NONPARAMETRIC ESTIMATION IN RANDOM COEFFICIENTS BINARY CHOICE MODELS

ERIC GAUTIER AND YUICHI KITAMURA

Preliminary Draft

ABSTRACT. This paper considers random coefficients binary choice models. The main goal is to estimate the density of the random coefficients nonparametrically. This is an ill-posed inverse problem characterized by an integral transform. A new density estimator for the random coefficients is developed, utilizing Fourier-Laplace series expansions on spheres. This approach offers a clear insight on the identification problem. More importantly, it leads to a closed form estimator formula. This allows a simple plug-in procedure that requires no numerical optimization. The new estimator, therefore, is easy to implement in empirical applications, while being flexible about the treatment of unobserved heterogeneity. Extensions including treatments of non-random coefficients and models with endogeneity are discussed.

1. Introduction

Consider a binary choice model

$$Y = \mathbb{I}\left\{X'\beta \geq 0\right\} \tag{1.1}$$

where $\mathbb{I}$ denotes the indicator function and X is a d-vector of covariates. We assume that the first element of X is 1, the vector X is thus of the form $X = (1, \tilde{X}')'$. The vector β is random. The random vector $(Y, \tilde{X}, \beta)$ is defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $(y_i, \tilde{x}_i, \beta_i), i = 1, ..., N$ denote its realizations. The econometrician observes $(y_i, \tilde{x}_i), i = 1, ..., N$, but $\beta_i, i = 1, ..., N$ remain unobserved. Therefore $\tilde{X}$ and β correspond to observed and unobserved heterogeneity across agents, respectively. Note that the first element of β in this formulation absorbs the usual scalar stochastic shock term as well as a constant in standard binary choice with non-random coefficients. This formulation is used

Date: This Version: April 29, 2008.
Keywords: Inverse problems, Discrete Choice Models.

JEL Classification Number: C14.

We thank seminar participants at Chicago, Harvard/MIT, the Henri Poincaré Institute, Hitotsubashi, LSE, NYU, Princeton, Simon Fraser, Tilburg University and the Tinbergen Institute for helpful comments. Kitamura acknowledges financial support from the National Science Foundation via grants SES-0241770 and SES-0551271.

in Ichimura and Thompson (1998), and is convenient for the subsequent development in the paper. Throughout the article we assume exogeneity

Assumption 1.1. *β is independent of $\tilde{X}$,*

Section 6.3 considers ways to relax this assumption.
The choice probability is given by

$$
\begin{aligned}
r(x) &= \mathbb{P}(Y = 1 | X = x) \\
&= \mathbb{E}_{\beta}[\mathbb{I}\{x'\beta > 0\}].
\end{aligned}
\tag{1.2}
$$

Discrete choice models with random coefficients variation are useful in applied research since it is often crucial to incorporate unobserved heterogeneity in the choice behavior of individuals. There is a vast and active literature on this topic. Recent contributions include Briesch, Chintagunta and Matzkin (1996), Brownstone and Train (1999), Chesher and Santos Silva (2002), Hess, Bolduc and Polak (2005), Harding and Hausman (2006), Athey and Imbens (2007), Bajari, Fox and Ryan (2007) and Train (2003). A common approach in estimating random coefficient discrete choice models is to assume parametric specifications. A leading example is the mixed Logit model, which is discussed in details by Train (2003). If one does not impose a parametric distributional assumption, the distribution of β itself is the structural parameter of interest. The goal for the econometrician is then to back out the distribution of β from the information about $r(x)$ obtained from the data.

Nonparametric treatments for unobserved heterogeneity distributions is an important issue in econometrics. Heckman and Singer (1984) study the issue of unobserved heterogeneity distributions in duration models and propose a treatment by a nonparametric maximum likelihood estimator (NPMLE). Elbers and Ridder (1982) also develop some identification results in such models. Beran and Hall (1992) and Hoderlein et al. (2007) discuss nonparametric estimation of random coefficients linear regression models. Despite the tremendous importance of random coefficient discrete choice models, as exemplified in the above references, nonparametrics in this area is relatively underdeveloped. An important paper by Ichimura and Thompson (1998) proposes a NPMLE estimator for the CDF of β. They present sufficient conditions for identification and prove the consistency of the NPMLE. The NPMLE requires high dimensional numerical maximization and can be computationally intensive even for a moderate sample size. Berry and Haile (2008) explore nonparametric identification problems in random coefficients multinomial choice models that often arise in empirical IO.

Here we develop a different approach that shares many similarities with standard deconvolution methods in the Euclidean space. This allows us to revisit the identification issue. Moreover, once sufficient constraints are imposed on the parameter, we obtain a general estimator of the density to be used in conjunction with an estimator of the choice probability. When a particular estimator of the choice probability is used, the estimator of the density can be expressed with a closed form formula. This is a simple plug-in procedure that requires no numerical optimization or integration. This estimator is easy to implement in empirical applications, while being flexible about the treatment of unobserved heterogeneity.

Since the scale of β is not identified in the binary choice model, we normalize the scale so that β is a vector of Euclidean norm 1 in $\mathbb{R}^d$. The vector β then belongs to the $d-1$ dimensional sphere $\mathbb{S}^{d-1}$. This is not a restriction as long as the probability that β is equal to 0 is 0. Also, since only the angle between X and β matters, we replace X by $X/\|X\|$ and assume X is on the sphere. Discrete choice models with random coefficients thus naturally fit the directional data literature, see for example Fisher et al. (1987). We aim to recover the joint probability density function f_β of the preferences β with respect to the spherical measure $d\sigma$ over $\mathbb{S}^{d-1}$ from the N observations $(y_1, x_1), \ldots, (y_N, x_N)$ of (Y, X).

The problem considered here is a linear ill-posed inverse problem. We can write

$$r(x) = \int_{b \in \mathbb{S}^{d-1}} \mathbb{I}\left\{x'b \geq 0\right\} f_\beta(b) d\sigma(b) = \int_{H(x)} f_\beta(b) d\sigma(b) := \mathcal{H}\left(f_\beta\right)(x) \tag{1.3}$$

where the set $H(x)$ is the hemisphere $\{b : x'b \geq 0\}$. The mapping $\mathcal{H}$ is called the hemispherical transformation. Inversion of this mapping was first studied by Funk (1916) and later by Rubin (1999). Groemer (1996) also recalls some of its properties. $\mathcal{H}$ is not injective without further restrictions and conditions need to be imposed to ensure identification. Even under an additional condition which guarantees identification, however, the inverse of $\mathcal{H}$ is not a continuous mapping, making the problem ill-posed. To see this, suppose we restrict f_β to be in $\mathrm{L}^2(\mathbb{S}^{d-1})$. Since the kernel is square integrable by compactness of the sphere, the operator is Hilbert-Schmidt and thus compact. Therefore if the inverse of $\mathcal{H}$ were continuous, $\mathcal{H}^{-1}\mathcal{H}$ would map the closed unit ball in $\mathrm{L}^2(\mathbb{S}^{d-1})$ to a compact set. But the Riesz theorem states that the unit ball is relatively compact if and only if the vector space has finite dimension. The fact that $\mathrm{L}^2(\mathbb{S}^{d-1})$ is an infinite dimensional space contradicts this. Therefore the inverse of $\mathcal{H}$ cannot be continuous. In order to overcome this problem, we use a one parameter family of regularized inverses that are continuous and converge to the inverse when the parameter goes to infinity. This is a common approach to ill-posed inverse problems in statistics (see, e.g. Carrasco et al., 2007).

Due to the particular form of the kernel of the operator $\mathcal{H}$ involving the scalar product $x'b$, the operator is a non commutative analogue of the convolution in $\mathbb{R}^d$. This analogy provides a clear insight into the identification issue. Note that our problem closely related to the so-called boxcar deconvolution (see, e.g. Groeneboom and Jongbloed (2003) and Johnstone and Raimondo (2004)), where identifiablity is a significant problem. The connection with deconvolution is also useful in deriving an estimator based on a series expansion on the Fourier basis or its extension to higher dimensional spheres called Fourier-Laplace series. These bases are defined via the Laplacian on the sphere, and they diagonalize the operator $\mathcal{H}$ on $\mathrm{L}^2\left(\mathbb{S}^{d-1}\right)$. Such techniques are used in Healy and Kim (1996) for nonparametric empirical Bayes estimation in the case of the sphere $\mathbb{S}^2$. The boxcar kernel of the integral operator $\mathcal{H}$, however, does not satisfy the assumptions made by Healy and Kim. In contrast to this paper, we make use of so-called "condensed" expressions. The approach replaces a full expansion on a Fourier-Laplace basis by an expansion in terms of the projections on the finite dimensional eigenspaces of the Laplacian on the sphere. This is useful since an explicit expression of the kernel of the projector is available. It allows us to work in any dimension and does not require a parametrization by hyperspherical coordinates nor the actual knowledge of an orthonormal basis. This approach, to the best of our knowledge, appears to be new in the econometrics literature.

The paper is organized as follows. In Section 2 we introduce a toy model and the tools from harmonic analysis that are used for the development of our estimation procedure and its asymptotic analysis. Section 3 deals with both the identification and a general procedure for the estimation of the density of the random coefficient relying on an estimator of the choice probability. In Section 4 we study a particular estimator of the choice probability and its derivatives and present their asymptotic properties. The corresponding estimator of the density of the random coefficients takes a simple closed form, and we derive its convergence rate in all the spaces $\mathrm{L}^p, p \in [1,\infty)$ and also prove a pointwise CLT in Section 5. Extensions such as estimation of marginals, models with non-random coefficients, and treatment of endogeneity are presented in Section 6.

2. Preliminaries

In this section we introduce some tools that are used to relate the estimation of the density of β to a deconvolution problem and results on the Hemispherical transform.

2.1. **A Toy Model.** We first study the case where X is of dimension 2 to gain basic insights. We parameterize the vector $b = (b_1, b_2)'$ of $\mathbb{S}^1$ by the angle $\phi = \arccos(b_1)$ in $[0, 2\pi)$. As it is often the case when standard Fourier series techniques are used, we consider spaces of complex valued functions. Let $\mathrm{L}^p(\mathbb{S}^1)$ denote the Banach spaces of Lebesgue p-integrable functions and its norm by $\|\cdot\|_p$. In the case of $\mathrm{L}^2(\mathbb{S}^1)$, the norm is derived from the hermitian product $\int_0^{2\pi} f(\theta)\overline{g(\theta)}d\theta$. With the parametrization by angles we obtain

$$\mathcal{H}(f_\beta)(\theta) = \int_0^{2\pi} \mathbb{I}\left\{|\theta - \phi| < \pi/2\right\} f_\beta(\phi)d\phi. \tag{2.1}$$

This expression suggests that the hemispherical transformation is a usual convolution of functions on $\mathbb{R}/(2\pi\mathbb{Z})$. Rewrite (2.1) as

$$\frac{\mathcal{H}(f_\beta)}{\pi}(\theta) = \int_0^{2\pi} \left(\frac{1}{\pi}\mathbb{I}\left\{|\theta - \phi| < \pi/2\right\}\right) f_\beta(\phi)d\phi. \tag{2.2}$$

It is then possible to link estimation of f_β with statistical deconvolution problems. $\mathcal{H}(f_\beta)/\pi$ is then interpreted as the density of θ, which is generated by adding (on $\mathbb{R}/(2\pi\mathbb{Z})$) a "noise" drawn from the uniform density $\frac{1}{\pi}\mathbb{I}\{|x| < \pi/2\}$ to the "signal" ϕ drawn from f_β. Let us relate inversion of the operator with differentiation. Differentiating the right hand-side of expression (2.1) we obtain $f_\beta(\theta + \pi/2) - f_\beta(\theta - \pi/2)$ where f_β is defined on the line by periodicity. Under an assumption such that f_β is supported on a hemisphere, this assumption is discussed further in Section 3.1, we obtain either $f_\beta(\theta + \pi/2)$ or $-f_\beta(\theta - \pi/2)$. When the model is identified properly the inverse is a differential operator and as such unbounded. It is typically the case that the inverse of kernel operator is a differential operator but, in order to generalize the inversion to any dimension, it is useful to work with Fourier series and their generalizations to higher dimensional spheres.

Fourier series is a useful tool for deconvolution problems on the circle. $\left(\exp(-int)/\sqrt{2\pi}\right)_{n\in\mathbb{Z}}$ is the orthonormal basis of $\mathrm{L}^2(\mathbb{S}^1)$ used to define Fourier series. This system is also complete in $\mathrm{L}^1(\mathbb{S}^1)$. Denoting by $c_n(f) = \int_0^{2\pi} f(t)\exp(-int)dt/(2\pi)$ the Fourier coefficients of $f \in \mathrm{L}^1(\mathbb{S}^1)$

$$f_\beta(\theta) = \sum_{n\in\mathbb{Z}} c_n(f_\beta)\exp(in\theta) \tag{2.3}$$

in the $\mathrm{L}^1(\mathbb{S}^1)$ sense. Recall also that for f and g in $\mathrm{L}^1(\mathbb{S}^1)$,

$$c_n(f * g) = 2\pi c_n(f)c_n(g). \tag{2.4}$$

Using equation (2.4) we obtain the following proposition.

