p+1)/2} \int_0^{2/c_n} t^{(p-1)/2}(2 - c_n t)^{(p-3)/2} w(t)\,dt,$$

where $D_{n,p} = 0(c_n^{-(p-1)/2})$. Since the bound in (3.25) is $0(c_n)$, the first assertion in (3.22) follows.

The calculation of $\operatorname{Var}(\hat{f}_n(x))$ is similar, the essential part being

$$(3.26) \quad E\left[w_n^2(1 - x^T X_i)\right] = A_{n,p}^{-2}\left[g(x) \int w^2\left[c_n^{-1}(1 - x^T y)\right] \mu(dy) + o(1)\right]$$

$$= O\left(c_n^{-(p-1)/2}\right)$$

uniformly in x, since $A2$ implies continuity of $g(x)$ on S_p.

To verify the third part of (3.22), let $U_n(x) = \hat{f}_n(x) - E(\hat{f}_n(x))$. If ξ_n denotes the (random) maximizing value of x, which exists by continuity of w,

$$(3.27) \quad \sup_{x \in S_p} |U_n(x)| = \left(n^{1/2} A_{n,p}\right)^{-1} n^{1/2} \left|\int w\left[c_n^{-1}(1 - \xi_n^T y)\right] d(P_n - P)\right|.$$

Since w is also bounded, $\sup_{y \in S_p} |w[c_n^{-1}(1 - \xi_n^T y)]| = O_p(1)$. Application of Lemma 1 to (3.27) yields the desired bound.

Let V_{1n} and V_{2n} denote, respectively, the integrals on the left side and right side of (3.23). The validity of (3.23) will be established by showing that

$$(3.28) \qquad \lim_{n \to \infty} E\left[b^T(V_{1n} - V_{2n})\right]^2 = 0$$

for every $p \times 1$ real column vector b. Let $s(x) = b^T t(x)$ and $s_n(z) = \int s(x) w_n(1 - x^T z)\mu(dx)$. Evidently $E(V_{1n}) = E(V_{2n}) = 0$ and $\operatorname{Var}(b^T V_{2n}) = E(s^2(X)) - [E(s(X))]^2$. Moreover,

$$(3.29) \qquad \operatorname{Var}(b^T V_{1n}) = E(s_n^2(X)) - \left[E(s_n(X))\right]^2$$

$$\operatorname{Cov}(b^T V_{1n}, b^T V_{2n}) = E(s_n(X) s(X)) - \left[E(s_n(X))\right]\left[E(s(X))\right].$$

Since $t(x)$ is a vector of monomials (see Section 1),

$$(3.30) \qquad \lim_{n \to \infty} \sup_{z \in S_p} |s_n(z) - s(z)| = 0$$

by the same expansion argument as was used in (3.24) for $E(\hat{f}_n(x))$. Hence,

$$(3.31) \qquad \lim_{n \to \infty} \operatorname{Var}(b^T V_{1n}) = \lim_{n \to \infty} \operatorname{Cov}(b^T V_{1n}, b^T V_{2n})$$

$$= \operatorname{Var}(b^T V_{2n}),$$

which implies (3.28).

LEMMA 3. *Suppose the random variables* $\{X_i : 1 \leqslant i \leqslant n\}$ *are* i.i.d. *with density* g *on* S_p, *assumptions* B1, B2 *are satisfied, and* $\lim_{n\to\infty} c_n = 0$. *Then*

$$\sup_{x\in S_p} |E_g(\hat{f}_n(x)) - g(x)| = O\big(c_n^{\frac{3}{2}}\big) \tag{3.32}$$

and the other conclusions of Lemma 2 *remain valid.*

PROOF. Because of assumption B2, the difference $E(\hat{f}_n(x)) - g(x)$ can be written as the sum of three terms:

$$\begin{aligned} W_{1n}(x) &= \big[A_{n,p}^{-1} \int w\big[c_n^{-1}(1 - x^T y)\big](y - x)\mu(dy)\big]^T \nabla g(x) \\ W_{2n}(x) &= A_{n,p}^{-1} \int w\big[c_n^{-1}(1 - x^T y)\big](y - x)^T \nabla^2 g(x)(y - x)\mu(dy) \\ |W_{3n}(x)| &\leqslant BA_{n,p}^{-1} \int w\big[c_n^{-1}(1 - x^T y)\big]\|y - x\|^3 \mu(dy). \end{aligned} \tag{3.33}$$

