Journal of the Korean Statistical Society (2004), **33**: *3, pp 255–269*

DIRECTIONAL LOG-DENSITY ESTIMATION

JIB HUH[1], PETER T. KIM[2], JA-YONG KOO[3] AND JINHO PARK[4]

ABSTRACT

This paper develops log-density estimation for directional data. The methodology is to use expansions with respect to spherical harmonics followed by estimating the unknown parameters by maximum likelihood. Minimax rates of convergence in terms of the Kullback-Leibler information divergence are obtained.

AMS 2000 subject classifications. Primary 62G05, 62G07; Secondary 58G25.
Keywords. Kullback-Leibler information, maximum likelihood, minimax rates of convergence, spherical harmonics.

1. INTRODUCTION

Approximation of log-densities has long been part of the statistical literature. For the Euclidean case, the idea dates as far back as Neyman (1937) where this method is used for assessing goodness-of-fit, while much of the theory is later developed in Crain (1974, 1976a, 1976b, 1977). By assuming a compact domain along with positivity of the density in question, one estimates the probability density by finite dimensional densities using some collection of basis functions.

Received August 2003; accepted March 2004.
Research of the second author was supported in part by NSERC Canada, OGP46204.
Research of the third author was supported (in part) by KOSEF through Statistical Research Center for Complex Systems at Seoul National University.
Research of the fourth author was supported by Inha University Research Grant (INHA-31630).
[1]Department of Statistics, Duksung Women's University, Seoul 132-714, Korea (e-mail : jhuh@duksung.ac.kr)
[2]Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario N1G 2W1, Canada (e-mail : pkim@uoguelph.ca)
[3]Department of Statistics, Korea University, Seoul 136-701, Korea (e-mail : jykoo@korea.ac.kr)
[4]Department of Statistics, Inha University, Incheon 402-751, Korea (e-mail : jhpark@anova.inha.ac.kr)

More recently, interest is in rates of convergence as well as practical model selection. Stone and Koo (1986), Stone (1989, 1990), Barron and Sheu (1991), Kooperberg and Stone (1991, 1992), Stone *et al.* (1997), and Koo *et al.* (1998), develop logspline density estimation.

The main advantages of this approach are that this method usually fits the data well with a small number of parameters and the estimator can take advantage of whatever degree of smoothness necessary while at the same time being a probability density. The consequence of the latter is that log-density estimators can be used in a natural way for inferential statements. Furthermore, classical maximum likelihood properties hold and simulations seem to reveal that they perform quite well, even for very irregular densities. The main disadvantage nevertheless is the computational complexity, however, with current computing capabilities, this no longer is a formidable task.

In directional statistics where the sample space is the unit sphere, certain parametric exponential models are of considerable importance *i.e.*, the von Mises-Fisher and the Bingham distributions. These distributions are important in that a great deal of directional data can be analyzed by these parametric models (*cf.* Fisher *et al.*, 1993; Mardia and Jupp, 2000). Nevertheless, these models can often be inadequate and so a richer class of models may be called for. Indeed, the idea is to expand functions on a unit sphere in terms of some basis functions where, the above parametric models have an explicit meaning in terms of these basis functions. Furthermore, by including more terms, one can nest the von Mises-Fisher and the Bingham distributions within a family of distributions and therefore provide a richer class of exponential models.

We now provide a summary of what is to follow. In Section 2, we introduce spherical Fourier analysis for functions defined on the $(p-1)$-dimensional unit sphere, S^{p-1}. The spherical Fourier basis provides the necessary tool for higher order expansions of functions on the unit $(p-1)$-sphere. This section also describes the log-density estimation based on an exponential family with the spherical harmonic basis. We study consistency properties of the log-density estimators through the Kullback-Leibler divergence in Section 3 for the $(p-1)$-dimensional unit sphere. As has been the situation in the Euclidean case, the difficulty of estimation depends on the smoothness of the underlying density. We extend the ideas from the Euclidean case to the directional framework. Calculation of the upper and lower bounds over Sobolev classes will also be made. It is shown that the upper and lower rates of convergence (with respect to the Kullback-Leibler divergence) are the same, therefore the convergence rate of the log-density estimator is minimax. All proofs are contained in Section 4.

2. DIRECTIONAL EXPONENTIAL FAMILY MODEL

We initially provide a brief discussion of the Fourier transform and its inverse for functions defined on S^{p-1}. A more thorough summary of spherical Fourier analysis can be found in Müller (1998). Other statistical works that use expansions in spherical harmonics include Giné (1975), Wahba (1981), Hendriks (1990), Healy and Kim (1996), Healy *et al.* (1998), and Kim and Koo (2000).

For some $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^t \in S^{p-1}$, the $p-1$ spherical coordinates can be represented by

$$\begin{aligned}
\omega_1 &= \sin\theta_{p-1} \cdots \sin\theta_2 \sin\theta_1 \\
\omega_2 &= \sin\theta_{p-1} \cdots \sin\theta_2 \cos\theta_1 \\
&\vdots \\
\omega_{p-1} &= \sin\theta_{p-1} \cos\theta_{p-2} \\
\omega_p &= \cos\theta_{p-1}
\end{aligned}$$

where $\theta_1 \in [0, 2\pi)$, $\theta_j \in [0, \pi)$ for $j = 2, \ldots, p-1$ and superscript t denotes transpose. The normalized invariant measure is

$$d\omega = \frac{\Gamma(p/2)}{2\pi^{p/2}} \sin^{p-2}\theta_{p-1} \cdots \sin\theta_2 d\theta_1 \cdots d\theta_{p-1}, \tag{2.1}$$

where $\Gamma(\cdot)$ denotes the gamma function.

Let $C_r^\mu(t)$, $t \in [-1, 1]$ be a polynomial of degree r determined by the power series

$$(1 - 2t\alpha + \alpha^2)^{-\mu} = \sum_{r=0}^{\infty} C_r^\mu(t)\alpha^r.$$

One notices that $C_r^{1/2}(t)$ are the classical Legendre polynomials. Thus for general μ, these polynomials are generalizations of the classical Legendre polynomials and are called the Gegenbauer (ultraspherical) polynomials.