Proposition 2.1. *$c_0(\mathcal{H}(f_\beta)) = \pi c_0(f_\beta)$ and for $n \in \mathbb{Z}\setminus\{0\}$, $c_n(\mathcal{H}(f_\beta)) = c_n(f_\beta)\, 2\sin(n\pi/2)/n$.*

As in classical deconvolution problems on the real line, our aim is to obtain f_β using equation (2.3) and Proposition 2.1. Notice that among the Fourier coefficients $c_n(f_\beta), n = 1, 2, ...$ it is only possible to recover the first coefficient $c_0(f_\beta)$ (which is easily seen to be $1/2\pi$, by integrating both sides of (2.1) and noting that f_β is a probability density function) and the odd coefficients. Indeed, Proposition 2.1 shows that $c_{2p}(\mathcal{H}(f_\beta)) = 0$ holds for all non-zero p's, regardless of the value of $c_{2p}(f_\beta)$. In other words, any f_β with the same coefficient $c_0(f_\beta)$ and odd coefficients gives rise to the same hemispherical transformation. Variations in r do not allow to identify the coefficients $c_{2p}(f_\beta)$ for a non zero p. The same phenomenon occurs in higher dimensions, as explained in Section 2.2.

Remark 2.1. If we make the stronger assumption that f_β belongs to $\mathrm{L}^2(\mathbb{S}^1)$, we may interpret this result in terms of operators. For $n \neq 0$ the vector spaces $H^{n,2} = \mathrm{span}\left\{\exp(int)/(2\pi), \exp(-int)/(2\pi)\right\}$ are eigenspaces of the compact self-adjoint operator $\mathcal{H}(f_\beta)$. These eigenspaces are associated with the eigenvalues $2\sin\left(n\pi/2\right)/n$. Also, $\bigoplus_{p\in\mathbb{Z}} H^{2p,2}$ is the null space $\ker \mathcal{H}$ of $\mathcal{H}$.

2.2. **Tools for Higher Dimensional Spheres.** Let us introduce some concepts used to treat the general case where $d \geq 2$. We consider functions defined on the sphere $\mathbb{S}^{d-1}$, which is a $d-1$ dimensional smooth submanifold of $\mathbb{R}^d$. The canonical measure on $\mathbb{S}^{d-1}$ (or spherical measure) is denoted by $d\sigma$ and is such that $\int_{\mathbb{S}^{d-1}} d\sigma = |\mathbb{S}^{d-1}|$ is the area of the sphere. It is given for $d \geq 1$ by $\left|\mathbb{S}^{d-1}\right| = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ where Γ is the usual Gamma function. $\mathrm{L}^p(\mathbb{S}^{d-1})$ with norm $\|\cdot\|_p$ are the usual spaces of integrable complex functions and $\mathrm{L}^2(\mathbb{S}^{d-1})$ is equipped with the hermitian product $(f,g)_{\mathrm{L}^2(\mathbb{S}^{d-1})} = \int_{\mathbb{S}^{d-1}} f(x)\overline{g}(x)d\sigma(x)$. We use the following notation throughout the paper

Notation. For two sequences of positive numbers $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, we write $a_n \asymp b_n$ when there exists a positive M such that $M^{-1}b_n \leq a_n \leq Mb_n$ for every n positive. □

The Laplacian Δ^S on the sphere allows to extend the Fourier basis to any dimension in the similar manner as the functions $\exp(-int)/\sqrt{2\pi}$ are eigenfunctions of $-\frac{\mathrm{d}}{\mathrm{d}t^2}$ associated with the eigenvalue n^2. Let Δ denote the Laplacian in $\mathbb{R}^d$, $\check{f}$ the radial extension of f, that is $\check{f}(x) = f(x/\|x\|)$, and $\hat{f}$ the restriction of f to $\mathbb{S}^{d-1}$. Δ^S, defined in terms of Riemannian geometry the usual way via a generalization of the formula "div∇^S" see the appendix, has a simple expression in the case of spheres

$$\Delta^S f = (\Delta \check{f})^{\frown} \tag{2.5}$$

also

$$\nabla^S f = (\nabla \check{f})^{\frown}. \tag{2.6}$$

Definition 2.1. A surface harmonic of degree n is the restriction to $\mathbb{S}^{d-1}$ of a homogeneous harmonic (solution of $\Delta f = 0$) polynomial of degree n in $\mathbb{R}^d$.

The reader is referred to Müller (1966) and Groemer (1996) for clear and detailed expositions on these concepts and important results concerning spherical harmonics used in this paper. Erdélyi et al. (1953, vol. 2, chapter 9) provide detailed accounts focusing on special functions. The proofs and results below can be found in the above references.

Lemma 2.1. *The following properties hold:*

(*i*) $-\Delta^S$ *is a positive self-adjoint unbounded operator on* $\mathrm{L}^2(\mathbb{S}^{d-1})$, *thus it has orthogonal eigenspaces and a basis of eigenfunctions;*

(*ii*) *Surface harmonics of positive degree* n *are eigenfunctions of* $-\Delta^S$ *for the eigenvalue* $n(n+d-2)$*;*

(*iii*) *The dimension of the vector space* $H^{n,d}$ *of spherical harmonics of degree* n *is*

$$\dim\, H^{n,d} = \frac{(2n+d-2)(n+d-2)!}{n!(d-2)!(n+d-2)}; \tag{2.7}$$

(*iv*) *A system formed of orthonormal bases of* $H^{n,d}$ *for each degree* $n = 0, \ldots, \infty$ *is complete in* $\mathrm{L}^1(\mathbb{S}^{d-1})$.

Notation. We let $h(n,d)$ denote $\dim\, H^{n,d}$ and $\zeta_{n,d} = n(n+d-2)$. □

Lemma 2.1 (i) and (iv) give the decomposition

$$\mathrm{L}^2(\mathbb{S}^{d-1}) = \bigoplus_{n\in\mathbb{N}} H^{n,d}$$

with orthogonal $H^{n,d}$'s being the eigenspaces of Δ^S. The space of surface harmonics of degree 0 is the one dimensional space spanned by 1. A series expansion on an orthonormal basis of surface harmonics is called a Fourier series when $d = 2$, a Laplace series when $d = 3$ and in the general case a Fourier-Laplace series.

Orthonormal bases of surface harmonics usually involve parametrization by angles, such as the spherical coordinates when $d = 3$ as used by Healy and Kim (1996) or hyperspherical coordinates for $d > 3$. In contrast, here we work with the decomposition of a function on the spaces $H^{n,d}$.

Definition 2.2. The condensed harmonic expansion of a function f in $\mathrm{L}^1(\mathbb{S}^{d-1})$ is the series $\sum_{n=0}^{\infty} Q_{n,d} f$.

This leads to a simple method both in terms of theoretical developments and practical implementations. The projector $Q_{n,d}$ on $H^{n,d}$ in $\mathrm{L}^2(\mathbb{S}^{d-1})$ can be expressed as an integral operator with kernel

$$q_{n,d}(x,y) = \sum_{l=1}^{h(n,d)} \overline{Y_{n,l}(x)} Y_{n,l}(y), \tag{2.8}$$

where $(Y_{n,l})_{l=1}^{h(n,d)}$ is any orthonormal basis of $H^{n,d}$. The kernel has a simple expression given by the addition formula.

Theorem 2.1 (Addition Formula)**.** *The following identity holds*

$$q_{n,d}(x,y) = \frac{h(n,d) C_n^{\nu(d)}(x'y)}{|\mathbb{S}^{d-1}| C_n^{\nu(d)}(1)} := {}^\flat q_{n,d}(x'y) \tag{2.9}$$

where C_n^ν are the Gegenbauer polynomials and here

$$\nu(d) = (d-2)/2.$$

The Gegenbauer polynomials are defined for $\nu > -1/2$ and are orthogonal with respect to the weight function $(1-t^2)^{\nu-1/2}dt$ on $[-1,1]$. They correspond to the $2/n$ times the Chebychev polynomials of the first kind when $d=2$, to the Legendre polynomials when $d=3$ and to the Chebychev polynomials of the second kind when $d=4$. Note that $C_0^\nu(t) = 1$ and $C_1^\nu(t) = 2\nu t$ for $\nu \neq 0$ while $C_1^0(t) = 2t$. Moreover, they satisfy the following recursion relation

$$(n+2)C_{n+2}^\nu(t) = 2(\nu+n+1)tC_{n+1}^\nu(t) - (2\nu+n)C_n^\nu(t). \tag{2.10}$$

In our approach the Gegenbauer polynomials will be evaluated at N points for a series of successive values of the degree n. The recursion relation (2.10) is therefore a powerful tool. Useful results on these polynomials are gathered in the appendix, see also Erdélyi et al. (1953, vol. 1, p. 175-179).

Definition 2.3. A zonal function f is a function that depends only on the geodesic distance to the north pole $\mathbf{n} = (1,0,\ldots,0)$ $d(x,\mathbf{n}) = \arccos(x'\mathbf{n})$. It can be written as $f(x) = {}^\flat f(x'\mathbf{n})$ where ${}^\flat f$ is defined on $[-1,1]$.
The convolution of a zonal function f with a function g is defined by

$$(f * g)(x) = \int_{\mathbb{S}^{d-1}} {}^\flat f(x'y) g(y) d\sigma(y).$$

Note that the convolution operation is commutative when two zonal functions are considered. Young inequalities are given for example in Kamzolov (1982).

Proposition 2.2 (Young inequalities)**.** *If f is zonal and belongs to $\mathrm{L}^p(\mathbb{S}^{d-1})$ and g belongs to $\mathrm{L}^r(\mathbb{S}^{d-1})$ then $f * g$ is well defined in $\mathrm{L}^q(\mathbb{S}^{d-1})$ when $p, q, r \geq 1$ and $1/q = 1/p + 1/r - 1$. Moreover*

$$\|f * g\|_q \leq \|f\|_r \|g\|_p.$$

It is interesting to note that we can write

$$\mathcal{H}f(x) = (\mathbb{I}\{\cdot'\mathbf{n} \geq 0\} * f)(x), \tag{2.11}$$

$$Q_{n,d}f(x) = (q_{n,d}(\cdot, \mathbf{n}) * f)(x),$$

and defining the projection onto $\bigoplus_{n=o}^{T} H^{n,d}$ by P_T

$$P_T f(x) = (D_T(\cdot, \mathbf{n}) * f)(x)$$

where

$$D_T(x, y) = \sum_{n=0}^{T} q_{n,d}(x, y)$$

is the Dirichlet kernel which extends the classical Dirichlet kernel on the circle. The sum over T in the definition of D_T also has the simple closed form (52) in Müller (1966) in terms of derivatives of Gegenbauer polynomials. Inversion of $\mathcal{H}$ corresponds to deconvolution. We can also note that the linear form $f \to (D_T(\cdot, \mathbf{n}) * f)(x)$ converges as T goes to infinity to the Dirac measure δ_x. The integral operator is a kernel operator. Generalization of trigonometric kernels are used as a regularization tool to estimate the choice probability and the coefficient of the random coefficient. The Dirichlet kernel corresponds to the best approximation in $\mathrm{L}^2(\mathbb{S}^{d-1})$ but is known to have flaws. It is not a *bona fide* approximation kernel (Katznelson (2004)). Indeed, the $\mathrm{L}^1(\mathbb{S}^{d-1})$ norm of the kernel is not uniformly bounded; more precisely, we have

$$\|D_T(\cdot, \mathbf{n})\|_1 \asymp T^{(d-2)/2} \tag{2.12}$$

when $d \geq 3$ and

$$\|D_T(\cdot, \mathbf{n})\|_1 \asymp \log T \tag{2.13}$$

when $d = 2$: See Gronwall (1914) for the $d = 3$ case and Ragozin (1972) and Colzani and Traveglini (1991) for higher dimensions. Furthermore, it is not a positive kernel either. One of the consequence is

Gibbs oscillations which deteriorate as the dimension increases. This suggests the use of other kernels as for Fourier series. The Cesàro kernel (see, e.g. Bonami and Clerc, 1973) is given by

(2.14)
$$S^{\delta}_{T,d}(x,y)=\sum_{k=0}^{T}\left(1-\frac{k}{T+1}\right)\left(1-\frac{k}{T+2}\right)\dots\left(1-\frac{k}{T+\delta}\right)q_{k,d}(x,y)=\frac{1}{A^{\delta}_{T}}\sum_{k=0}^{T}A^{\delta}_{T-k}q_{k,d}(x,y)$$

where

$$\sum_{n=0}^{\infty}A^{\delta}_{n}x^{n}=(1-x)^{-\delta-1},\ \text{i.e. } A^{\delta}_{n}=\left(\begin{array}{c}n+\delta\\ n\end{array}\right)=\frac{(n+\delta)(n+\delta-1)\dots(\delta+1)}{n(n-1)\dots1}\asymp n^{\delta}.$$

The Cesàro kernel is obtained by taking Cesàro means of the Dirichlet kernel. It puts more weight than the Dirichlet kernel on the lower frequencies and provides more smoothing. The Fejèr kernel in the $d=2$ case is a Cesàro kernel. Kogbetliantz (1924) proved that ${}^{\flat}S^{\delta}_{T,d}$ is everywhere non-negative when $\delta\ge d-1$. We will now choose $\delta\ge d-1$ and $\delta=d-1$ for the estimation of a function (density, regression function, here the choice probability, or f_β), $\delta=d+1$ for the estimation of a function and its derivatives and so on. Positiveness is very convenient in our case to treat the plug-in. An important result (see, e.g. Kamzolov, 1982) is

$$\forall\delta>(d-2)/2,\ \forall p\ge1,\ \left\|S^{\delta}_{T,d}(\cdot,\mathbf{n})\right\|_{p}\asymp T^{(d-1)(1-1/p)} \tag{2.15}$$

which implies that for our choice of δ the $\mathrm{L}^1(\mathbb{S}^{d-1})$ norms of the Cesàro kernels are uniformly bounded. Note that $\mathrm{L}^1(\mathbb{S}^{d-1})$ norms are of the same order in T for Riesz kernels (see, e.g. Colzani and Traveglini, 1991) but here working with Cesàro kernels we also obtain positive kernels. The following proposition is proved in the appendix.