Switching to polar coordinates for S_p, with x as pole, and arguing as in Lemma 2 yields the facts $W_{1n}(x) \equiv 0$, $\sup_{x\in S_p}|W_{3n}(x)| = O(c_n^{\frac{3}{2}})$ and

$$\begin{aligned} W_{2n}(x) = A_{n,p}^{-1} &\int_0^\pi w\big[c_n^{-1}(1 - \cos(\theta))\big] \\ &\times \big[D_1(x)\sin^2(\theta) + D_2(x)(1 - \cos(\theta))^2\big]\sin^{p-2}(\theta)\,d\theta \end{aligned} \tag{3.34}$$

for some bounded nonnegative functions D_1, D_2. Since

$$\begin{aligned} &\int_0^\pi w\big[c_n^{-1}(1 - \cos(\theta))\big]\sin^p(\theta)\,d\theta \\ &\qquad = c_n^{(p+1)/2} \int_0^{2/c_n} t^{(p-1)/2}(2 - c_n t)^{(p-1)/2} w(t)\,dt \\ &\qquad = c_n^{(p+1)/2} \int_0^{2/c_n} t^{(p-1)/2} w(t)\,dt + O\big(c_n^{(p+3)/2}\big) \end{aligned} \tag{3.35}$$

and

$$\int_0^\pi w\big[c_n^{-1}(1 - \cos(\theta))\big](1 - \cos(\theta))^2 \sin^{p-2}(\theta)\,d\theta = O\big(c_n^{(p+3)/2}\big) \tag{3.36}$$

it follows, using (3.19) and B1, that $\sup_{x\in S_p}|W_{2n}(x)| = o(c_n^{\frac{3}{2}})$. Hence (3.32) holds. The rest of the lemma is proved like Lemma 2, allowing for the fact that $w(t)$ assumes both positive and negative values in this case.

The two lemmas just proved enable us to ensure that the density estimator (3.20) has the properties (3.4) which are needed for Theorem 2. If $p = 2$ and w satisfies A1, the fulfillment of (3.4) follows from Lemma 2, provided $\lim_{n\to\infty} n^{1/2} c_n = 0$, $\lim_{n\to\infty} n^{1/2} c_n^{1/2} = \infty$. This construction fails when $p = 3$, because the bias in $\hat{f}_n$ is then too large relative to the variance of $\hat{f}_n$. However, if $p = 3$ and w satisfies B1, then (3.4) follows from Lemma 3, provided $\lim_{n\to\infty} n^{1/2} c_n^{3/2} = 0$, $\lim_{n\to\infty} n^{1/2} c_n = \infty$. Choosing w is not difficult; $w(t) = \exp(-t)$ satisfies A2 while $w(t) = (1 - t/2)\exp(-t)$ satisfies B1 if $p = 3$.

4. Robustness of the estimators. A good estimator of the parameter β should not only be asymptotically efficient if the postulated exponential family model (1.6) were in fact correct, but should also be robust against the small departures from this ideal model that will occur in practice. By robust, we mean that the distribution of the estimator is not greatly perturbed if the assumed model is only approximately true. A mathematical formulation of this robustness concept was first set forth by Hampel (1971); a different technical formulation, introduced in Beran (1978) will be used in this section.

Since it is not feasible to calculate the exact distributions of an estimator over a very large set of alternative models, asymptotic methods become indispensable in analyzing robustness of a procedure. However, we cannot assess an estimator's robustness by studying only its asymptotic distributions over a neighborhood of possible models for the data. The convergence in law to the limit distributions might not be uniform over the neighborhood, in which event some of the limit distributions would approximate poorly some of the exact distributions. The theorem proved in this section shows that the convergence in law established in Theorem 1 for the MLE $\hat{\beta}_{M,n}$ and in Theorem 2 for the regression estimator $\hat{\beta}_{R,n}$ is locally uniform at the exponential family model f_β. Thus, the asymptotic distributions of these estimators under models near f_β provide reliable approximations to the exact distributions and can be examined to assess robustness.