For $\mathbf{k} = (k_1, k_2, \ldots, k_{p-2})$, let $\mathcal{K}_\ell = \{\ell \geq k_1 \geq k_2 \geq \cdots \geq k_{p-2} \geq 0\}$. Define

$$Y_{\mathbf{k}}^{\ell,i} = A_{\mathbf{k}}^\ell \prod_{j=1}^{p-2} C_{k_{j-1}-k_j}^{k_j+(p-j-1)/2}(\cos\theta_{p-j})(\sin\theta_{p-j})^{k_j} Y_i^{k_{p-2}}(\theta_1), \tag{2.2}$$

where

$$Y_1^h(\psi) = C_h^0(\cos\psi), \quad Y_2^h(\psi) = \sqrt{2}\sin\psi C_{h-1}^1(\cos\psi),$$

$$Y_2^0(\psi) = 0 \text{ and } [A_{\mathbf{k}}^\ell]^2 = \frac{1}{\Gamma(p/2)} \prod_{j=1}^{p-2} \frac{\Gamma(k_j + (p-j+1)/2)}{\Gamma(k_j + (p-j)/2)}.$$

The collection

$$\{Y_{0,\ldots,0}^{0,1}, Y_{\mathbf{k}}^{\ell,i} : \mathbf{k} \in \mathcal{K}_\ell,\ \ell \geq 1,\ i = 1,2\}, \tag{2.3}$$

are the eigenfunctions of Δ, the Laplace-Beltrami operator on S^{p-1}, $p \geq 3$ and

$$\Delta Y_{\mathbf{k}}^{\ell,i} = \lambda_\ell Y_{\mathbf{k}}^{\ell,i}, \text{ where } \lambda_\ell = \ell(\ell + p - 2),\ \ \ell \geq 0.$$

Thus each $\ell \geq 0$, determines the eigenspace $\mathcal{E}_\ell$, where

$$\dim \mathcal{E}_\ell = \frac{(2\ell + p - 2)(\ell + p - 3)!}{\ell!(p-2)!}.$$

Collectively, (2.3) is called the spherical harmonics for $L^2(S^{p-1})$ and (2.3) forms a complete orthonormal basis (*cf.* Xu, 1997).

We will need the addition formula for functions on S^{p-1}. For $\boldsymbol{\omega}, \boldsymbol{\nu} \in S^{p-1}$

$$\sum_{\mathbf{k} \in \mathcal{K}_\ell} \left[Y_{\mathbf{k}}^{\ell,1}(\boldsymbol{\omega}) Y_{\mathbf{k}}^{\ell,1}(\boldsymbol{\nu}) + Y_{\mathbf{k}}^{\ell,2}(\boldsymbol{\omega}) Y_{\mathbf{k}}^{\ell,2}(\boldsymbol{\nu}) \right] = \frac{\ell + (p-2)/2}{(p-2)/2} C_\ell^{(p-2)/2}(\boldsymbol{\omega}^t \boldsymbol{\nu}). \tag{2.4}$$

In particular, any point on the 2-dimensional sphere, S^2 can be represented by

$$\boldsymbol{\omega} = (\cos\phi \sin\theta, \sin\phi \sin\theta, \cos\theta)^t,$$

where $\theta \in [0, \pi)$ and $\phi \in [0, 2\pi)$. Let

$$Y_q^\ell(\boldsymbol{\omega}) = \begin{cases} (-1)^q \sqrt{\dfrac{(2\ell+1)(\ell-q)!}{2\pi(\ell+q)!}} P_q^\ell(\cos\theta) \cos(q\phi), & q = 1, 2, \ldots, \ell, \\[2ex] \sqrt{\dfrac{(2\ell+1)}{4\pi}} P_0^\ell(\cos\theta), & q = 0, \\[2ex] (-1)^q \sqrt{\dfrac{(2\ell+1)(\ell-q)!}{2\pi(\ell+q)!}} P_q^\ell(\cos\theta) \sin(q\phi), & q = -1, -2, \ldots, -\ell, \end{cases} \tag{2.5}$$

where P_q^ℓ are the associated Legendre functions and we use the unnormalized invariant measure $d\boldsymbol{\omega} = \sin\theta\, d\theta d\phi$. Then,

$$\{Y_q^\ell : -\ell \leq q \leq \ell,\ \ell = 0, 1, \ldots\}$$

is a complete orthonormal basis for $L^2(S^2)$. For each fixed $\ell \geq 0$, the eigenspace $\mathcal{E}_\ell$ is the span of $\{Y_q^\ell : -\ell \leq q \leq \ell\}$ which is called an invariant irreducible subspace (see Beran, 1979). It is worth noting that in (2.5) when $q = 0$, then Y_0^ℓ is independent of ϕ. Furthermore, there is antipodal symmetry when ℓ is even since the exponent will involve even powers. In particular, the function is the same on one half of the sphere as well as the other half.

Let $f : S^{p-1} \to \mathbb{R}$. We define the spherical Fourier transform of f by

$$\hat{f}_{\mathbf{k}}^{\ell,i} = \int_{S^{p-1}} f(\omega) Y_{\mathbf{k}}^{\ell,i}(\omega) d\omega$$

for $\mathbf{k} \in \mathcal{K}_\ell$, $\ell \geq 0$ and $i = 1, 2$ where $d\omega$ is defined according to (2.1). The spherical inversion can be obtained by

$$f(\omega) = \sum_{\ell \geq 0} \sum_{\mathbf{k} \in \mathcal{K}_\ell} \sum_{i=1}^{2} \hat{f}_{\mathbf{k}}^{\ell,i} Y_{\mathbf{k}}^{\ell,i}(\omega)$$

for $\omega \in S^{p-1}$. We note that spherical inversion should be interpreted in the L^2 sense although it can hold pointwise with additional smoothness conditions.