Proposition 2.3. *For all δ non-negative, there exists a constant K such that for all $x,y,z\in\mathbb{S}^{d-1}$,*

$$\left|{}^{\flat}S^{\delta}_{T,d}(z'x)-{}^{\flat}S^{\delta}_{T,d}(z'y)\right|\le K|x-y|T^{d+1}.$$

Lemma 2.1 (ii) allows us to provide the following definion of the Sobolev spaces in terms of the distribution $g_s(x)=\sum_{k=1}^{\infty}\zeta_{k,d}^{-s/2}q_{k,d}(x,\mathbf{n})$:

Definition 2.4. The Sobolev space $\mathrm{W}^s_p(\mathbb{S}^{d-1})$ is the set of functions f for which there exists f_s in $\mathrm{L}^p(\mathbb{S}^{d-1})$ satisfying $\int_{\mathbb{S}^{d-1}}f_s(x)d\sigma(x)=0$ and a constant C such that

$$f(x)=C+(g_s*f_s)(x).$$

The f_s is called the s-th fractional derivative of f. If s is an integer, then we equip the space with the norm

$$\|f\|_{p,s} = \max\left\{\|f\|_p, \|f_1\|_p, \dots, \|f_s\|_p\right\}.$$

For the case of $s = 2$, that is, for the Sobolev space $\mathrm{H}^s(\mathbb{S}^{d-1}) := \mathrm{W}_2^s(\mathbb{S}^{d-1})$, it is also possible to use an equivalent norm, (the square of) which is equal to

$$\sum_{n=0}^{\infty} (1+\zeta_{n,d})^s \|Q_{n,d}f\|_2^2$$

and consider a continuum of values for s. We use these spaces to make smoothness assumptions. Useful bounds on the approximation are given as follows (see, e.g. Kamzolov (1982) and Kushpel et al. (1997)). Note that its part (ii) provides upper bounds in terms of the Sobolev norm $\|\cdot\|_{p,s}$ defined above.

Proposition 2.4 (approximation error). *(i) For f in $\mathrm{H}^s(\mathbb{S}^{d-1})$ and $v < s$ where v and s take continuous values,*

$$\|f - D_T(\cdot, \mathbf{n}) * f\|_{2,v} \le T^{-(s-v)} \|f\|_{2,s};$$

(ii) For $d \ge 2$, p in $[1, \infty)$ and integers v and s, $v < s$, there exists a constant $A(d, \delta, s, v, p)$ such that for every f in $\mathrm{W}_p^s(\mathbb{S}^{d-1})$,

$$\left\| f - S_{T,d}^{\delta}(\cdot, \mathbf{n}) * f \right\|_{p,v} \le A(d, \delta, s, v, p) T^{-(s-v)} \|f\|_{p,s}.$$

The odd and even part of a function defined on the sphere are important notions in the development of our analysis of the identification.

Definition 2.5. We define the odd part and the even part of a function f by:

$$f^-(b) = (f(b) - f(-b))/2$$

and

$$f^+(b) = (f(b) + f(-b))/2,$$

for every b in $\mathbb{S}^{d-1}$

If the function f is in $\mathrm{L}^2(\mathbb{S}^{d-1})$ then using equations (2.9) and (9.10) we can check that for p non-negative $Q_{2p,d}f(x) = Q_{2p,d}f(-x)$ and $Q_{2p+1,d}f(x) = -Q_{2p+1,d}f(-x)$. Thus the sum of the odd terms in the condensed harmonic expansion corresponds to f^- and the sum of the even terms corresponds to f^+. If a non-negative function f has its support included in some hemisphere then

$$f(x) = 2f^-(x)\mathbb{I}\left\{f^-(x) > 0\right\}. \tag{2.16}$$

If we denote by supp$f$ the support of f, this follows from the fact that $f^-(x) = f^+(x) \geq 0$ on supp$f$ while $f^-(x) = -f^+(x) \leq 0$ on $-$supp$f$ and both f^- and f^+ are 0 on $\mathbb{S}^{d-1} \setminus (\text{supp}f \bigcup -\text{supp}f)$.
If f is a probability density function, the coefficient of degree 0 in the expansion of f on surface harmonics is $1/|\mathbb{S}^{d-1}|$.

Remark 2.2. Conversely, any harmonic polynomial or series such that the degree 0 coefficient is $1/|\mathbb{S}^{d-1}|$ integrates to one. Thus, truncation used below as a regularization procedure, preserves the probability mass. Non-negativity can be ensured working with suitablly chosen Cesàro kernels.

The next theorem shows that Fourier-Laplace series on spheres is a very natural tool for the study of our operator which as we have seen corresponds to convolution.

Theorem 2.2 (Funk-Hecke Theorem)**.** *If g belongs to $H^{n,d}$ for some n and F is such that*

$$\int_{-1}^{1} |F(t)|^2(1-t^2)^{(d-3)/2}dt < \infty,$$

then

$$\int_{\mathbb{S}^{d-1}} F(x'y)g(y)d\sigma(y) = \lambda_n(F)g(x) \tag{2.17}$$

where

$$\lambda_n(F) = \int_{-1}^{1} F(t)^\flat q_{n,d}(t)(1-t^2)^{(d-3)/2}.$$

In other words, the kernel operator K defined by

$$f \in \mathrm{L}^2(\mathbb{S}^{d-1}) \mapsto \left(x \mapsto \int_{\mathbb{S}^{d-1}} F(x'y)f(y)d\sigma(y)\right) \in \mathrm{L}^2(\mathbb{S}^{d-1})$$

is, when restricted to a subspace $H^{n,d}$, the multiplication by $\lambda_n(F)$. Thus a basis of surface harmonics diagonalizes any integral operator where the kernel function involves the scalar product $x'y$.

Remark 2.3. Healy and Kim (1996) use Fourier-Laplace expansions to analyze a deconvolution problem on the sphere in dimension $d = 3$. As we shall see below, the Addition Formula along with condensed harmonic expansions provide a general treatment that works for cases with arbitrary dimension.

2.3. **The Hemispherical Transform.** The Hemisphercial transform corresponds to a particular case of the kernel $F(t) = \mathbb{I}\{t \in [0,1]\}$ in the Funk-Hecke theorem.

Notation. We define $\lambda(n,d) = \lambda_n\left(\mathbb{I}\left\{t \in [0,1]\right\}\right)$ for $d \geq 3$ and $\lambda(n,2) = \frac{2\sin(n\pi/2)}{n}$ of Proposition 2.1. □

Proposition 2.5. *When $d \geq 2$, the coefficients $\lambda(n,d)$ have the following expression*

(*i*) $\lambda(0,d) = \frac{2}{|\mathbb{S}^{d-1}|}$
(*ii*) $\lambda(1,d) = \frac{|\mathbb{S}^{d-2}|}{d-1}$
(*iii*) $\forall p > 0,\ \lambda(2p,d) = 0$
(*iv*) $\forall p > 0,\ \lambda(2p+1,d) = \frac{(-1)^p|\mathbb{S}^{d-2}|1\cdot 3\cdots(2p-1)}{(d-1)(d+1)\cdots(d+2p-1)}$.

For the sake of completeness we give a simple proof of this result in the appendix (see also Groemer (1996) and Rubin (1999)). The following corollary corresponds to an observation made in Remark 2.1 for the $d = 2$ case.

Corollary 2.1. *The null space of the himispherical transform $\mathcal{H}$, seen as an operator on $\mathrm{L}^2(\mathbb{S}^{d-1})$, is*

$$\ker \mathcal{H} = \bigoplus_{p=1}^{\infty} H^{2p,d}.$$

The spaces $H^{0,d}$ and $H^{2p+1,d}$ for p non negative are the eigenspaces associated with non zero eigenvalues.

As a consequence of Proposition 2.5 $\mathcal{H}$ is not injective and restrictions have to be imposed in order to ensure identification. In Section 3 we present conditions for identification which allow us to reconstruct f_β from f_β^-.

Define $\mathrm{L}^2{}_{\mathrm{odd}}(\mathbb{S}^{d-1})$ and $\mathrm{H}^s_{\mathrm{odd}}(\mathbb{S}^{d-1})$ as the restrictions of $\mathrm{L}^2(\mathbb{S}^{d-1})$ and $\mathrm{H}^s(\mathbb{S}^{d-1})$ to odd functions. The following proposition can be found in Rubin (1999).

Proposition 2.6. *$\mathcal{H}$ is a bijection from $\mathrm{L}^2_{\mathrm{odd}}(\mathbb{S}^{d-1})$ to $\mathrm{H}^{d/2}_{\mathrm{odd}}(\mathbb{S}^{d-1})$.*

We can also easily check (see the proof in the appendix) that

Proposition 2.7. *For all s non-negative, there exists positive constants C_l and C_u such that for all f in $\mathrm{H}^s(\mathbb{S}^{d-1})$*

$$C_l \left\|f^-\right\|_{2,s} \leq \left\|\mathcal{H}(f^-)\right\|_{2,s+d/2} \leq C_u \left\|f^-\right\|_{2,s}.$$

The factor $d/2$ corresponds to the degree of "regularization" due to smoothing by $\mathcal{H}$. Now the inverse of an odd function R^- is given by the distribution

$$\mathcal{H}^{-1}(R^-)(b) = \left(i_s * R^-\right)(b) \tag{2.18}$$

where the distribution i_s is given by:

$$i_s(x) = \sum_{p=0}^{\infty} \frac{1}{\lambda(2p+1,d)} q_{2p+1,d}(x,\mathbf{n}). \tag{2.19}$$

When R^- belongs to $\mathrm{H}^{d/2}(\mathbb{S}^{d-1})$ then $\mathcal{H}^{-1}(R^-)(b)$ is a well defined $\mathrm{L}^2(\mathbb{S}^{d-1})$ function, otherwise the distribution is only defined in a Sobolev space with negative exponent. Moreover, for some values of d in addition to the case of $d=2$ we presented earlier, it is still possible to relate the inverse of the operator $\mathcal{H}$ with differentiation. If we consider the case where d is even, we know from Proposition 2.5, that

$$\frac{1}{\lambda(2p+1,d)} = (-1)^p|\mathbb{S}^{d-2}|(2p+1)(2p+3)\dots(d+2p-1).$$

Thus when 4 divides d,

$$\frac{1}{\lambda(2p+1,d)} = |\mathbb{S}^{d-2}| \prod_{k=1}^{d/4}[-\zeta_{2p+1,d} + 2(k-1)(d-2k)]$$

and when d is even but 4 does not divide d,

$$\frac{1}{\lambda(2p+1,d)} = -|\mathbb{S}^{d-2}|(2p+d/2) \prod_{k=1}^{(d-2)/4}[-\zeta_{2p+1,d} + 2(k-1)(d-2k)].$$

Hence we have obtained the following result

Proposition 2.8. *When 4 divides d,*

$$\mathcal{H}^{-1} = |\mathbb{S}^{d-2}| \prod_{k=1}^{d/4}[-\Delta^S + 2(k-1)(d-2k)].$$

As a consequence of Bernstein type inequalities on the sphere (see, e.g. Ditzian, 1998), at least when 4 divides d and from Section 2.1 when $d=2$, we know that

$$\forall q \in [1,\infty],\ \exists C>0:\ \forall P \in \bigoplus_{p=o}^{T} H^{2p+1,d},\ \|\mathcal{H}^{-1}P\|_q \le CT^{d/2}\|P\|_q. \tag{2.20}$$

But the results in Ditzian (1998) allow us to prove directly the following Bernstein-type inequalities for all dimensions, as shown in the appendix.