Technically, the theorem proved in this section describes the asymptotic behavior of $\hat{\beta}_{M,n}$ and $\hat{\beta}_{R,n}$ under general sequences of densities $\{\Pi_{i=1}^n g_n(x_i)\}$ which are contiguous to the model densities $\{\Pi_{i=1}^n f_\beta(x_i)\}$. Contiguity is achieved by requiring that

$$g_n^{1/2}(x) = \cos(bn^{-1/2})f_\beta^{1/2}(x) + \sin(bn^{-1/2})\delta(x) \tag{4.1}$$

for some scalar b and some function $\delta \in L_2(S_p)$ which has unit length in L_2-norm and is orthogonal to $f_\beta^{1/2}$. Both b and δ are allowed to vary over their domains so as to generate different sequences of densities $\{g_n\}$. As in Le Cam (1969), the log-likelihood ratio $L_n = \log[\Pi_{i=1}^n(g_n(X_i)/f_\beta(X_i))]$ can be approximated, under the model distribution $\Pi_{i=1}^n f_\beta(x_i)$, by

$$L_n = 2n^{-1/2}b\Sigma_{i=1}^n\delta(X_i)f_\beta^{-1/2}(X_i) - 2b^2 + o_p(1). \tag{4.2}$$

Thus, the limiting distribution of L_n under $\{\Pi_{i=1}^n f_\beta(x_i)\}$ is $N(-2b^2, 4b^2)$, which implies the contiguity of $\{\Pi_{i=1}^n g_n(x_i)\}$ and $\{\Pi_{i=1}^n f_\beta(x_i)\}$.

Let $\|\cdot\|_P$ denote the Prohorov metric on probabilities and let $\mathcal{D}_{g_n}(T_n)$ stand for the distribution of the argument statistic $T_n = T_n(\mathbf{X}_n)$, when $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)$ is distributed according to the density $\Pi_{i=1}^n g_n(x_i)$. Let $\beta_R(g)$ be the $q \times 1$ vector obtained by deleting the first component of $\alpha(g)$, which was defined in (3.6). The phrase "weakly compact" in the theorem statement below means compact in the product topology generated by weak convergence in $L_2(S_p)$ and ordinary convergence on the real line.

THEOREM 3. *Suppose* $\{g_n\}$ *is defined by* (4.1). *Let* K *be an arbitrary weakly compact subset of* $\{(b, \delta) \in R \times L_2(S_p)$: $\int\delta^2(x)\mu(dx) = 1$, $\delta \perp f_\beta^{1/2}$, $\sup_{x\in S_p}|\delta(x)| \leqslant C < \infty\}$. *Then*

$$\lim_{n\to\infty}\sup_{(b,\delta)\in K}\|\mathcal{D}_{g_n}\left[n^{1/2}\left(\hat{\beta}_{M,n} - \beta_M(g_n)\right)\right] - N\left(0, \left[\nabla^2 c(\beta)\right]^{-1}\right)\|_P = 0 \tag{4.3}$$

and

$$\lim_{n\to\infty}\sup_{(b,\delta)\in K}\|\mathcal{D}_{g_n}\left[n^{1/2}(\hat{\beta}_{R,n}-\beta_R(g_n)\right] - N\left(0,\left[\nabla^2 c(\beta)\right]^{-1}\right)\|_P = 0. \tag{4.4}$$

Since $\hat{\beta}_{M,n}$ is centered by the same functional β_M in both (2.3) of Theorem 1 and (4.3) of Theorem 3, Theorem 3 indicates that the convergence to the limiting distributions in Theorem 1 occurs uniformly over a rich set of local (in Hellinger metric) perturbations of f_β. Moreover, the distributions of the centered estimator under these local perturbations are all approximately $N(0,[\nabla^2 c(\beta)]^{-1})$ for n large enough. From this and the property $\lim_{n\to\infty}\beta(g_n)=\beta$, which follows from the proof below, we may conclude that sufficiently small, fairly arbitrary perturbations of f_β do not affect the exact distributions of $\hat{\beta}_{M,n}$ very much, at least in large samples. This conclusion is an asymptotic version of the qualitative robustness property discussed at the beginning of this section.