2.1. Likelihood estimation

Let $\mathcal{D}_m = \{(\ell, \mathbf{k}, i) : \ell = 1, \ldots, m, \ \mathbf{k} \in \mathcal{K}_\ell, \ i = 1, 2\}$ and define $\mathcal{B}$ to be the collection of all $|\mathcal{D}_m|$-dimensional vectors $\boldsymbol{\beta}$, where for a given finite set, $|\cdot|$ denote its cardinality. An exponential family of densities based on the spherical harmonic basis (2.2) is defined by

$$f(\omega; \boldsymbol{\beta}) = \exp\{s(\omega; \boldsymbol{\beta}) - \psi(\boldsymbol{\beta})\}, \quad \boldsymbol{\beta} \in \mathcal{B}, \tag{2.6}$$

where

$$s(\omega; \boldsymbol{\beta}) = \sum_{(\ell, \mathbf{k}, i) \in \mathcal{D}_m} \beta_{\mathbf{k}}^{\ell,i} Y_{\mathbf{k}}^{\ell,i}(\omega) \quad \text{and} \quad \psi(\boldsymbol{\beta}) = \log\Big\{ \int_{S^{p-1}} \exp\{s(\omega; \boldsymbol{\beta})\} d\omega \Big\}.$$

Here the constant function $Y_{0,\ldots,0}^{0,1} = 1$ is absorbed in $\psi(\boldsymbol{\beta})$. For notational convenience, let $s(\boldsymbol{\beta})$ and $f(\boldsymbol{\beta})$ denote the functions $s(\cdot\,; \boldsymbol{\beta})$ and $f(\cdot\,; \boldsymbol{\beta})$, $\boldsymbol{\beta} \in \mathcal{B}$, respectively. Furthermore, let $\mathcal{S}_m$ be the span of $\{Y_{\mathbf{k}}^{\ell,i} : (\ell, \mathbf{k}, i) \in \mathcal{D}_m\}$. We note that (2.6) is the usual definition for an exponential family, see for example Brown (1986). For the directional case, this construction is formalized in Beran (1979), see also Chapter 9 in Diaconis (1988).

Let $X_1, \dots, X_n$ be a random sample from some density f on S^{p-1}. The log-likelihood function corresponding to the exponential family is defined by

$$\mathcal{L}(\beta) = \sum_{j=1}^{n} \sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m} \beta_{\mathbf{k}}^{\ell,i} Y_{\mathbf{k}}^{\ell,i}(X_j) - n\psi(\beta), \quad \beta \in \mathcal{B}.$$

Let

$$\beta_{\text{ml}} = \underset{\beta\in\mathcal{B}}{\text{argmax}}\, \mathcal{L}(\beta)$$

be the maximum likelihood estimator (MLE) of $\beta \in \mathcal{B}$. Since the Hessian matrix of $\psi(\cdot)$ is strictly positive definite, $\mathcal{L}(\cdot)$ is a strictly concave function on $\mathcal{B}$; thus the MLE β_{ml} is unique if it exists. We set $f_{\text{ml}} = f(\beta_{\text{ml}})$ and refer to f_{ml} as the MLE of f. A formal list of likelihood properties are summarized in Beran (1979).

2.2. Entropy estimation

The relative entropy, or Kullback-Leibler (KL) divergence from density f_1 to density f_2 on S^{p-1} is defined by

$$D(f_1\|f_2) = \int_{S^{p-1}} f_1(\omega) \log \frac{f_1(\omega)}{f_2(\omega)} d\omega.$$

Let $\mathcal{S}_m$ be the linear span of $\{Y_{\mathbf{k}}^{\ell,i} : (\ell, \mathbf{k}, i) \in \mathcal{D}_m\}$. Define

$$\beta_{\text{kl}} = \underset{\beta\in\mathcal{B}}{\text{argmax}} D(f\|f(\beta)),$$

and set $f_{\text{kl}} = f(\beta_{\text{kl}})$. We will refer to f_{kl} as the information projection of f onto $\mathcal{S}_m$. The density f_{kl} (among the class of exponential densities) minimizes the KL-divergence to f. From Koo and Chung (1998), the information projection f_{kl} exists uniquely.

The KL-divergence decomposes into the sum of two terms which correspond to approximation error and estimation error, respectively. It is characterized by an orthogonality-like relation

$$D(f\|f(\beta)) = D(f\|f_{\text{kl}}) + D(f_{\text{kl}}\|f(\beta)) \tag{2.7}$$

valid for all densities (2.6).

3. ASYMPTOTIC RESULTS

In this section, we demonstrate that the MLE is rate minimax in the Kullback-Leibler sense. We note that such results are important in physics, see for example Chandler and Gibson (1989) and Taijeron *et al.* (1994) where their requirements are for data analysis on hyperspheres.

We will use the following notation. For two positive sequences $\{a_n\}$ and $\{b_n\}$, let $a_n \ll b_n$ ($\ll_{\mathrm{p}}$) mean $a_n = O(b_n)$ (O_{p}) as $n \to \infty$. If $a_n \ll b_n$ ($\ll_{\mathrm{p}}$) and $b_n \ll a_n$ ($\ll_{\mathrm{p}}$) then denote this by $a_n \asymp b_n$ ($\asymp_{\mathrm{p}}$).

Let

$$\gamma_2 = \inf_{s\in\mathcal{S}_m} \big\|\log f - s\big\|_2 \ \text{ and } \ \gamma_\infty = \inf_{s\in\mathcal{S}_m} \big\|\log f - s\big\|_\infty$$

be the L^2 and L^∞ error approximations for some positive function f by some $s \in \mathcal{S}_m$. We can establish an upper bound approximation error for $D(f\|f_{\mathrm{kl}})$ in terms of γ_2 under bounded conditions on f.