Theorem 2.3 (Bernstein type inequalities)**.** *For all dimensions $d \geq 2$, all $q \in [1, \infty]$, there exists a positive C such that for all P in $\bigoplus_{p=0}^{T} H^{2p+1,d}$,*

$$\|\mathcal{H}^{-1}P\|_q \leq CT^{d/2}\|P\|_q. \tag{2.21}$$

This proves to be very important for our subsequent analysis of the estimation of the density of the random coefficient. In addition to the bound implied by Proposition 2.7, we use it in our analysis of various bounds involving the $\mathrm{L}^1(\mathbb{S}^{d-1})$ and $\mathrm{L}^\infty(\mathbb{S}^{d-1})$ norms.

Rubin (1999) gives other inversion formulas for the Hemispherical transform. For example, when d is even, the inverse of $\mathcal{H}^2$ is a polynomial of degree $d/2$ in the Laplacian. This is straightforward from the above computations. When d is odd, the inverse involves a differential operator as well as an operator involving the principal value. It is also shown that a wavelet transform also allows to invert the hemispherical transform. The fact that the inversion corresponds roughly to a differential operator is another manifestation, besides invertibility or identification, of the ill-posedness. Indeed, it implies that the operator is unbounded. We call the factor $d/2$ in (2.21) the degree of ill-posedness of the inverse problem. Note that (2.21) holds with the equality sign for $q = 2$, therefore the definition of the ill-posedness $d/2$ is exact for the $\mathrm{L}^2(\mathbb{S}^{d-1})$-norm.

3. General Results

3.1. Identification in the Random Coefficient Model.

This section analyzes the identifiability of f_β and discusses sufficient conditions for identification. We make the following assumption which also appears in Ichimura and Thompson (1998). It is used to extend the choice probability $r(x)$ to a function on the whole sphere and as a result to identify f_β.

Assumption 3.1. *The support of X is the whole northern hemisphere $H^+ = \{x \in \mathbb{S}^{d-1} : \ x'\mathbf{n} \geq 0\}$.*

This assumption demands that $\tilde{X}$ is supported on the whole space $\mathbb{R}^{d-1}$. It rules out discrete or bounded $\tilde{X}$ (See Section 6 for a potential approach to deal with such regressors as dummy variables). We now assume that the law of X is absolutely continuous with respect to $d\sigma$ and denote its density by f_X.

We now consider choice probabilities $r(x)$ given by (1.2) which are invariant by dilatation

$$\forall x \in \mathbb{R}^d, \ \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 1|X = x/\|x\|).$$

As such they can be studied as function on the sphere. The invariance by dilatation is satisfied in the case of the random coefficient model (1.1). They are only defined on the support of X. Under

Assumption 3.1 it is possible to extend such functions $r(x)$ to a *bona fide* function on the whole sphere. If we again think that the choice probability si consistent with the random coefficients model (1.1), then noting that f_β is a probability density function, we obtain for x in H^+

$$\mathcal{H}(f_\beta)(-x) = \int_{H(-x)} f_\beta(b)d\sigma(b) = 1 - r(x) = 1 - \mathcal{H}(f_\beta)(x). \tag{3.1}$$

We thus consider an extension R as follows:

$$\forall x \in H^+, R(x) = r(x), \text{ and } \forall x \in -H^+, R(x) = 1 - r(-x) = 1 - R(-x). \tag{3.2}$$

Note that

$$\begin{aligned} R(x) &= R^+(x) + R^-(x) \\ &= \frac{1}{2}\left[R(x) + R(-x)\right] + R^-(x) \\ &= \frac{1}{2}\left[R(x) + (1 - R(x))\right] + R^-(x) \quad \text{by (3.2)} \\ &= \frac{1}{2} + R^-(x) \end{aligned} \tag{3.3}$$

thus R is then entirely determined by its odd part. Now, provided that the extension R belongs to $\mathrm{H}^{d/2}(\mathbb{S}^{d-1})$ (the Sobolev imbedding of $\mathrm{H}^s(\mathbb{S}^{d-1})$ into the space of continuous functions for $s > (d-1)/2$ implies it is continuous), there exists a unique odd function f^- in $\mathrm{L}^2(\mathbb{S}^{d-1})$ such that

$$R = \frac{1}{2} + \mathcal{H}\left(f^-\right) = \mathcal{H}\left(\frac{1}{|\mathbb{S}^{d-1}|} + f^-\right),$$

This follows from Proposition 2.6. Moreover as $\forall x \in \mathbb{S}^{d-1},\ 0 \le R(x) \le 1$, $\int_{\{f^-(b)\ge 0\}} f^-(b)d\sigma(x) = -\int_{\{f^-(b)\ge 0\}} f^-(b)d\sigma(b) \le 1$, thus $\int_{\mathbb{S}^{d-1}} |f^-(b)|d\sigma(b) \le 1$. Also, following the discussion of Section 2.2, $\frac{1}{|\mathbb{S}^{d-1}|} + f^-$ integrates to 1. Proposition 2.5 implies that whatever the even function g having 0 as coefficient of degree 0 (i.e. integrating to zero over the sphere),

$$R = \mathcal{H}\left(g + \frac{1}{|\mathbb{S}^{d-1}|} + f^-\right).$$

Now the function

$$g = |f^-| - \frac{1}{|\mathbb{S}^{d-1}|}\int_{\mathbb{S}^{d-1}} |f^-(b)|d\sigma(b)$$

is even, integrates to zero and it allows us to define a function:

$$f_\beta^* := g + \frac{1}{|\mathbb{S}^{d-1}|} + f^- = 2f^-\mathbb{I}\{f^- > 0\} + \frac{1}{|\mathbb{S}^{d-1}|}\left(1 - \int_{\mathbb{S}^{d-1}} |f^-(b)|d\sigma(b)\right) \ge 0.$$

Obviously $f_\beta^{*-} = f_\beta^- = f^-$. This function f_β^* is non-negative, and it is indeed bounded from below by (and equal to on at least a whole hemisphere) $\frac{1}{|\mathbb{S}^{d-1}|}\left(1 - \int_{\mathbb{S}^{d-1}} |f_\beta^-(b)|d\sigma(b)\right)$.

The above discussion highlights the central identification problem in this model. Without further assumptions, we cannot rule out other functions like g that give rise to legitimate pdf's while being consistent with R. Only the odd part f_β^- of the density of the random coefficient, besides the known coefficient of degree 0, is identified. We thus present a sufficient condition on f_β so that when it is satisfied, f_β is identified from f_β^-. Ichimura and Thompson (1998, Theorem 1) give a set of conditions that imply the identification of the model (1.1). One of the assumptions postulates that there exists c on $\mathbb{S}^{d-1}$ such that $\mathbb{P}(c'\beta > 0) = 1$. This, in our terminology, means that:

Assumption 3.2. *The support of β is a subset of some hemisphere.*

As noted by Ichimura and Thompson (1998), Assumption 3.2 does not seem too stringent in many economic applications. It is often reasonable to assume that an element of the random coefficients vector, such as a price coefficient, has a known sign. If the j-th element of β has a known sign (and positive), then Assumption 3.2 holds with c being a unit vector with its j-th element being 1. This is a case in which the location of the hemisphere in Assumption 3.2 is known *a priori*, though the knowledge about its location is not necessary for identification. Assumption 3.2 implies the following mapping from f_β^- to f_β developed in (2.16):

$$f_\beta(b) = 2f_\beta^-(b)\mathbb{I}\left\{f_\beta^-(b) > 0\right\}, \tag{3.4}$$

it corresponds to the above case where $\frac{1}{|\mathbb{S}^{d-1}|}\left(1 - \int_{\mathbb{S}^{d-1}} |f_\beta^-(b)|d\sigma(b)\right) = 0$. This relation is useful because (i) it shows that Assumption 3.2 guarantees identification if f_β^- is identified, (ii) it enables us to derive a key formula that leads to a simple and practical estimation algorithm and (iii) it can be used to obtain an estimator for f_β that is guaranteed to be non-negative. Hence we have obtained

Proposition 3.1. *If Assumption 3.1 is satisfied and if r is such that its extension R belongs to $\mathrm{H}^{d/2}(\mathbb{S}^{d-1})$, then there exists a pdf f_β^* such that*

$$R = \mathcal{H}\left(f_\beta^*\right) = \frac{1}{2} + \mathcal{H}\left(f_\beta^{*-}\right)$$

and for all x in H^+, $r(x) = \mathcal{H}\left(f_\beta^*\right)(x)$. Moreover, if Assumption 3.2 holds then f_β is uniquely identified.

Remark 3.1. Assumption 3.2 is testable since it yields implications in terms of f_β^- which is identified under weak conditions. For example, we can compare the positivity of f_β^- with its negativity on the corresponding points on the opposite side of the sphere. Or, it implies that f_β^- integrates to

$1/(2|\mathbb{S}^{d-1}|)$ on H and $-1/(2|\mathbb{S}^{d-1}|)$ on $-H$. An estimator for f_β^- and its asymptotic properties are presented in the next section.

3.2. **Nonparametric Estimation of f_β.** If f_β^- belong to $\mathrm{H}^s(\mathbb{S}^{d-1})$ then R belongs to $\mathrm{H}^{d/2+s}(\mathbb{S}^{d-1})$. Moreover, if we use an estimator $\hat{R}^{-,N}$ of R^- to define

$$\hat{f}_\beta^{-,N} = \mathcal{H}^{-1}\left(\hat{R}^{-,N}\right) \tag{3.5}$$

$$= i_s * \hat{R}^{-,N}$$

$$= \sum_{p=0}^{\infty} \frac{1}{\lambda(2p+1,d)} \int_{\mathbb{S}^{d-1}} q_{2p+1,d}(\cdot,x)\hat{R}^{-,N}(x)d\sigma(x) \tag{3.6}$$

as an estimator of f_β^-, then Proposition 2.7 implies that for $v \in [0,s]$,

$$\|\hat{f}_\beta^{-,N} - f_\beta^-\|_{2,v} \asymp \|\hat{R}^{-,N} - R^-\|_{2,v+d/2}. \tag{3.7}$$

Also, setting

$$\hat{f}_\beta^N(b) = 2\hat{f}_\beta^{-,N}(b)\mathbb{I}\left\{\hat{f}_\beta^{-,N}(b) > 0\right\} \tag{3.8}$$

as suggested in Section 3.1, we obtain that

$$\|\hat{f}_\beta^N - f_\beta\|_2 \asymp \|\hat{R}^{-,N} - R^-\|_{2,d/2}. \tag{3.9}$$

This is explained in the proof of Theorem 5.1 of Section 5 given in the appendix. Thus, the rate of convergence for the estimation of f_β is directly related to the rate of convergence for the estimation of R^-. In particular, if R^- is estimated at the nonparametric rate

$$O_p\left(N^{-\frac{d/2+s-d/2}{2(d/2+s)+d-1}}\right) = O_p\left(N^{-\frac{s}{2s+2d-1}}\right)$$

in the $\|\cdot\|_{2,d/2}$-norm (recall that R^- is defined on a $d-1$ dimensional manifold), it implies estimation of f_β at the same rate in $\mathrm{L}^2(\mathbb{S}^{d-1})$.

As already mentioned, $d/2$ is the degree of ill posedness (the definition is different from the one in Kim and Koo (2000) where it would be $d/2-1$). It corresponds to the rate at which eigenvalues $|\lambda(2p+1,d)|$ converges to zero. This is very similar to results in standard deconvolution problems on $\mathbb{R}$ as obtained by Fan (1991).

We now give an example of an estimator $\hat{R}^{-,N}$ in Section 4 that implies a very simple closed form expression for $\hat{f}_\beta$ which avoids numerical evaluation of the integrals in (3.6). For other nonparametric estimators for R^-, numerical integration can be carried out to obtain $\hat{f}_\beta^N$. Also, we truncate the infinite sum in (3.6) at to some integer T, that is, we use the new estimator which could be written in the three equivalent forms

$$\hat{f}_{\beta}^{-,N} = \mathcal{H}^{-1}\left(P_{T_N}\hat{R}^{-,N}\right) \tag{3.10}$$

$$= i_s * \left(D_{T_N}(\cdot,\mathbf{n}) * \hat{R}^{-,N}\right)$$

$$= \sum_{p=0}^{T_N} \frac{1}{\lambda(2p+1,d)} \int_{\mathbb{S}^{d-1}} q_{2p+1,d}(\cdot,x)\hat{R}^{-,N}(x)d\sigma(x) \tag{3.11}$$

for suitably chosen T_N that goes to infinity with N. This approach amounts to the spectral cut-off method used in the statistics of inverse problems.