Similarly, Theorem 2 and (4.4) of Theorem 3 justify the assertion that the regression estimator $\hat{\beta}_{R,n}$ is robust, at least in large samples.

PROOF OF THEOREM 3. Since the two results are analogous, we give only the argument for (4.4). Suppose (4.4) were false. Then, by weak compactness of K, there would exist a weakly convergent sequence $\{(b_n,\delta_n)\in K\}$ such that $\mathcal{D}_{g_n}[n^{1/2}(\hat{\beta}_{R,n}-\beta_R(g_n))]\not\Rightarrow N(0,[\nabla^2 c(\beta)]^{-1})$ for

$$g_n^{1/2}(x)=\cos(b_n n^{-1/2})f_\beta^{1/2}(x)+\sin(b_n n^{-1/2})\delta_n(x). \tag{4.5}$$

Thus, to prove (4.4), it suffices to show that

$$\mathcal{D}_{g_n}\left[n^{1/2}(\hat{\beta}_{R,n}-\beta_R(g_n))\right]\Rightarrow N\left(0,\left[\nabla^2 c(\beta)\right]^{-1}\right) \tag{4.6}$$

for every sequence $\{(b_n,\delta_n)\in K\}$ converging weakly to some $\{(b,\delta)\in K\}$, with $\{g_n\}$ defined by (4.5).

For every such sequence $\{g_n\}$, the approximation (4.2) to the log-likelihood ratio L_n continues to hold. Thus, by a standard contiguity argument, the limiting distribution of $n^{1/2}(\hat{\beta}_{R,,n}-\beta)$ under $\{\Pi_{i=1}^n g_n(x_i)\}$ is $N(2b[\nabla^2 c(\beta)]^{-1}\int v(x)\delta(x)f_\beta^{1/2}(x)\mu(dx),[\nabla^2 c(\beta)]^{-1})$. Hence, (4.6) is true if and only if

$$\lim_{n\to\infty}n^{1/2}(\beta_R(g_n)-\beta)=2b\left[\nabla^2 c(\beta)\right]^{-1}\int v(x)\delta(x)f_\beta^{1/2}(x)\mu(dx) \tag{4.7}$$

for every sequence $\{(b_n,\delta_n)\in K\}$ which converges weakly to some $(b,\delta)\in K$. In the notation of Section 3.1, (4.7) is equivalent to

$$\lim_{n\to\infty}n^{1/2}(\alpha(g_n)-\alpha)=2bV^{-1}(f_\beta)\int t(x)\delta(x)f_\beta^{1/2}(x)\mu(dx), \tag{4.8}$$

which will now be established.

From the definition (3.6), $\alpha(g_n) = V^{-1}(g_n)\Sigma_{i=1}^{3} T_{in}$, where

$$\begin{aligned} T_{1n} &= \int t(x)\log(f_\beta(x))f_\beta(x)\mu(dx) = V(f_\beta)\alpha \\ T_{2n} &= \int t(x)\log(f_\beta(x))\left[g_n(x) - f_\beta(x)\right] \mu(dx) \\ T_{3n} &= \int t(x)\left[\log(g_n(x)) - \log(f_\beta(x))\right] g_n(x)\mu(dx). \end{aligned} \tag{4.9}$$

Straightforward calculations based upon (4.1) yield

(4.10)

$$\begin{aligned} V^{-1}(g_n) &= \left[I + V^{-1}(f_\beta)\int t(x)t^T(x)(g_n(x) - f_\beta(x))\mu(dx)\right]^{-1} V^{-1}(f_\beta) \\ &= V^{-1}(f_\beta) - V^{-1}(f_\beta)\left[\int t(x)t^T(x)(g_n(x) - f_\beta(x))\mu(dx)\right] V^{-1}(f_\beta) + O(n^{-1}) \end{aligned}$$

and

$$\begin{aligned} n^{1/2}T_{2n} &= 2b\int t(x)\log(f_\beta(x))\delta(x)f_\beta^{1/2}(x)\mu(dx) + o(1) \\ &= 2b\left[\int t(x)t^T(x)\delta(x)f_\beta^{1/2}(x)\mu(dx)\right]\alpha + o(1). \end{aligned} \tag{4.11}$$