THEOREM 3.1. *Suppose there exists an $A > 0$ such that $\|s\|_\infty \leq A\|s\|_2$ for all $s \in \mathcal{S}_m$, $A\gamma_2 = o(1)$ and $\gamma_\infty < \infty$. If $M_1 > 0$ such that $M_1^{-1} \leq f \leq M_1$ then for n sufficiently large*

(i) $D\big(f\|f_{\mathrm{kl}}\big) \leq \dfrac{M_1}{2}\mathrm{e}^{\gamma_\infty}\gamma_2^2$;

(ii) *If* $\dfrac{A^2 m^{p-1}}{n} = o(1)$, *then* $D(f_{\mathrm{kl}}\|f_{\mathrm{ml}}) \ll_{\mathrm{p}} \dfrac{m^{p-1}}{n}$.

For a differentiable function $f : S^{p-1} \to \mathbb{R}$ stronger results can be obtained. We would like to present the discussion in terms of the parameter space being some function class of continuously differentiable functions on S^{p-1}. The Sobolev space of order $\nu > (p-1)/2$, is defined to be the collection of continuously differentiable functions up to and including ν with the ν^{th} derivative being square integrable. For some function $h = \sum_{\ell,\mathbf{k},i} \hat{h}_{\mathbf{k}}^{\ell,i} Y_{\mathbf{k}}^{\ell,i}$, denote its Sobolev norm of order ν by

$$\big\|h\big\|_\nu^2 = \sum_{\ell,\mathbf{k},i}(1+\lambda_\ell)^\nu \big|\hat{h}_{\mathbf{k}}^{\ell,i}\big|^2.$$

For some constant $M > 0$ define $\mathcal{F}_\nu(M)$ by

$$\mathcal{F}_\nu(M) = \big\{f : \log f \text{ is Sobolev of order } \nu > (p-1)/2 \text{ and } \|\log f\|_\nu \leq 1 + M\big\}.$$

Consider an unknown distribution P_f depending on the density function $f \in \mathcal{F}_\nu(M)$. Suppose $\{b_n\}$ is some sequence of positive numbers. This sequence is called a lower bound for f if

$$\lim_{c\to 0}\liminf_{n}\inf_{\tilde f}\sup_{f\in\mathcal{F}_\nu(M)} P_f\big(D(f\|\tilde f)\ge cb_n\big)=1,$$

where the last infimum is over all possible estimators $\tilde f$ based on $X_1,\dots,X_n$. Alternatively, the sequence in question is said to be an upper bound for f if there is a sequence of estimators $\tilde f$ such that

$$\lim_{c\to\infty}\limsup_{n}\sup_{f\in\mathcal{F}_\nu(M)} P_f\big(D(f\|\tilde f)\ge cb_n\big)=0.$$

The sequence of numbers $\{b_n\}$ is called the optimal rate of convergence for f if it is both a lower bound and an upper bound with the associated estimators $\tilde f$, being called asymptotically optimal.

We have the following results.

THEOREM 3.2. *If $f\in\mathcal{F}_\nu(M)$ with $\nu>(p-1)/2$, then*

$$D(f\|f_{\mathrm{ml}})\ll_{\mathrm{p}} n^{-2\nu/(2\nu+p-1)}$$

where $m\asymp n^{1/(2\nu+p-1)}$, as $n\to\infty$.

THEOREM 3.3. *Suppose that $f\in\mathcal{F}_\nu(M)$ with $\nu>(p-1)/2$. Then*

$$n^{-2\nu/(2\nu+p-1)}\ll_{\mathrm{p}} D(f\|\tilde f)$$

where $\tilde f$ is any density estimator based on $X_1,\dots,X_n$, as $n\to\infty$.

We can now use this lower bound along with the upper bound to obtain optimal rates of convergence for f in the KL-divergence sense. Putting together Theorems 3.2 and 3.3 we immediately get the following.

COROLLARY 3.1. *Suppose that $f\in\mathcal{F}_\nu(M)$ with $\nu>(p-1)/2$. Then the optimal rate of convergence for estimating f equals $n^{-2\nu/(2\nu+p-1)}$ and f_{ml} is asymptotically optimal.*

4. PROOFS

Let $\|\boldsymbol\beta\|$ be the Euclidean norm of a vector $\boldsymbol\beta\in\mathcal{B}$. For $g=\log f$, let

$$s(g)=\sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m^0}\hat g_{\mathbf{k}}^{\ell,i}Y_{\mathbf{k}}^{\ell,i},$$

where $\mathcal{D}_m^0=\mathcal{D}_m\cup\big\{\ell=0,\;\;\mathbf{k}=(0,\dots,0),\;\;i=1\big\}$.

4.1. Proof of Theorem 3.1

The first task is to show that $\boldsymbol{\beta}_{\mathrm{kl}}$ exists with $\int f_{\mathrm{kl}} Y_{\mathbf{k}}^{\ell,i} = \int f Y_{\mathbf{k}}^{\ell,i}$ for all $(\ell,\mathbf{k},i) \in \mathcal{D}_m$ when n is sufficiently large. Set $\theta^* = (\int Y_{\mathbf{k}}^{\ell,i} f)$ and $\theta = (\int Y_{\mathbf{k}}^{\ell,i} f(\boldsymbol{\delta}))$, where $\boldsymbol{\delta} = (\hat{g}_{\mathbf{k}}^{\ell,i})$ and $(\ell,\mathbf{k},i) \in \mathcal{D}_m$. The entries in the vector $\theta^* - \theta$ are seen to be coefficients in the $L^2(S^{p-1})$ orthonormal projection of $f - f(\boldsymbol{\delta})$ onto $\mathcal{S}_m$. By Bessel's inequality, the boundedness of f, we have

$$\begin{aligned}
\|\theta^* - \theta\|^2 &\le \|f - f(\boldsymbol{\delta})\|_2^2 \\
&\le M_1^2 \int \frac{(f - f(\boldsymbol{\delta}))^2}{f} \\
&\le M_1^2 \|g - s(g)\|_2^2 \exp\left\{2\|g - s(g)\|_\infty - 2\left(\hat{g}_{0,\dots,0}^{0,1} + \psi(\boldsymbol{\delta})\right)\right\} \\
&\le M_1^2 \mathrm{e}^{4\gamma_\infty} \gamma_2^2.
\end{aligned}$$