4. Estimators for the Choice Probability and Its Derivatives

We have seen so far that the model implies invariance by dilatation of the vector of covariates (augmented by 1) of the choice probability. Here we present estimators for its derivatives as well. They are, given x in $[0,\infty)\times\mathbb{R}^{d-1}$, the partial derivatives $\frac{\partial}{\partial x_j}R^{\check{}}$ which are the components of the gradient in the Euclidian space which satisfies

$$\nabla_x R^{\check{}} = \frac{1}{\|x\|}\nabla^S_{x/\|x\|}R.$$

Since R is square integrable, it has a condensed harmonic expansion which enables us to obtain the expressions in the next theorem.

Theorem 4.1. *We have for x in $\mathbb{S}^{d-1}$,*

$$R(x) = \frac{1}{2} + \sum_{p=0}^{\infty} \mathbb{E}\left[\frac{(2Y-1)}{f_X(X)}^{\flat} q_{2p+1,d}(X'x)\right] \tag{4.1}$$

and for x in $[0,\infty)\times\mathbb{R}^{d-1}$ and X on the sphere,

$$\nabla_x R^{\check{}} = \frac{d|\mathbb{S}^{d+1}|}{|\mathbb{S}^{d-1}|\|x\|} \sum_{p=0}^{\infty} \mathbb{E}\left[\frac{(2Y-1)}{f_X(X)}^{\flat} q_{2p,d+2}(X'x/\|x\|)X\right]. \tag{4.2}$$

Remark 4.1. Note that we can replace above $(2Y-1)$ by $2Y$ since $\int_{\mathbb{S}^{d-1}} q_{2p+1,d}(x,v)d\sigma(x) = 0$ for all v in $\mathbb{S}^{d-1}$. However, simulation results suggest that the symmetrization seems to provide better estimates.

This suggests an estimator of the form $\hat{R}^N(x) = \frac{1}{2} + \hat{R}^{-,N}$ with

$$\hat{R}^{-,N}(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{(2y_i-1)}{\hat{f}_X^N(x_i)} \sum_{p=0}^{T_N} q_{2p+1,d}(x_i,x)$$

where $\hat{f}_X^N$ is an estimator of f_X and T_N is a suitably chosen sequence converging to infinity with N. Note that the second summation corresponds to a truncated version of the Dirichlet kernel. We can generalize this, by introducing a class of estimators of the form

$$\hat{R}^{-,N}(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{(2y_i-1)}{\hat{f}_X^N(x_i)} K_{T_N}^{-}(x_i,x) \tag{4.3}$$

where $K_{T_N}^{-}$ is the odd part of a trigonometric kernel. For example, one can use a Cesàro kernel in place of the Dirichlet kernel in the definition (4.3).

Remark 4.2. Many other estimators of R^-, such as kernel regression estimator, are available. As noted before, however, we then need to use numerical integration to evaluate (3.11).

The treatment of the plug-in of $\hat{f}_X^N$ in (4.3) could be quite involved if one wishes to obtain the same rates of convergence for the plug-in estimator as the infeasible estimator that uses f_X under mild smoothness conditions on f_X. We choose $K_{T_N} = S_{T_N,d}^{\delta\ -}$ for $\delta \geq d-1$. Here the kernels are uniformly bounded and non-negative. Because of (2.12) and (2.13), if we use the Dirichlet kernel instead, achieving such a rate of convergence for the plug-in estimator seems to require very stringent assumptions on the smoothness of f_X.

We could consider the following two cases

(I) $\exists\, m_X > 0 :\ \forall x \in H^+,\ f_X(x) \geq m_X$

(II) Assumption 3.1 is satisfied but condition (I) is not.

Condition (I) is technical but not realistic for usual distributions of $\tilde{X}$ in $\mathbb{R}^d$ (see, e.g. Hoderlein et al., 2007).

Remark 4.3. We need to make an assumption of the type of Assumption 4.1 below in order that the estimator converges fast enough to f_X. It usually requires that f_X belongs to $\mathrm{H}^{\sigma}(\mathbb{S}^{d-1})$ where σ is large enough. When f_X is bounded from below on H^+ it is for example impossible that it is continuous even though it is on the interior of H^+.

We now restrict ourselves to the case (II) as it is more relevant. Since f_X is not bounded from below, we use a trimmed version of (4.3)

$$\hat{R}^{-,N}(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{(2y_i-1) S_{2T_N+1}^{\delta\ -}(x_i,x)}{\max\left(\hat{f}_X^N(x_i), a_N\right)} \tag{4.4}$$

with a sequence of the form

$$a_N = \log(N)^{-r} \tag{4.5}$$

for some positive r, and define

$$\widehat{R}^N = \frac{1}{2} + \hat{R}^{-,N}. \tag{4.6}$$

Concerning derivatives we use the estimator

$$\widehat{\nabla_x \check{R}}^N = \nabla_x \left(\left(\widehat{R}^N \right)\check{} \right) = \frac{d|\mathbb{S}^{d+1}|}{N|\mathbb{S}^{d-1}| \|x\|} \sum_{i=1}^{N} \frac{2y_i - 1}{\max\left(\hat{f}_X(x_i), (\log N)^{-r} \right)} {}^\flat S^{\delta\,+}_{2T_N, d+2}(x_i' x / \|x\|) x_i. \tag{4.7}$$

For a mathematical treatment of the plug-in procedure it is very convenient if both ${}^\flat S^{\delta}_{2T_N,d+2}$ and ${}^\flat S^{\delta}_{2T_N+1,d}$ are non-negative. This can be achieved by taking $\delta = d+1$. If one only wants to estimate R, $\delta = d-1$ provides enough smoothing, while in order to be able to estimate derivatives it is useful to work with higher order kernels involving higher order Cesàro summation.

Estimation of densities on compact manifolds have been studied by several authors, using histogram (Ruymgaart (1989)), projection estimators (see, e.g. Devroye and Gyorfi (1985) for the circle and Hendriks (1990) for general compact Riemannian manifolds) or kernel estimators (see, e.g. Devroye and Gyorfi (1985) for the case of the circle, Hall et al. (1987) and Klemelä (2000) for higher dimensional spheres). We now assume that the following holds for f_x and its estimator $\hat{f}_X^N$.

Assumption 4.1. *f_X is smooth enough and its estimator $\hat{f}_X^N$ is such that, depending on the type of result,* (i) *or* (ii) *holds:*

(*i*)

$$\max_{i=1,\dots,N} \left| \frac{\max\left(f_X(x_i), \log(N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i), \log(N)^{-r}\right)} - 1 \right| = O_p\left(\left(\frac{N}{(\log N)^{2r + (1-2/q)\mathbb{I}\{q \geq 2\}}} \right)^{-\frac{\sigma}{2\sigma+d-1}} (\log N)^{-r} \right)$$

(*ii*) *There is a constant C such that*

$$\overline{\lim}_{N\to\infty} \left(\frac{N}{(\log N)^{2r}} \right)^{\frac{\sigma}{2\sigma+d-1}} (\log N)^r \max_{i=1,\dots,N} \left| \frac{\max\left(f_X(x_i), \log(N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i), \log(N)^{-r}\right)} - 1 \right| \leq C \quad a.s.$$

for some r, q and σ that will be specified later.

This rate can easily be achieved when f_X is smooth enough. In Section 7 we use

$$\hat{f}_X^N(x) = \frac{1}{N} \sum_{i=1}^{N} S^{d-1}_{T_N,d}(x_i, x) \tag{4.8}$$

for a suitably chosen T_N that depends on the sample size and the smoothness of f_X. Theoretical properties of this estimator will appear elsewhere but note that its rate of convergence in sup-norm can be obtained in a similar manner as here in the proof of Theorem 5.1. This estimator is in the spirit of the projection estimators of Hendriks (1990) but here we are able to obtain a closed form using the condensed harmonic expansions together with the Addition Formula and consider a modification of the Dirichlet kernel in order to have a *bona fide* approximation kernel.

We now present the asymptotic properties of the estimators for R and its derivative. The proofs are very similar to those of Theorems 5.1 and 5.2 of Section 5 given in the appendix and thus omitted. We first state results on the rates of convergence, including strong uniform convergence rates. Besides the log correction due to trimming of f_X, the rate is comparable to the usual nonparametric rates.

Theorem 4.2 (Convergence rates in $\mathrm{L}^q(\mathbb{S}^{d-1})$)**.** *Suppose Assumptions 3.1, 4.1* (i) *and condition* (II) *hold. If R belongs to $\mathrm{W}^{\sigma}_q(\mathbb{S}^{d-1})$ with q in $[1,\infty)$ and σ positive, T_N satisfies*

$$T_N \asymp \left(\frac{N}{(\log N)^{2r+(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{\frac{1}{2\sigma+d-1}},$$

and for a positive r

$$\sigma\left(\left\{0<f_X<(\log N)^{-r}\right\}\right)=o\left(\left(\frac{N}{(\log N)^{2r+(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{-\frac{\sigma+(d-1)(1-1/q)}{2\sigma+d-1}}\right)$$

holds, then

$$\left\|\hat{R}^N-R\right\|_q=O_p\left(\left(\frac{N}{(\log N)^{2r+(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{-\frac{\sigma}{2\sigma+d-1}}\right),$$

$$\forall j=1,\dots,d,\ \left\|\widehat{\frac{\partial}{\partial x_j}R^{\check{}}}^{N}-\frac{\partial}{\partial x_j}R^{\check{}}\right\|_q=O_p\left(\left(\frac{N}{(\log N)^{2r+(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{-\frac{\sigma-1}{2\sigma+d-1}}\right).$$

Moreover, if Assumption 4.1 (ii) *holds then there exists a constant C such that*

$$\overline{\lim}_{N\to\infty}\left(\frac{N}{(\log N)^{-2r-1}}\right)^{\frac{\sigma}{2\sigma+d-1}}\left\|\hat{R}^N-R\right\|_\infty\leq C\quad a.s.$$

$$\overline{\lim}_{N\to\infty}\left(\frac{N}{(\log N)^{-2r-1}}\right)^{\frac{\sigma-1}{2\sigma+d-1}}\left\|\widehat{\frac{\partial}{\partial x_j}R^{\check{}}}^{N}-\frac{\partial}{\partial x_j}R^{\check{}}\right\|_\infty\leq C\quad a.s.$$

Theorem 4.3 (Asymptotic normality)**.** *Suppose R belongs $\mathrm{W}^{\sigma}_{\infty}(\mathbb{S}^{d-1})$ with σ positive and Assumption 3.1 and condition* (II) *hold. If f_X, $\hat{f}_X^N$, T_N and r satisfy*

$$\max_{i=1,\dots,N}\left|\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right|=O_p\left((\log N)^{-2r}\right),$$

$$T_N=O\left(\left(\frac{N}{(\log N)^{2r}}\right)^{1/(d-1)}\right),$$

$$T_N N^{-\frac{1}{2\sigma+d-1}}=o(1)\qquad\text{(under smoothing)},$$

$$N^{1/2}T_N^{(d-1)/2}\sigma\left(\left\{0<f_X<(\log N)^{-r}\right\}\right)=o(1),$$

then

$$N^{\frac{1}{2}}s_{1N}^{-1}\left(\hat{R}^N(x)-R(x)\right)\xrightarrow{d}N(0,1)$$

and

$$N^{\frac{1}{2}}s_{2N}^{-1}\left(\frac{d|\mathbb{S}^{d+1}|}{|\mathbb{S}^{d-1}|\|x\|}\right)^{-1}\left(\widehat{\nabla_x R^{\check{}}}^N(x)-\nabla_x R^{\check{}}\right)\xrightarrow{d}N(0,1)$$

where

$$s_{1N}^2:=\operatorname{var}\left(\frac{(2Y_i-1)^{\flat}S^{\delta\ -}_{2T_N+1,d}(X_i'x)}{\max\left(f_X(X_i),(\log N)^{-r}\right)}\right)$$

$$s_{2N}^2:=\operatorname{var}\left(\frac{(2Y_i-1)^{\flat}S^{\delta\ +}_{2T_N,d+2}(X_i'x)}{\max\left(f_X(X_i),(\log N)^{-r}\right)}X_i\right)$$

5. A Closed Form Estimator of f_β

This section considers

$$\hat{f}_\beta^{-,N}=\mathcal{H}^{-1}\left(\hat{R}^{-,N}\right)$$

as an estimator of f_β^-. Note that the estimator (4.4) lives in a finite dimensional space, that is, $P_{T_N}\hat{R}^{-,N}=\hat{R}^{-,N}$ holds, therefore we do not need additional spectral cut-off prior to the inversion of $\mathcal{H}$. As far as we are only interested in f_β and not in its derivatives, we can choose $\delta=d-1$ below. The estimator takes a simple closed form and requires no numerical integration since the formula

$$\mathcal{H}^{-1}\left(S^{\delta\ -}_{2T_N+1}(x_i,\cdot)\right)(b)=\frac{1}{A^{\delta}_{T_N}}\sum_{p=0}^{T_N}\frac{A^{\delta}_{2(T_N-p)}}{\lambda(2p+1,d)}q_{2p+1,d}(x_i,b)$$

can be used in

$$\hat{f}_{\beta}^{-,N}(b) = \frac{1}{N}\sum_{i=1}^{N} \frac{(2y_i - 1)\mathcal{H}^{-1}\left(S^{\delta}_{2T_N+1}{}^{-}(x_i,\cdot)\right)(b)}{\max\left(\hat{f}_X^N(x_i), a_N\right)}$$

The final estimator of f_β is obtained using (3.8).