A more complicated series of approximations, described below, shows that

$$n^{1/2}T_{3n} = 2b\int t(x)\delta(x)f_\beta^{1/2}(x)\mu(dx) + o(1). \tag{4.12}$$

Combining equations (4.9) through (4.12) gives

(4.13)
$$\begin{aligned} &n^{1/2}(\alpha(g_n) - \alpha) \\ &\quad= V^{-1}(f_\beta)\cdot 2b\int t(x)\left[-t^T(x)\alpha + t^T(x)\alpha + 1\right]\delta(x)f_\beta^{1/2}(x)\mu(dx) + o(1), \end{aligned}$$

and hence (4.8).

It remains to prove (4.12). For any $\varepsilon > 0$, let

$$A_{n,\varepsilon} = \left\{x \in S_p : |f_\beta^{1/2}(x)g_n^{-1/2}(x) - 1| > \varepsilon\right\}, \tag{4.14}$$

and let $\bar{A}_{n,\varepsilon}$ denote the complement in S_p of $A_{n,\varepsilon}$. Then

(4.15)
$$\begin{aligned} &\left|n^{1/2}\int_{A_{n,\varepsilon}} t(x)\left[\log(g_n(x)) - \log(f_\beta(x))\right] g_n(x)\mu(dx)\right| \\ &\quad\leqslant n^{1/2}\int_{A_{n,\varepsilon}}\left|t(x)\log(g_n(x)) + t(x)t^T(x)\alpha\right| g_n(x)\mu(dx) = O(n^{-1/2}), \end{aligned}$$

because $t(x)$ is bounded, $\log(g_n(x))$ is bounded by definition of K, and

$$\begin{aligned} \int_{A_{n,\varepsilon}} g_n(x)\mu(dx) &= P_{g_n}\left[|f_\beta^{1/2}(x)g_n^{-1/2}(x) - 1| > \varepsilon\right] \\ &\leqslant \varepsilon^{-2}\int\left[f_\beta^{1/2}(x) - g_n^{1/2}(x)\right]^2\mu(dx) \\ &= O(n^{-1}) \end{aligned} \tag{4.16}$$

by Chebyshev's inequality and (4.5).

On the other hand, by Taylor expansion,

$$\begin{aligned} & n^{1/2}\textstyle\int_{\bar{A}_{n,\varepsilon}} t(x)\left[\log(g_n(x)) - \log(f_\beta(x))\right] g_n(x)\mu(dx) \\ (4.17)\qquad & = -2n^{1/2}\textstyle\int_{\overline{A_{n,\varepsilon}}} t(x)\log\left[1 + \left(f_\beta^{1/2}(x)g_n^{-1/2}(x) - 1\right)\right] g_n(x)\mu(dx) \\ & = -\left(W_{1n} + W_{2n}\right) \end{aligned}$$

where

$$\begin{aligned} & W_{1n} = 2n^{1/2}\textstyle\int_{\bar{A}_{n,\varepsilon}} t(x)\left[f_\beta^{1/2}(x)g_n^{-1/2}(x) - 1\right] g_n(x)\mu(dx) \\ (4.18)\qquad & W_{2n} = 2n^{1/2}\textstyle\int_{\bar{A}_{n,\varepsilon}} t(x)r_n(x)g_n(x)\mu(dx) \\ & |r_n(x)| \leqslant 2^{-1}(1-\varepsilon)^{-2}\left[f_\beta^{1/2}(x)g_n^{-1/2}(x) - 1\right]^2, \qquad x \in \bar{A}_{n,\varepsilon}. \end{aligned}$$

Evidently, $W_{2n} = O(n^{-1/2})$ and, using (4.16),

$$\begin{aligned} & W_{1n} = 2n^{1/2}\textstyle\int t(x)\left[f_\beta^{1/2}(x)g_n^{-1/2}(x) - 1\right] g_n(x)\mu(dx) \\ (4.19)\qquad & \qquad -2n^{1/2}\textstyle\int_{A_{n,\varepsilon}} t(x)\left[f_\beta^{1/2}(x)g_n^{-1/2}(x) - 1\right] g_n(x)\mu(dx) \\ & = -2b\textstyle\int t(x)\delta(x)f_\beta^{1/2}(x)\mu(dx) + o(1). \end{aligned}$$

The limit (4.12) follows from the approximations of the last two paragraphs.