For the last inequality we have used the fact that $\left|\psi(\boldsymbol{\delta}) + \hat{g}_{0,\dots,0}^{0,1}\right| \le \|g - s(g)\|_\infty$, since $\psi(\boldsymbol{\delta}) + \hat{g}_{0,\dots,0}^{0,1} = \log\left\{\int \exp(s(g) - g) f\right\}$. From this same fact we have $\|\log f/f(\boldsymbol{\delta})\|_\infty \le 2\|g - s(g)\|_\infty = 2\gamma_\infty$ and together with $M_1^{-1} \le f \le M_1$, we obtain

$$\|\log f(\boldsymbol{\delta})\|_\infty \le \log M_1 + 2\gamma_\infty.$$

Now if $M_1 \mathrm{e}^{2\gamma_\infty} \gamma_2 \le 1/(4\mathrm{e}bA)$, that is, if $A\gamma_2 = o(1)$, then we may conclude that the solution $\boldsymbol{\beta}_{\mathrm{kl}}$ to the equation $\left(\int Y_{\mathbf{k}}^{\ell,i} f(\boldsymbol{\beta})\right) = \theta^*$ exists and that

$$\|\log f_{\mathrm{kl}}/f(\boldsymbol{\delta})\|_\infty \le \epsilon,$$

where $\epsilon = 4M_1^2 \exp(4\gamma_\infty + 1) A\gamma_2$. So by the triangle inequality, we obtain $\|\log f/f_{\mathrm{kl}}\|_\infty \le 2\gamma_\infty + \epsilon$, and

$$\|\log f_{\mathrm{kl}}\|_\infty \le \log M_1 + 2\gamma_\infty + \epsilon.$$

Therefore, by boundedness of f, we have

$$D(f\|f_{\mathrm{kl}}) \le D(f\|f(\boldsymbol{\delta})) \le \frac{1}{2} \mathrm{e}^{\|g - s(g)\|_\infty} M_1 \|g - s(g)\|_2^2 \le \frac{1}{2} M_1 \mathrm{e}^{\gamma_\infty} \gamma_2^2.$$

For the proof of the second part of Theorem 3.1, we have to show that $D(f_{\mathrm{kl}}\|f_{\mathrm{ml}})$ is small with high probability. Let $\theta^* = \left(\int Y_{\mathbf{k}}^{\ell,i} f_{\mathrm{kl}}\right)$ which is the same as $\left(\int Y_{\mathbf{k}}^{\ell,i} f\right)$. Also let $\theta_{\mathrm{ml}} = \left(\sum_{j=1}^n Y_{\mathbf{k}}^{\ell,i}(X_j)/n\right)$. Whenever a solution $\boldsymbol{\beta}_{\mathrm{ml}} \in \mathcal{B}$ to the equation $\left(\int Y_{\mathbf{k}}^{\ell,i} f(\boldsymbol{\beta})\right) = \theta_{\mathrm{ml}}$ exists, we recognize $f_{\mathrm{ml}} = f(\boldsymbol{\beta}_{\mathrm{ml}})$ as the MLE. With these choices

$$\|\boldsymbol{\beta}_{\mathrm{ml}} - \boldsymbol{\beta}_{\mathrm{kl}}\|^2 = \sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m} \left\{\frac{1}{n}\sum_{j=1}^n \left(Y_{\mathbf{k}}^{\ell,i}(X_j) - E_f Y_{\mathbf{k}}^{\ell,i}(X_j)\right)\right\}^2.$$

By Chebyshev's inequality and the boundedness assumption on f, for some $C_1 > 0$ we have

$$\begin{aligned}
&P_f\Big[\sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m}\Big\{\frac{1}{n}\sum_{j=1}^{n}\big(Y_{\mathbf{k}}^{\ell,i}(X_j)-E_fY_{\mathbf{k}}^{\ell,i}(X_j)\big)\Big\}^2 > C_1|\mathcal{D}_m|/n\Big]\\
&< \frac{n}{C_1|\mathcal{D}_m|}E_f\Big[\sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m}\Big\{\frac{1}{n}\sum_{j=1}^{n}\big(Y_{\mathbf{k}}^{\ell,i}(X_j)-E_fY_{\mathbf{k}}^{\ell,i}(X_j)\big)\Big\}^2\Big]\\
&< \frac{1}{C_1|\mathcal{D}_m|}\sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m}E_f\big(Y_{\mathbf{k}}^{\ell,i}\big)^2\\
&\le \frac{M_1}{C_1}.
\end{aligned}$$

Then, $\|\beta_{\mathrm{ml}}-\beta_{\mathrm{kl}}\|^2 \ll_{\mathrm{p}} |\mathcal{D}_m|/n$. If $\big(C_1|\mathcal{D}_m|/n\big)^{1/2} \le 1/(4ebA)$, or equivalently, if $A\sqrt{|\mathcal{D}_m|/n} = o(1)$, then except on the set above (whose probability is less than M_1/C_1) it can be shown β_{ml} exists and $D(f_{\mathrm{kl}}\|f_{\mathrm{ml}}) = O_{\mathrm{p}}(|\mathcal{D}_m|/n)$ except on a set of probability less than M_1/C. This completes the proof of Theorem 3.1.