The proof of the following result is given in the appendix.

Theorem 5.1 (Convergence rates in $\mathrm{L}^q(\mathbb{S}^{d-1})$)**.** *Suppose condition* (II), *Assumptions 3.1 and 4.1 with* $\sigma = s + \frac{d}{2}$ *hold. If* f_β^- *belongs to* $\mathrm{W}^s_q(\mathbb{S}^{d-1})$ *with* q *in* $[1,\infty)$ *and* $s > 0$, T_N *satisfies*

$$T_N \asymp \left(\frac{N}{(\log N)^{2r+(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{\frac{1}{2s+2d-1}}$$

and for a positive r

$$\sigma\left(\left\{0 < f_X < (\log N)^{-r}\right\}\right) = O\left(\left(\frac{N}{(\log N)^{2r+(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{-\frac{s+d/2+(d-1)(1-1/q)}{2s+2d-1}}\right) \tag{5.1}$$

holds, then

$$\left\|\hat{f}_\beta^N - f_\beta\right\|_q = O_p\left(\left(\frac{N}{(\log N)^{2r+(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{-\frac{s}{2s+2d-1}}\right). \tag{5.2}$$

Moreover, if Assumption 4.1 (ii) *holds then there exists a constant* C *such that*

$$\overline{\lim}_{N\to\infty}\left(\frac{N}{(\log N)^{-2r-1}}\right)^{\frac{s}{2s+2d-1}}\left\|\hat{f}_\beta^N - f_\beta\right\|_\infty \leq C \quad a.s. \tag{5.3}$$

The rate $N^{-\frac{s}{2s+2d-1}}$ is in accordance with the L^2 rate in Healy and Kim (1996) who study deconvolution on $\mathbb{S}^2$ for non degenerate kernels. Kim and Koo (2000) prove that the rate in Healy and Kim (1996) is optimal in the minimax sense. Their statistical problem though does not involve plug-in and trimming. Also, somewhat less importantly, it does not cover the case when the convolution kernel is given by an indicator function, which appears in our operator $\mathcal{H}$. Hoderlein et al. (2007) study estimation of densities in a linear model with random coefficients and obtain the same rate when f_X is bounded from below and thus no trimming is required (we need to replace their dimension d by our dimension $d-1$). The upper bound on the rate of consistency is logarithmicaly close to that rate and the closeness all depends on the decay to zero of the density $f_X(x)$ as x approaches the boundary of H^+.

The next theorem is concerned with pointwise asymptotic normality. The proof is given in the appendix.

Theorem 5.2 (Asymptotic normality)**.** *Suppose f_β^- belongs to $\mathrm{W}^s_\infty(\mathbb{S}^{d-1})$ with $s>0$, and Assumption 3.1 and* (II) *holds. If $\hat{f}_X^N$, T_N and r satisfy*

$$\max_{i=1,\dots,N}\left|\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right|=O_p\left((\log N)^{-2r}\right), \tag{5.4}$$

$$T_N=O\left(\left(\frac{N}{(\log N)^{2r}}\right)^{1/(d-1)}\right), \tag{5.5}$$

$$T_N N^{-\frac{1}{2s+2d-1}}=o(1), \tag{5.6}$$

$$N^{1/2}T_N^{(d-1)/2}\sigma\left(\left\{0<f_X<(\log N)^{-r}\right\}\right)=o(1), \tag{5.7}$$

then

$$N^{\frac{1}{2}}s_N^{-1}\left(\hat{f}_\beta^N(b)-f_\beta(b)\right)\xrightarrow{d}N(0,1) \tag{5.8}$$

holds for b such that $f_\beta(b)\neq 0$, where $s_N^2:=4\mathrm{var}(Z_{N,i})$, $Z_{N,i}=\frac{(2Y_i-1)\mathcal{H}^{-1}\left(S^{\delta}_{2\overline{T}_N+1}(X_i,\cdot)\right)(b)}{\max(f_X(X_i),(\log N)^{-r})}$.

Condition (5.4) is very mild. The rate in Condition (5.6) is faster than the optimal rate (undersmoothing), which guarantees that the bias is asymptotically negligible.

6. Discussion

6.1. **Estimation of Marginals.** In Section 3 we have provided an expression for the estimator of the full joint density of β, from which an estimator for a marginal density can be obtained. Let $d\sigma_k$ denote the surface measure and $d\underline{\sigma}_k=d\sigma_k/|\mathbb{S}^k|$ the uniform measure on $\mathbb{S}^k$. We write $\beta=\left(\overline{\beta}',\overline{\overline{\beta}}'\right)'$ and wish to obtain the density of the marginal of $\overline{\beta}$ which is a vector of dimension $d-k$. We also define $\overline{P}$ and $\overline{\overline{P}}$ the projectors such that $\overline{\beta}=\overline{P}\beta$ and $\overline{\overline{\beta}}=\overline{\overline{P}}\beta$ and denote by $d\overline{P}_*\underline{\sigma}_{d-1}$ and $d\overline{\overline{P}}_*\underline{\sigma}_{d-1}$ the direct image probability measures. One possibility is to define the marginal law of $\overline{\overline{\beta}}$ as the measure $\overline{\overline{P}}_*f_\beta d\sigma$. This may not be convenient, however, since then a uniform distribution would have U-shaped marginals. The U-shape becomes more pronounced as the dimension of β increases. In order to obtain a flat density for the marginals of the uniform joint distribution on the sphere it is enough to consider densities with respect to the dominating measure $d\overline{\overline{P}}_*\underline{\sigma}_{d-1}$. Notice that sampling U uniformly on $\mathbb{S}^{d-1}$ is equivalent to sampling $\overline{\overline{U}}$ according to $\overline{\overline{P}}_*\underline{\sigma}_{d-1}$ and then given $\overline{\overline{U}}$ forming $\rho\left(\overline{\overline{U}}\right)V$ where V is a draw from the uniform distribution $\underline{\sigma}_{d-1-k}$ on $\mathbb{S}^{d-1-k}$ and $\rho\left(\overline{\overline{U}}\right)=\sqrt{1-\left\|\overline{\overline{U}}\right\|^2}$. Indeed given $\overline{\overline{U}}$, $\overline{U}/\rho\left(\overline{\overline{U}}\right)$ is uniformly distributed on $\mathbb{S}^{d-1-k}$. Thus, when g is an element of $\mathrm{L}^1(\mathbb{S}^{d-1})$ we can write for k in $\{1,\dots,d-1\}$,

$$\int_{\mathbb{S}^{d-1}} g(b)d\underline{\sigma}_{d-1}(b) = \int_{\mathbb{B}^k}\left[\int_{\mathbb{S}^{d-1-k}} g\left(\rho\left(\overline{\overline{b}}\right)u,\overline{\overline{b}}\right)d\underline{\sigma}_{d-1-k}(u)\right]d\overline{\overline{P}}_{*}\underline{\sigma}_{d-1}\left(\overline{\overline{b}}\right) \tag{6.1}$$

where $\mathbb{B}^k$ is the k dimensional ball of radius 1. Setting $g = |\mathbb{S}^{d-1}|f_\beta(b)\mathbb{I}\left\{\overline{\overline{b}}\in A\right\}$ for A Borel set of $\mathbb{B}^k$ shows that the marginal density of $\overline{\overline{\beta}}$ with respect to the dominating measure $d\overline{\overline{P}}_{*}\underline{\sigma}_{d-1}$ is given by

$$f_{\overline{\overline{\beta}}}\left(\overline{\overline{b}}\right) = |\mathbb{S}^{d-1}|\int_{\mathbb{S}^{d-1-k}} f_\beta\left(\rho\left(\overline{\overline{b}}\right)u,\overline{\overline{b}}\right)d\underline{\sigma}_{d-1-k}(u). \tag{6.2}$$

In the particular case where $k=d-1$, *i.e.* we are interested in the marginal of $\tilde{\beta}$, we use $d\underline{\sigma}_0 = (\delta_1+\delta_{-1})/2$ where δ denotes the Dirac mass.

When the dimension of the variables in the integral is small we can use hyperspherical parametrization (polar coordinates when $k=d-2$ and spherical coordinates when $k=d-3$) and deterministic numerical integration methods. When it is not, one may use Monte-Carlo methods, by forming

$$\hat{f}^{N,T,M}_{\overline{\overline{\beta}}}\left(\overline{\overline{b}}\right) = \frac{1}{M}\sum_{j=1}^{M}\hat{f}^{N,T}_{\beta}\left(\rho\left(\overline{\overline{b}}\right)u_j,\overline{\overline{b}}\right) \tag{6.3}$$

where u_j are draws from independent uniform random variables on $\mathbb{S}^{d-1-k}$. Draws u_j could be obtained by computing $u_j = v_j/\|v_j\|$ where v_j are draws from a standard Gaussian random vector of dimension $d-1-k$. When $\overline{\overline{\beta}}$ is of dimension 2 we could draw contour plots on the disk, that is, level sets of the density. When β is of dimension 3 it is possible to draw contour plots on $\mathbb{S}^2$.

6.2. **Treatment of non-random coefficients.** It may be useful to develop an extension of the method described in the previous sections to models that have non-random coefficients, at least for two reasons. First, the convergence rate of our estimator of the joint density of β slows down as the dimension d of β grows, which is a manifestation of the curse of dimensionality. Treating some coefficient as fixed parameters alleviates this problem. Second, our identification assumption in Section 3.1 precludes covariates with discrete or bounded support. This may not be desirable as many random coefficient discrete choice models in Economics involve dummy variables as covariates. The following identification/estimation strategy allows such covariates as far as their coefficients are non-random. Note that Hoderlein et al. (2007) suggest a method to deal with non-random coefficients in their treatment of random coefficient linear regression models. Identification in random coefficient binary choice models with covariates with limited support is somewhat tricky. As we shall see shortly, identification is possible in a model where the coefficients on covariates with limited support are non-random, provided that at least one of the covariates with "large support" has a non-random coefficient as well. More precisely, consider the model:

$$Y_i = \mathbb{I}\{\beta_{1i} + \beta_{2i}' X_{2i} + \alpha_1 Z_{1i} + \alpha_2' Z_{2i} \geq 0\} \tag{6.4}$$

where $\beta_1 \in \mathbb{R}$ and $\beta_2 \in \mathbb{R}^{d_X - 1}$ are random coefficients, whereas the coefficients $\alpha_1 \in \mathbb{R}$ and $\alpha_2 \in \mathbb{R}^{d_Z - 1}$ are nonrandom. The covariate vector $(Z_1, Z_2')'$ is in $\mathbb{R}^{d_Z}$, though the $(d_Z - 1)$-subvector z_2 might have limited support: for example, it can be a vector of dummies. The covariate vector $(X_2', Z_1)'$ is assumed to be, among other things, continuously distributed. Normalizing the coefficients vector and the vector of covariates to be elements of the unit sphere works well for the development of our procedure, as we have seen in the prevous sections. The model (6.4), however, is presented "in the original scale" to avoid confusion.