5. Testing goodness-of-fit.

The question of whether model (1.3) is appropriate for a given set of observations can be formulated as a hypothesis testing problem. Under the hypothesis H_n, the $\{X_i : 1 \leqslant i \leqslant n\}$ are i.i.d. with density

$$f_h(x) = \exp\left[h(x) - d(h)\right] \tag{5.1}$$

for some function h belonging to a specified invariant subspace M of $C(S_p)$. Since tests with optimal power against all possible alternatives do not exist, we are led to consider a more restricted alternative K_n, which still allows for a broad range of deviations from H_n. Under K_n, the $\{X_i\}$ are i.i.d. with density (5.1) for some function $h \in N - M$, where $N \supset M$ is another specified invariant subspace of $C(S_p)$. The subspace N is chosen so as to model anticipated departures from H. The idea of constructing alternatives to H_n in this way is due to Neyman (1937).

To complete the discussion begun in the introduction, we describe below a simple goodness-of-fit test for H_n versus K_n which is asymptotically equivalent to the likelihood ratio test. By appropriate choice of the basis functions $\{v_i(x)\}$ (see Section 1), the testing problem can be reduced to the following canonical form: the $\{X_i : 1 \leqslant i \leqslant n\}$ are i.i.d. with density

$$f_\beta(x) = \exp\left[\beta^T v(x) - c(\beta)\right], \tag{5.2}$$

where the components of $v(x) = (v_1(x), v_2(x), \cdots, v_q(x))^T$ form a basis for N, the subset $\{v_1(x), v_2(x), \cdots, v_r(x)\}$, $r < q$, constitutes a basis for M, and $\{1, v_1(x), \cdots, v_q(x)\}$ are linearly independent. Thus the hypothesis to be tested is $H_n : \beta_j = 0, r + 1 \leqslant j \leqslant q$.

Let $\hat{\beta}_n$ denote any estimator of β, such as $\hat{\beta}_{M,n}$ or $\hat{\beta}_{R,n}$, which has the property

$$(5.3)\quad n^{1/2}(\hat{\beta}_n - \beta) = \left[\mathrm{Cov}_\beta(v(X))\right]^{-1} n^{-1/2}\Sigma_{i=1}^n\left[v(X_i) - E_\beta(v(X))\right] + o_p(1)$$

under $\{\Pi_{i=1}^n f_\beta(x_i)\}$. Let $\hat{\beta}_{n,2}$ denote the $(q - r) \times 1$ lower subvector of $\hat{\beta}_n$, let

$$(5.4)\quad W_n = \left[n^{-1}\Sigma_{i=1}^n v(X_i)v^T(X_i) - \left(n^{-1}\Sigma_{i=1}^n v(X_i)\right)\left(n^{-1}\Sigma_{i=1}^n v^T(X_i)\right)\right]^{-1},$$

and let $W_{n,22}$ denote the $(q - r) \times (q - r)$ lower right-hand submatrix of W_n. The proposed test is to reject H_n for sufficiently large values of

$$(5.5)\qquad A_n = n\hat{\beta}_{n,2}^T W_{n,22}^{-1}\hat{\beta}_{n,2}^T.$$

Suppose $\beta_0 = (\beta_1, \beta_2, \cdots, \beta_r, 0, \cdots, 0)^T$ and $W_{22}(\beta_0)$ is the $(q - r) \times (q - r)$ lower right-hand submatrix of $W(\beta_0) \equiv [\mathrm{Cov}_{\beta_0}(v(X))]^{-1}$. Under H_n, $n^{1/2}(\hat{\beta}_n - \beta_0)$ is asymptotically $N(0, W(\beta_0))$ and $W_n \to_p W(\beta_0)$. Hence, the distribution of A_n under H_n is asymptotically chi-square with $q - r = \dim(N) - \dim(M)$ degrees of freedom; moreover A_n is asymptotically equivalent to the normalized likelihood ratio statistic (Wald (1943)).