4.2. Proof of Theorem 3.2.

First we show that if $f \in \mathcal{F}_\nu(M)$, then the boundedness condition on f is satisfied. Write $g = \sum \hat{g}_{\mathbf{k}}^{\ell,i}Y_{\mathbf{k}}^{\ell,i}$. Observe that

$$\begin{aligned}
|g(\omega)|^2 &\le \Big(\sum_{(\ell,\mathbf{k},i)}(1+\lambda_\ell)^{\nu}|\hat{g}_{\mathbf{k}}^{\ell,i}|^2\Big)\Big(\sum_{(\ell,\mathbf{k},i)}(1+\lambda_\ell)^{-\nu}|Y_{\mathbf{k}}^{\ell,i}|^2\Big)\\
&\le (1+M)\sum_{l}(1+\lambda_\ell)^{-\nu}\dim\mathcal{E}_\ell.
\end{aligned}$$

In the above, we use the addition formula (2.4). Since $\nu > (p-1)/2$, the series $\sum_\ell(1+\lambda_\ell)^{-\nu}\dim\mathcal{E}_\ell$ converges thus giving us a bound.

To prove Theorem 3.2, we need the bounds A, γ_2 and γ_∞. To determine A, choose any element $s = \sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m^0}\beta_{\mathbf{k}}^{\ell,i}Y_{\mathbf{k}}^{\ell,i}$ in $\mathcal{S}_m$. By the Cauchy-Schwarz and the Parseval inequalities, we have that, uniformly in $\omega \in S^{p-1}$,

$$|s(\omega)| \le \Big(\sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m^0}|Y_{\mathbf{k}}^{\ell,i}(\omega)|^2\Big)^{1/2}\Big(\sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m^0}|\beta_{\mathbf{k}}^{\ell,i}|^2\Big)^{1/2}$$

$$\leq \Big(\sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m^0} \big|Y_{\mathbf{k}}^{\ell,i}(\omega)\big|^2 \Big)^{1/2} \|s\|_2$$
$$\ll \Big(\sum_{l=0}^{m} \dim\mathcal{E}_\ell \Big)^{1/2} \|s\|_2$$
$$\leq m^{(p-1)/2} \|s\|_2.$$

Let $\mathcal{D}_m^{0,c} = \big\{(\ell,\mathbf{k},i) : \ell > m, \mathbf{k} \in \mathcal{K}_\ell, i = 1,2\big\}$. Since

$$(1+\lambda_m)^\nu \sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m^{0,c}} |\hat{g}_{\mathbf{k}}^{\ell,i}|^2 \leq \sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m^{0,c}} (1+\lambda_\ell)^\nu |\hat{g}_{\mathbf{k}}^{\ell,i}|^2 < 1 + M,$$

we have the bound on γ_2 as follows

$$\gamma_2^2 \leq \sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m^{0,c}} |\hat{g}_{\mathbf{k}}^{\ell,i}|^2 < (1+M)(1+\lambda_m)^{-\nu} \ll m^{-2\nu}.$$

It follows from the Cauchy-Schwarz inequality and the addition formula (2.4) that the error γ_∞^2 is bounded by

$$\Big| \sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m^{0,c}} \hat{g}_{\mathbf{k}}^{\ell,i} \Big|^2 \leq \Big(\sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m^{0,c}} |Y_{\mathbf{k}}^{\ell,i}(\omega)|^2 (1+\lambda_\ell)^{-\nu} \Big) \Big(\sum_{(\ell,\mathbf{k},i)\in\mathcal{D}_m^{c}} (1+\lambda_\ell)^\nu |\hat{g}_{\mathbf{k}}^{\ell,i}|^2 \Big)$$
$$\ll (1+M) \sum_{l=m+1}^{\infty} (1+\lambda_\ell)^{-\nu} \dim\mathcal{E}_\ell$$
$$\ll m^{-2\nu+p-1}. \tag{4.1}$$

Choose $m \asymp n^{1/(2\nu+p-1)}$. Then $\gamma_\infty \ll m^{-2\nu+p-1)} = o(1)$, $\gamma_2 \asymp n^{-2\nu/(2\nu+p-1)}$, $A\gamma_2 \ll m^{(p-1)/2-\nu}$, $|\mathcal{D}_m|/n \asymp n^{-2\nu/(2\nu+p-1)}$ and $A\sqrt{|\mathcal{D}_m|/n} \asymp m^{p-1}/\sqrt{n} = n^{((p-1)/2-\nu)/2(\nu+(p-1)/2)}$. Consequently, from Theorem 3.1 and the Kullback-Leibler decomposition (2.7), the proof of Theorem 3.2 is now complete.

4.3. Proof of Theorem 3.3.

The approach we will follow is to first specify a subproblem, followed by using Fano's lemma to calculate the difficulty of the subproblem. The lower bound will then appear.

Let N be a positive integer depending on n and define

$$V = \big\{(\ell,\mathbf{k},i) : \ell = N+1,\ldots,2N, \mathbf{k} \in \mathcal{K}_\ell, i = 1,2\big\}.$$

Let $\tau_{\mathbf{k}}^{\ell,i}$ be either 0 or 1 for $(\ell,\mathbf{k},i) \in V$ and define $\boldsymbol{\tau} = \{\tau_{\mathbf{k}}^{\ell,i} : (\ell,\mathbf{k},i) \in V\}$. Consider the function

$$f_{\boldsymbol{\tau}} = \exp\Big\{M_2 N^{-\nu-(p-1)/2} \sum_{(\ell,\mathbf{k},i)\in V} \tau_{\mathbf{k}}^{\ell,i} Y_{\mathbf{k}}^{\ell,\mathbf{k},i} - \psi(\boldsymbol{\tau})\Big\} \tag{4.2}$$

where M_2 is a constant to be determined below. Finally, let

$$\mathcal{F}_{\nu,n}(M) = \big\{f_{\boldsymbol{\tau}} : \boldsymbol{\tau} \in \{0,1\}^{|V|}\big\},$$

and assume that $N \to \infty$ as $n \to \infty$.