Define $\beta_1^*(Z_2) := \beta_1 + \alpha_2' Z_2$, $\tau(Z_2) = (\beta_1^*(Z_2), \alpha_1, \beta_2)'$ and $W = (1, Z_1, X_2')'$. We also use the notation

$$\tau(Z_2) := \frac{(\beta_1^*(Z_2), \alpha_1, \beta_2)'}{\|(\beta_1^*(Z_2), \alpha_1, \beta_2')\|} \in \mathbb{S}^{d_X+1}, W := \frac{(1, Z_1, X_2')'}{\|(1, Z_1, X_2')'\|} \in \mathbb{S}^{d_X+1}.$$

Then (6.4) is equivalent to:

$$\begin{aligned} Y &= \mathbb{I}\{\beta_1^*(Z_2) + (\alpha_1, \beta_2)(Z_1, X_2')' \geq 0\} \\ &= \mathbb{I}\{(\beta_1^*(Z_2), \alpha_1, \beta_2)(1, Z_1, X_2')' \geq 0\} \\ &= \mathbb{I}\left\{\frac{(\beta_1^*(Z_2), \alpha_1, \beta_2)}{\|(\beta_1^*(Z_2), \alpha_1, \beta_2)'\|} \frac{(1, Z_1, X_2')'}{\|(1, Z_1, X_2')'\|} \geq 0\right\} \\ &= \mathbb{I}\left\{\tau(Z_2)' W \geq 0\right\}. \end{aligned}$$

This has the same form as our original model if we condition on $Z_2 = z_2$. We can then apply previous results for identification and estimation under the following assumptions. First, suppose $(\beta_1, \beta_2')'$ and W are independent, instead of Assumption 1.1. Second, we impose some condition on $f_{W|Z_2=z_2}$, the conditional density of W given $Z_2 = z_2$. More specifically, suppose there exists a set $\mathcal{Z}_2 \in \mathbb{R}^{d_Z-1}$, such that Assumption 3.1 holds if we replace f_X and d with $f_{W|Z_2=z_2}$ and $d_X + 1$ for all $z_2 \in \mathcal{Z}_2$. If Z_2 is a vector of dummies, for example, $\mathcal{Z}_2$ would be a discrete set. By (4.1) and (2.18) we obtain

$$f^-_{\tau(Z_2)|Z_2=z_2}(t) = \sum_{p=0}^{\infty} \frac{1}{\lambda(2p+1, d_X+1)} \mathbb{E}\left[\left. \frac{(2Y-1) q_{2p+1,d_X}(W,t)}{f_{W|Z_2=z_2}(W)} \right| Z_2 = z_2\right] \tag{6.5}$$

for all $z_2 \in \mathcal{Z}_2$, where the right hand side consists of observables. This determines $f_{\tau(Z_2)|Z_2=z_2}$. That is, the conditional density

$$f\left(\left. \frac{(\beta_1^*(Z_2), \alpha_1, \beta_2)}{\|(\beta_1^*(Z_2), \alpha_1, \beta_2)'\|} \right| Z_2 = z_2\right)$$

is identified for all $z_2 \in \mathcal{Z}_2$ (Here and henceforth we use the notation $f(\cdot|\cdot)$ to denote conditional densities with appropriate arguments when adding subscripts is too cumbersome). This obviously identifies

$$f\left(\left.\frac{(\beta_1^*(Z_2), \alpha_1, \beta_2)}{\|\beta_2\|}\right| Z_2 = z_2\right) \tag{6.6}$$

for all $z_2 \in \mathcal{Z}_2$ as well. If we are only interested in the joint distribution of β_2 under a suitable normalization, we can stop here. The presence of the term $\alpha_1 Z_1$ in (6.4) is unimportant so far.

Some more work is necessary, however, if one is interested in the joint distribution of the coefficients on all the regressors. Notice that the distribution (6.6) gives

$$f\left(\left.\frac{\beta_1^*(Z_2)}{\|\beta_2\|}\right| Z_2 = z_2\right) = f\left(\left.\frac{\beta_1 + \alpha_2' Z_2}{\|\beta_2\|}\right| Z_2 = z_2\right),$$

from which we can, for example, get

$$\mathbb{E}\left(\left.\frac{\beta_1^*(Z_2)}{\|\beta_2\|}\right| Z_2 = z_2\right) = \mathbb{E}\left(\frac{\beta_1}{\|\beta_2\|}\right) + \mathbb{E}\left(\frac{1}{\|\beta_2\|}\right)\alpha_2' z_2 \quad \text{for all } z_2 \in \mathcal{Z}_2.$$

Define a constant

$$c := \mathbb{E}\left(\frac{1}{\|\beta_2\|}\right)$$

then we can identify $c\alpha_2$ as far as $z_2 \in \mathcal{Z}_2$ has enough variation and

$$\mathbb{E}\left(\frac{\alpha_1}{\|\beta_2\|}\right) = c\alpha_1$$

is identified as well. Let

$$f\left(\frac{(\beta_{2i}', \alpha_1, \alpha_2')'}{\|\beta_{2i}\|}\right) \tag{6.7}$$

denote the joint density of all the coefficient (except for β_1, which corresponds to the conventional disturbance term in the original model (6.4), normalized by the length of β_{2i}). Then

$$f\left(\frac{(\beta_{2i}', \alpha_1, \alpha_2')'}{\|\beta_{2i}\|}\right) = f\left(\begin{bmatrix} I_{d_X-1} & 0 \\ 0 & 1 \\ \vdots & \frac{c\alpha_2}{c\alpha_1} \end{bmatrix} \begin{bmatrix} \frac{\beta_{2i}}{\|\beta_{2i}\|} \\ \frac{\alpha_1}{\|\beta_{2i}|} \end{bmatrix}\right).$$

In the expression on the right hand side, $f\left((\beta_{2i}', \alpha_1)'/\|\beta_{2i}\|\right)$ is already available from (6.6), and $c\alpha_1$ and $c\alpha_2$ are identified already, therefore the desired joint density (6.7) is identified. Obviously (6.7) also determines the joint density of $(\beta_{2i}', \alpha_1, \alpha_2')'$ under other suitable normalizations as well.

The density (6.5) is estimable: when Z_2 is discrete, one can use the estimator of Section 5 to each subsample corresponding to each value of Z_2. If Z_2 is continuous we can estimate $f_{W|z_2}$ and the conditional expectation by nonparametric smoothing. An estimator for the density (6.6) can be then obtained numerically.

6.3. **Endogenous Regressors.** Assumption 1.1 is violated if some of the regressors are endogenous in the sense that the random coefficients and the covariates are not independent. This problem can be solved if an appropriate vector of instruments is available. To be more specific, suppose we observe (Y, X, Z) generated from the following model

$$Y = \mathbb{I}\{\beta_1 + \tilde{\beta}' X \geq 0\} \tag{6.8}$$

with

$$X = \Gamma Z + V \tag{6.9}$$

where V is a vector of reduced form residuals and Z is independent of (β, V). The equations (6.8) and (6.9) yield

$$Y = \mathbb{I}\{\left(\beta_1 + V'\tilde{\beta}\right) + Z'\Gamma'\tilde{\beta}\}.$$

Suppose the distribution of ΓZ satisfy Assumption 3.1. It is then possible to estimate the density of $\overline{\tau} = \tau/\|\tau\|$ where $\tau = \left(\beta_1 + V'\tilde{\beta}, \tilde{\beta}\right)'$ by replacing Γ with a consistent estimator, which is easy to obtain under the maintained assumptions. This yields an estimator for the joint density of $\tilde{\beta}/\|\tau\|$, the random coefficients on the covariates under scale normalization.

7. Application

To be added.

8. Conclusion

To be added.

9. Appendix

Let us start by recalling some notions of Riemannian geometry to enlighten the notions of gradient and Laplacian on the sphere. The tangent space $T_x\mathbb{S}^{d-1}$ at a point x on the sphere is the vector space of tangents $\frac{\mathrm{d}}{\mathrm{d}t}\gamma(t)\big|_{t=0}$ of curves $\gamma : (-\epsilon, \epsilon) \to U$ where $\epsilon > 0$ and U is a neighborhood of x in $\mathbb{R}^d$, drawn on $\mathbb{S}^{d-1}$. We can easily check that it is the orthogonal in $\mathbb{R}^d$ of x. Given a map f from $\mathbb{S}^{d-1}$ to $\mathbb{R}$, its differential $d_x f$ at x in $\mathbb{R}^d$ is a linear form acting on $T_x\mathbb{S}^{d-1}$. It is such that for h in $T_x\mathbb{S}^{d-1}$ corresponding to a curve γ, $d_x f.h$ is defined as $\frac{\mathrm{d}}{\mathrm{d}t}[f(\gamma)]\big|_{t=0}$. A useful example in the case of derivatives of choice probabilities is the height function, see do Carmo (1976) p.86, defined for z in $\mathbb{S}^{d-1}$ as $x \in \mathbb{S}^{d-1} \mapsto z'x$. Its differential is the mapping

$$h \in T_x\mathbb{S}^{d-1} \mapsto z'h. \tag{9.1}$$

As in the Euclidian plane, the gradient on the sphere is related to the above defined differential using the scalar product. The gradient of f at x is denoted by $\nabla_x^S f$ and defined as the vector of $T_x\mathbb{S}^{d-1}$ such that for h in $T_x\mathbb{S}^{d-1}$, $\nabla_x^S f'h = d_x f.h$. The scalar product on the tangent spaces is the restriction of the scalar product in $\mathbb{R}^d$. This is a general construction of a gradient on smooth submanifolds of $\mathbb{R}^d$. It matches in the particular case of the sphere the definition provided by identity (2.6). The Laplace operator on a smooth submanifolds of $\mathbb{R}^d$ is defined through the generalization of the formula div∇. The generalization of the divergence is defined as follows. A vector field X is a map which to x in $\mathbb{S}^{d-1}$ assigns a vector $X(x)$ of $T_x\mathbb{S}^{d-1}$. It is differentiable if given a local parametrization of $\mathbb{S}^{d-1}$, for example using the stereographic projection, consisting of to maps φ from an open set U in $\mathbb{R}^{d-1}$ to $V \cap \mathbb{S}^{d-1}$ where V is an open set of $\mathbb{R}^d$, $X(\varphi)$ is differentiable. The linear mapping which to v in $T_x\mathbb{S}^{d-1}$ corresponding to some curve $\gamma(-\epsilon, \epsilon) \to U$ and X a vector field, assigns the orthogonal projection of $\frac{\mathrm{d}}{\mathrm{d}t}X(\gamma)\big|_{t=0}$ on $T_x\mathbb{S}^{d-1}$ is denoted by D. Then Δ^S is defined as tr$D\nabla^S$. Also, see for example Gallot et al (2004) p.209, we have

$$-\int_{\mathbb{S}^{d-1}} f(x)\Delta^S f(x) d\sigma(x) = \int_{\mathbb{S}^{d-1}} |||df_x|||^2 d\sigma(x) = \int_{\mathbb{S}^{d-1}} \nabla_x^S f' \nabla_x^S f d\sigma(x) \tag{9.2}$$

where $|||\cdot|||$ denotes the operator norm. We can check using the condensed harmonic expansion, Lemma 2.1 (ii) and relation (9.2) that

$$\|f\|_{2,1}^2 = \|f\|_2^2 + \|\nabla^S f\|_2^2$$

where the last term denotes the right hand-side of (9.2). The definition of the Sobolev spaces based on $\mathrm{L}^2(\mathbb{S}^{d-1})$ matches the classical space defined in terms of derivatives.

We now present some results on the Gegenbauer polynomials. These results can be found in Erdélyi et al. (1953) and Groemer (1996). The Gegenbauer polynomials have the following explicit representation

$$C_n^\nu(t) = \sum_{l=0}^{[n/2]} \frac{(-1)^l(\nu)_{n-l}}{l!(n-2l)!}(2t)^{n-2l} \tag{9.3}$$

where $(a)_0 = 1$ and for n in $\mathbb{N} \setminus \{0\}$, $(a)_n = a(a+1)\cdots(a+n-1) = \Gamma(a+n)/\Gamma(a)$. When $\nu = 0$, case $d = 2$, it is related to the Chebychev polynomials of the first kind as follows

$$\forall n \in \mathbb{N} \setminus \{0\},\ C_n^0(t) = \frac{2}{n} T_n(t)$$

and

$$C_0^0(t) = T_0(t) = 1$$

where

$$\forall n \in \mathbb{N},\ T_n(t) = \cos\left(n \arccos(t)\right).$$

When $\nu = 1$, case $d = 4$, $C_n^1(t)$ coincides with the Chebychev polynomial of the second kind $U_n(t)$ which is such that

$$\forall n \in \mathbb{N},\ U_n(t) = \frac{\sin[(n+1)\arccos(t)]}{\sin[\arccos(t)]}.$$

The Gegenbauer polynomials are stable by differentiation, they satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t} C_n^\nu(t) = 2\nu C_{n-1}^{\nu+1}(t) \tag{9.4}$$

for $\nu > 0$ and

$$\frac{\mathrm{d}}{\mathrm{d}t} C_n^0(t) = 2C_{n-1}^1(t). \tag{9.5}$$

For $\nu \neq 0$ the Rodrigues formula states that

$$C_n^\nu(t) = (-2)^{-n}(1-t^2)^{-\nu+1/2} \frac{(2\nu)_n}{(\nu+1/2)_n n!} \frac{\mathrm{d}^n}{\mathrm{d}t^n}(1-t^2)^{n+\nu-1/2}. \tag{9.6}$$

The following results are also used in the paper

$$\sup_{t\in[-1,1]} \left| \frac{C_n^\nu(t)}{C_n^\nu(1)} \right| \leq 1, \tag{9.7}$$

$$\forall\ \nu > 0,\ \forall n \in \mathbb{N},\ C_n^\nu(1) = \begin{pmatrix} n+2\nu-1 \\ n \end{pmatrix} \tag{9.8}$$

$$C_0^0(1) = 1 \text{ and } \forall n \in \mathbb{N} \setminus \{0\},\ C_n^0(1) = \frac{2}{n}, \tag{9.9}$$

$$C_n^\nu(-t) = (-1)^n C_n^\nu(t). \tag{9.10}$$

The normalization of these orthogonal polynomials is such that

$$\| C_n^{\nu(d)}\left(x'\cdot\right) \|_2 = \int_{-1}^1 (C_n^{\nu(d)}(t))^2 (1-t^2)^{(d-3)/2} dt = \frac{|\mathbb{S}^{d-1}|(C_n^{\nu(d)}(1))^2}{|\mathbb{S}^{d-2}| h(n,d)}. \tag{9.11}$$

In the proofs of the results we denote by C any constant depending only on the dimension, it thus takes different values for different inequalities.