The goodness-of-fit test based upon A_n is much easier to perform than the likelihood ratio test, particularly if $\hat{\beta}_n$ is the regression estimator of β. The main computational difficulty with the likelihood ratio test is the need to evaluate $c(\beta)$. Even in relatively simple special cases of (5.2), such as the Bingham distribution, closed forms for $c(\beta)$ may not exist.

Let $\{g_n\}$ be any sequence of densities on S_p such that $\lim_{n\to\infty}\int[n^{1/2}(g_n^{1/2}(x) - f_{\beta_0}^{1/2}(x)) - \gamma(x)]^2\mu(dx) = 0$ for some $\gamma \in L_2(S_p)$; necessarily, γ is orthogonal to $f_{\beta_0}^{1/2}$ in $L_2(S_p)$. Indeed,

$$\begin{aligned}\int\gamma f_{\beta_0}^{1/2}\mu(dx) &= \lim_{n\to\infty} n^{1/2}\int\left(g_n^{1/2} - f_{\beta_0}^{1/2}\right)f_{\beta_0}^{1/2}\mu(dx)\\ (5.6)\qquad &= \lim_{n\to\infty} - 2^{-1}n^{1/2}\int\left(g_n^{1/2} - f_{\beta_0}^{1/2}\right)^2\mu(dx)\\ &= 0.\end{aligned}$$

By a contiguity argument (cf. Section 4), the limiting distribution of A_n under the sequence of local alternatives $\{\Pi_{i=1}^n g_n(x_i)\}$ is noncentral chi-square with $\dim(N) - \dim(M)$ degrees of freedom and noncentrality parameter $d^T W_{22}^{-1}(\beta_0)d$, where d is the $(q - r) \times 1$ lower subvector of $2W(\beta_0)\int v(x)\gamma(x)f_{\beta_0}^{1/2}(x)\mu(dx)$. The asymptotic powers of the A_n-test and the likelihood ratio test are the same under these local alternatives.

REFERENCES

Barndorff-Nielsen, O. (1973). Exponential families: exact theory. Aarhus Univ. monograph.

Beran, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* **6** 292–313.

Berk, R. H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *Ann. Math. Statist.* **43** 193–204.

Bingham, C. (1974). An antipodally symmetric distribution on the sphere. *Ann. Statist.* **2** 1201–1225.

Crain, B. R. (1974). Estimation of distributions using orthogonal expansions. *Ann. Statist.* **2** 454–463.

Crain, B. R. (1976a). Exponential models, maximum likelihood estimation, and the Haar condition. *J. Amer. Statist. Assoc.* **71** 737–740.

CRAIN, B. R. (1976b). More on estimation of distributions using orthogonal expansions. *J. Amer. Statist. Assoc.* **71** 741–745.

DUNKL, C. F. and RAMIREZ, D. E. (1971). *Topics in Harmonic Analysis*. Appleton-Century-Crofts.

GRIZZLE, J. E., STARMER, C. F., and KOCH, G. G. (1969). Analysis of categorical data by linear models. *Biometrics*. **25** 489–504.

HABERMAN, S. (1974). *The Analysis of Frequency Data*. Univ. Chicago Press.

HÁJEK, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*. **14** 323–330.

HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887–1896.

HUBER, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probability* **I** 221–233.

LE CAM, L. (1969). *Théorie Asymptotique de la Décision Statistique*. Les Presses de l'Université de Montréal.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. John Wiley and Sons.

MARDIA, K. V. (1972). *Statistics of Directional Data*. Academic Press.

MARDIA, K. V. (1975). Statistics of directional data. *J. Roy. Statist. Soc. Ser. B* **37** 349–393.

NEYMAN, J. (1937). "Smooth" test for goodness of fit. *Skand. Aktuarietidskr.* **20** 149–199.

RAO, C. R. (1965). *Linear Statistical Inference and its Applications*. John Wiley and Sons.

WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54** 426–482.

WATSON, G. S. (1960). More significance tests on the sphere. *Biometrika*. **47** 87–91.

WATSON, G. S. (1970). Orientation statistics in the earth sciences. *Bull. Geol. Inst. Univ. Uppsala. N.S.* **2** 73–89.

WHITTAKER, E. T. and WATSON, G. N. (1927). *A Course of Modern Analysis* (4th edition). Cambridge Univ. Press.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720