Applying the Sobolev norm to (4.2), we get

$$\begin{aligned}
\Big\|M_2 N^{-\nu-(p-1)/2} \sum_{(\ell,\mathbf{k},i)\in V} \tau_{\mathbf{k}}^{\ell,i} Y_{\mathbf{k}}^{\ell,i}\Big\|_{\nu}^{2} &= M_2^2 N^{-2\nu-(p-1)} \sum_{(\ell,\mathbf{k},i)\in V} (1+\lambda_\ell)^{\nu} (\tau_{\mathbf{k}}^{\ell,i})^2 \\
&\le M_2^2 N^{-2\nu-(p-1)} \sum_{(\ell,\mathbf{k},i)\in V} (1+\lambda_\ell)^{\nu} \\
&= M_2^2 N^{-2\nu-(p-1)} \sum_{\ell=N+1}^{2N} (1+\lambda_\ell)^{\nu} \dim\mathcal{E}_\ell \\
&\le C_1 M_2^2,
\end{aligned} \tag{4.3}$$

$$\begin{aligned}
|\psi(\boldsymbol{\tau})| &= \Big|\log\Big[\int \exp\Big\{M_2 N^{-\nu-(p-1)/2} \sum_{(\ell,\mathbf{k},i)\in V} (\tau_{\mathbf{k}}^{\ell,i})^2 Y_{\mathbf{k}}^{\ell,i}\Big\}\Big]\Big| \\
&\le \Big\|M_2 N^{-\nu-(p-1)/2} \sum_{(\ell,\mathbf{k},i)\in V} \tau_{\mathbf{k}}^{\ell,i} Y_{\mathbf{k}}^{\ell,i}\Big\|_{\infty} \\
&\le C_1 M_2.
\end{aligned} \tag{4.4}$$

In the above, by using the same argument as in (4.1) together with (4.3) and (4.4) we have

$$\begin{aligned}
\|\log f_{\boldsymbol{\tau}}\|_{\nu}^{2} &= \Big\|M_2 N^{-\nu-(p-1)/2} \sum_{(\ell,\mathbf{k},i)\in V} \tau_{\mathbf{k}}^{\ell,i} Y_{\mathbf{k}}^{\ell,i} - \psi(\boldsymbol{\tau}) Y_{0,\ldots,0}^{0,1}\Big\|_{\nu}^{2} \\
&= \psi(\boldsymbol{\tau})^2 + \sum_{(\ell,\mathbf{k},i)\in V} (1+\lambda_\ell)^{\nu} M_2^2 N^{-2\nu-p+1} (\tau_q^l)^2 \\
&\le C_1 M_2^2 \\
&\le (1+M)^2
\end{aligned}$$

if M_2 is chosen sufficiently small. Consequently, we have shown that $f_{\boldsymbol{\tau}} \in \mathcal{F}_{\nu}(M)$, for sufficiently large n, so that

$$\mathcal{F}_{\nu,n}(M) \subset \mathcal{F}_{\nu}(M).$$

Now let us write $g_r = M_2 N^{-\nu-(p-1)/2} \sum_{(\ell,\mathbf{k},i)\in V} \tau_{\mathbf{k}}^{\ell,i}(r) Y_{\mathbf{k}}^{\ell,i} - \psi(\boldsymbol{\tau}(r))$, for $r = 1, 2$, such that $g_1 \neq g_2$. Since $\{Y_{0,\ldots,0}^{0,1}\} \cup \{Y_{\mathbf{k}}^{\ell,i} : (\ell,\mathbf{k},i) \in V\}$ are orthogonal,

$$\begin{aligned}\|g_1 - g_2\|_2^2 &= M_2^2 N^{-2\nu-(p-1)} \Big\| \sum_{(\ell,\mathbf{k},i)\in V} (\tau_{\mathbf{k}}^{\ell,i}(1) - \tau_{\mathbf{k}}^{\ell,i}(2)) Y_{\mathbf{k}}^{\ell,i} \Big\|_2^2 \\ &\quad + \big(\psi(\boldsymbol{\tau}(1)) - \psi(\boldsymbol{\tau}(2))\big)^2 \\ &\geq C_1 M_2^2 N^{-2\nu-(p-1)}.\end{aligned}$$

For $f_r = \mathrm{e}^{g_r} \in \mathcal{F}_{\nu,n}(M)$, it follows from the boundedness condition that

$$D(f_1 \| f_2) \geq \frac{1}{2} \mathrm{e}^{-2\log M_1} M_1^{-1} \|g_1 - g_2\|_2^2 \geq C_1 N^{-2\nu-(p-1)}.$$

It follows from Lemma 3.1 of Koo (1993) that there exists a subset $\mathcal{F}_{\nu,n}^0(M)$ of $\mathcal{F}_{\nu,n}(M)$ such that

$$D(f_1 \| f_2) > C_1 N^{-\nu},\ f_1 \neq f_2 \in \mathcal{F}_{\nu,n}^0(M) \text{ and } \log\big(|\mathcal{F}_{\nu,n}^0(M)| - 1\big) \geq C_1 N^{p-1}$$

for sufficiently large n. Since $\|f\|_\infty \leq M_1$ for large n, we have

$$\begin{aligned}D(f_1 \| f_2) &\leq \mathrm{e}^{\|\log f_1/f_2\|_\infty} \int f_1 \Big(\frac{\log f_1}{f_2}\Big)^2 \\ &\leq C_1 \Big\| \sum_{(\ell,\mathbf{k},i)\in V} M_2 N^{-\nu-(p-1)/2} \tau_{\mathbf{k}}^{\ell,i} Y_{\mathbf{k}}^{\ell,i} \Big\|_2^2 \\ &\leq C_1 N^{-2\nu}.\end{aligned}$$

By Fano's lemma, see Birgé (1983), Yatracos (1988) and Koo (1993), if $\tilde{f}$ is any estimator of f, then

$$\begin{aligned}\sup_{f\in\mathcal{F}_\nu(M)} P_f\big(D(f\|\tilde{f}) > cN^{-\nu}\big) &\geq \sup_{f\in\mathcal{F}_{\nu,n}(M)} P_f\big(D(f\|\tilde{f}) > cN^{-\nu}\big) \\ &\geq \sup_{f\in\mathcal{F}_{\nu,n}^0(M)} P_f\big(D(f\|\tilde{f}) > cN^{-\nu}\big) \\ &\geq 1 - \frac{nD(f_1\|f_2) + \log 2}{\log(|\mathcal{F}_{\nu,n}^0(M)| - 1)}.\end{aligned}$$

It therefore follows that

$$\sup_{f\in\mathcal{F}_\nu(M)} P_f\big(D(f\|f_{\mathrm{ml}}) > cN^{-\nu}\big) > 0$$

for $c > 0$ as $n \to \infty$. Now choose N such that $N \asymp n^{1/(2\nu+(p-1))}$ to obtain the desired result.

REFERENCES

BARRON, A. R. AND SHEU, C. H. (1991). "Approximation of density functions by sequences of exponential families", *The Annals of Statistics*, **19**, 1347–1369.

BERAN, R. (1979). "Exponential models for directional data", *The Annals of Statistics*, **7**, 1162–1178.

BIRGÉ, L. (1983). "Approximation dans les espaces metriques et theorie de l'estimation", *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **65**, 181–237.

BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families*, Institute of Mathematical Statistics, Hayward.

CHANDLER, C. AND GIBSON, A. G. (1989). "N-body quantum scattering theory in two Hilbert spaces. V. Computation strategy", *Journal of Mathematical Physics*, **30**, 1533–1544.

CRAIN, B. R. (1974). "Estimation of distributions using orthogonal expansions", *The Annals of Statistics*, **2**, 454-463.

CRAIN, B. R. (1976a). "Exponential models, maximum likelihood estimation, and the Haar condition", *Journal of the American Statistical Association*, **71**, 737–740.

CRAIN, B. R. (1976b). "More on estimation of distributions using orthogonal expansions", *Journal of the American Statistical Association*, **71**, 741–745.

CRAIN, B. R. (1977). "An information theoretic approach to approximating a probability distribution", *SIAM Journal on Applied Mathematics*, **32**, 339–346.

DIACONIS, P. (1988). *Group Representations in Probability and Statistics*, Institute of Mathematical Statistics, Hayward.

FISHER, N. I., LEWIS, T. AND EMBLETON, B. J. (1993). *Statistical Analysis of Spherical Data*, Cambridge University Press, Cambridge.

GINÉ, E. (1975). "Invariant tests for uniformity on compact Riemannian manifolds based on Sobolev norms", *The Annals of Statistics*, **3**, 1243–1266.

HEALY, D. M., HENDRIKS, H. AND KIM, P. T. (1998). "Spherical deconvolution", *Journal of Multivariate Analysis*, **67**, 1–22.

HEALY, D. M. AND KIM, P. T. (1996). "An empirical Bayes approach to directional data and efficient computation on the sphere", *The Annals of Statistics*, **24**, 232–254.

HENDRIKS, H. (1990). "Nonparametric estimation of a probability density on a Riemannian manifold using Fourier expansions", *The Annals of Statistics*, **18**, 832–849.

KIM, P. T. AND KOO, J.-Y. (2000). "Directional mixture models and optimal estimation of the mixing density", *The Canadian Journal of Statistics*, **28**, 383–398.

KOO, J.-Y. (1993). "Optimal rates of convergence for nonparametric statistical inverse problems", *The Annals of Statistics*, **21**, 590–599.

KOO, J.-Y. AND CHUNG, H. Y. (1998). "Log-density estimation in linear inverse problems", *The Annals of Statistics*, **26**, 335–362.

KOO, J.-Y., KOOPERBERG, C. AND PARK, J. (1998). "Logspline density estimation under censoring and truncation", *Scandinavian Journal of Statistics*, **26**, 87–105.

KOOPERBERG, C. AND STONE, C. J. (1991). "A study of logspline density estimation", *Computational Statistics and Data Analysis*, **12**, 327–347.

KOOPERBERG, C. AND STONE, C. J. (1992). "Logspline density estimation for censored data", *Journal of Computational and Graphical Statistics*, **1**, 301–328.

MARDIA, K. V. AND JUPP, P. E. (2000). *Directional Statistics*, John Wiley & Sons, New York.

MÜLLER, C. (1998). *Analysis of Spherical Symmetries in Euclidean Spaces*, Springer-Verlag, New York.

NEYMAN, J. (1937). " 'Smooth' test for goodness of fit", *Scandinavian Actuarial Journal*, **20**, 149–199.

STONE, C. J. (1989). "Uniform error bounds involving logspline models", In *Probability, Statistics, and Mathematics : Papers in Honor of Samuel Karlin* (T. W. Anderson, K. B. Athrya and D. L. Iglehart, eds.), 335–355, Academic Press, Boston.

STONE, C. J. (1990). "Large-sample inference for logspline models", *The Annals of Statistics*, **18**, 717–741.

STONE, C. J., HANSEN, M. H., KOOPERBERG, C. AND TRUONG, Y. K. (1997) "Polynomial splines and their tensor products in extended linear modeling (with discussion)", *The Annals of Statistics*, **25**, 1371–1470.

STONE, C. J. AND KOO, C.-Y. (1986). "Logspline density estimation", *Contemporary Mathematics*, **59**, 1–15, American Mathematical Society, Providence.

TAIJERON, H. J., GIBSON, A. G. AND CHANDLER, C. (1994). "Spline interpolation and smoothing on hyperspheres", *SIAM Journal on Scientific and Statistical Computing*, **15**, 1111–1125.

WAHBA, G. (1981). "Spline interpolation and smoothing on the sphere", *SIAM Journal on Scientific and Statistical Computing*, **2**, 5–16.

XU, Y. (1997). "Orthogonal polynomials for a family of product weight functions on the spheres", *Canadian Journal of Mathematics*, **49**, 175–192.

YATRACOS, Y. G. (1988). "A lower bound on the error in nonparametric regression type problems", *The Annals of Statistics*, **16**, 1180–1187.