Lemma 9.1. *For p positive and $d \geq 2$,*

$$\frac{\mathrm{d}}{\mathrm{dt}}\left({}^{\flat}q_{2p+1,d}\right) = \frac{d|\mathbb{S}^{d+1}|}{|\mathbb{S}^{d-1}|}q_{2T,d+2}$$

Proof. Using (2.9), (9.4), (9.5), (9.8) and (2.7)

$$\begin{aligned}\left(\frac{\mathrm{d}}{\mathrm{dt}}\left({}^{\flat}q_{2p+1,d}\right)\right)(t) &= \frac{h(2p+1,d)}{|\mathbb{S}^{d-1}|C_{2p+1}^{\nu(d)}(1)}(d-2)C_{2p}^{\nu(d)+1}(t)\\ &= \frac{4p+d}{|\mathbb{S}^{d-1}|(d-2)}(d-2)C_{2p}^{\nu(d+2)}(t).\end{aligned}$$

We conclude since, using again (9.8) and (2.7),

$$\frac{h(2p,d+2)}{C_{2p}^{\nu(d+2)}(1)} = \frac{4p+d}{d}.$$

□

Proof of Proposition 2.3. Using Lemma 9.1 and the expression of the Cesàro kernel we obtain

$$\frac{\mathrm{d}}{\mathrm{dt}}\left({}^{\flat}C_{2T+1,d}^{\delta\ -}\right) = \frac{d|\mathbb{S}^{d+1}|}{|\mathbb{S}^{d-1}|}C_{2T,d+2}^{\delta\ +}.$$

Using also (2.15), the following inequalities for some constant K depending only on δ and d follow

$$\begin{aligned}S_{T,d}^{\delta}(z'x) - S_{T,d}^{\delta}(z'y) &= \int_{z'x}^{z'y}\left(\frac{\mathrm{d}}{\mathrm{dt}}S_{T,d}^{\delta}\right)(t)dt\\ &\leq \left\|\frac{\mathrm{d}}{\mathrm{dt}}S_{T,d}^{\delta}\right\|_{\infty}|x-y|\\ &= \left\|\frac{d|\mathbb{S}^{d+1}|}{|\mathbb{S}^{d-1}|}C_{2T,d+2}^{\delta\ +}\right\|_{\infty}|x-y|\\ &\leq KT^{d+1}|x-y|.\end{aligned}$$

□

Let us present some useful inequalities.

Lemma 9.2.

$$h(n,d) \asymp n^{d-2}, \tag{9.12}$$

$$|\lambda(2p+1,d)| \asymp p^{-d/2}. \tag{9.13}$$

Proof. Estimate (9.12) is clearly satisfied when $d = 2$ and 3 since $h(n, 2) = 2$ and $h(n, 3) = 2n + 1$. When $d \geq 4$ we have

$$h(n, d) = \frac{2}{(d-2)!}(n + (d-2)/2)[(n+1)(n+2)\cdots(n+d-3)],$$

the lower bound is straightforward and the upper bound follows from

$$h(n, d) \leq \frac{2}{(d-2)!}(n + d - 3)^{d-2}$$

and $2/((d-2)!)$ by a constant large enough.

When d is even and $p \geq d/2$

$$|\lambda(2p+1, d)| = \frac{\kappa_d}{(2p+1)(2p+3)\cdots(2p+d-1)}$$

where

$$\kappa_d = \frac{|\mathbb{S}^{d-2}| 1 \cdot 3 \cdots (d-1)}{d-1}.$$

The upper bound is straightforward and we can write

$$|\lambda(2p+1, d)| \geq \frac{\kappa_d}{(2p+d-1)^{d/2}}$$

and conclude replacing κ_d by a small enough constant.

Sterling's double inequality, see Feller (1968) p.50-53

$$\sqrt{2\pi} n^{n+1/2} \exp\left(-n + \frac{1}{12n+1}\right) < n! < \sqrt{2\pi} n^{n+1/2} \exp\left(-n + \frac{1}{12n}\right)$$

implies that

$$\frac{(2^p p!)^2}{(2p)!} \asymp \sqrt{p}$$

thus

$$1 \cdot 3 \cdots (2p-1) \asymp \sqrt{p} 2 \cdot 4 \cdots (2p).$$

Therefore, for $p \geq d/2$ and d odd we have

$$|\lambda(2p+1, d)| \asymp \frac{\sqrt{p}}{(2p+2)(2p+4)\cdots(2p+d-1)}$$

and (9.13) holds for d even and odd. $\square$

Proof of Proposition 2.5. From the Funk-Hecke theorem we know that the coefficients $\alpha(n, d) = C_n^{\nu(d)}(1)|\mathbb{S}^{d-2}|^{-1}\lambda_n(\mathbb{I}\{t \in [0, 1]\})$ are given by

$$\alpha(n, d) = \int_0^1 C_n^{\nu(d)}(t)(1-t^2)^{(d-3)/2} dt$$

using (9.6),

$$\alpha(n,d)=\frac{(-2)^{-n}(d-2)_n}{n!\,((d-1)/2)_n}\int_0^1\frac{\mathrm{d}^n}{\mathrm{d}t^n}(1-t^2)^{n+(d-3)/2}dt.$$

Thus for $n\geq 1$ and $d\geq 3$,

$$\alpha(n,d)=-\frac{(-2)^{-n}(d-2)_n}{n!\,((d-1)/2)_n}\left.\frac{\mathrm{d}^{n-1}}{\mathrm{d}t^{n-1}}(1-t^2)^{n-1+(d-3)/2}dt\right|_{t=0}$$

since the term on the right hand-side is equal to 0 for $t=1$. To prove that the coefficients $\alpha(2p,d)$ are equal to zero for p positive it is enough to prove

$$\left.\frac{\mathrm{d}^{2p+1}}{\mathrm{d}t^{2p+1}}(1-t^2)^{2p+1+m}\right|_{t=0}=0,\quad \forall m\geq 1,\ p\geq 0.$$

The Faá di Bruno formula gives that this quantity is equal to

$$\sum_{k_1+2k_2=2p+1}\left.\frac{(-1)^{2p+1-k_2}(2p+1)!(m+1)\cdots(2p+1+m)}{k_1!k_2!}(1-t^2)^{m+k_2}(2t)^{k_1}\right|_{t=0}.$$

and we conclude since k_1 in the sum cannot be equal to 0.

When $n=2p+1$ for p non-negative we obtain, using again the Faá di Bruno formula, that the derivative at $t=0$ is equal to

$$(-1)^p\frac{(2p)!}{p!}\left[(2p+1+(d-3)/2)(2p+(d-3)/2)\cdots(p+2+(d-3)/2)\right].$$

We obtain the result of Proposition 2.5 using identity (9.8). For the case $d=2$ we use Proposition 2.1. □

Proof of Proposition 2.7. By definition we have

$$\|\mathcal{H}\left(f^-\right)\|^2_{2,s+d/2}=\sum_{p=0}^{\infty}(1+\zeta_{2p+1,d})^{s+d/2}\|Q_{2p+1,d}\mathcal{H}(f^-)\|_2^2$$

where according to the Funk-Hecke Theorem

$$\begin{aligned}Q_{2p+1,d}\mathcal{H}(f^-)&=Q_{2p+1,d}\mathcal{H}\left(\sum_{q=0}^{\infty}Q_{2q+1,d}f\right)\\&=Q_{2p+1,d}\left(\sum_{q=0}^{\infty}\lambda(2q+1,d)Q_{2q+1,d}f\right)\\&=\lambda(2p+1,d)Q_{2p+1,d}f.\end{aligned}$$

We conclude since Lemma 9.2 gives that $(1+\zeta_{2p+1,d})^{s+d/2}\lambda^2_{2p+1,d}\asymp(1+\zeta_{2p+1,d})^{s}$. □

Proof of Theorem 2.3. We apply Theorem 3.2. of Ditzian (1998) to $-P(D) = \mathcal{H}^{-2}$ choosing $\alpha = 1$ and $B = \mathrm{L}^q(\mathbb{S}^{d-1})$ and obtain that there exists a positive number $B(d,q)$ such that for all P in $\bigoplus_{p=0}^{T} H^{2p+1,d}$,

$$\begin{aligned}\|\mathcal{H}^{-2}P\|_q &\le B(d,q)\frac{1}{\lambda_{2T+1}^2}\|P\|_q\\ &\le CT^d\|P\|_q.\end{aligned}$$

The last inequality follows from (9.13). We deduce the result concerning $\mathcal{H}^{-1}$ using the Kolmogorov type inequality corresponding to Theorem 8.1 of Ditzian (1998). □

Proof of Theorem 4.1. R has the following condensed harmonic expansion

$$R(x) = \frac{1}{2} + \sum_{p=1}^{\infty}(Q_{2p+1,d}R)(x).$$

We then write using (3.2), changing variables and using (9.10),

$$\begin{aligned}(Q_{2p+1,d}R)(x) &= \int_{\mathbb{S}^{d-1}} q_{2p+1,d}(x,z)R(z)d\sigma(z)\\ &= \int_{H^+} q_{2p+1,d}(x,z)r(z)d\sigma(z) + \int_{-H^+} q_{2p+1,d}(x,z)(1-r(-z))d\sigma(z)\\ &= \int_{H^+} q_{2p+1,d}(x,z)r(z)d\sigma(z) - \int_{H^+} q_{2p+1,d}(x,z)(1-r(z))d\sigma(z)\end{aligned}$$

$$\begin{aligned}(Q_{2p+1,d}R)(x) &= \int_{H^+} q_{2p+1,d}(x,z)(2r(z)-1)d\sigma(z)\\ &= \int_{H^+} q_{2p+1,d}(x,z)\mathbb{E}\left[\left.\frac{2Y-1}{f_X(z)}\right| X=z\right] f_X(z)d\sigma(z)\\ &= \mathbb{E}\left[\frac{(2Y-1)q_{2p+1,d}(x,Z)}{f_X(Z)}\right].\end{aligned}$$

Lebesgue differentiation theorem along with (9.1) gives

$$\nabla^S_{x/\|x\|}R = \sum_{p=0}^{\infty}\mathbb{E}\left[\frac{(2Y-1)}{f_X(X)}\nabla^S_{x/\|x\|}{}^{\flat}q_{2p+1,d}(X'x/\|x\|)X\right]$$

the expression for the gradient of the radial extension of R follows then from Lemma 9.1. □

Proof of Theorems 4.2 and 4.3. The proofs for the estimation of R is the same as for f_β below. For the later we use one more tool being Theorem 2.3. The proof for derivatives is the same and we only need to replace σ by $\sigma-1$ and d by $d+2$. Note that $1/(2\sigma+d-1) = 1/(2(\sigma-1)+d+2-1)$. The multivariate CLT for derivatives is obtained using the Cramer-Wold device. □

Now we turn to the proofs of Theorems 5.1 and 5.2. For notational convenience we simply write $\hat{f}_\beta := \hat{f}_\beta^{N,T}$, $\hat{f}_\beta^- := \hat{f}_\beta^{-,N,T}$, $\mathbb{I} := \mathbb{I}\{f_\beta^-(b) > 0\}$ and $\hat{\mathbb{I}} := \mathbb{I}\{\hat{f}_\beta^-(b) > 0\}$. Then $f_\beta = 2f_\beta^-\mathbb{I}$ and $\hat{f}_\beta = 2\hat{f}_\beta^-\hat{\mathbb{I}}$. We denote by

$$\begin{aligned}\overline{f}_{\beta,T}^{\,-} &= \mathcal{H}^{-1}\overline{R}_T^{\,-}\\ \overline{f}_\beta^{\,-} &= \mathcal{H}^{-1}\overline{R}^{\,-}.\end{aligned}$$

where

$$\begin{aligned}\overline{R}_T^{\,-}(x) &= \frac{1}{N}\sum_{i=1}^N \frac{(2y_i-1)S_{T,d}^{d-1\;-}(x_i,x)}{\max\left(f_X(x_i),(\log N)^{-r}\right)}\\ \overline{R}^{\,-}(x) &= \frac{1}{N}\sum_{i=1}^N \frac{(2y_i-1)S_{T,d}^{d-1\;-}(x_i,x)}{f_X(x_i)}.\end{aligned}$$

Here we set $\delta = d-1$ since this is the order of the Cesàro summation which is sufficient for the estimation of f_β. If one is also interested in derivatives of f_β one should use higher order kernels but the proofs below work for any $\delta \geq d-1$.
We use several times the decomposition

$$\hat{f}_\beta^- - f_\beta^- = \left(\hat{f}_\beta^- - \overline{f}_{\beta,T}^{\,-}\right) - \left(\overline{f}_{\beta,T}^{\,-} - \mathbb{E}\left[\overline{f}_{\beta,T}^{\,-}\right]\right) - \left(\mathbb{E}\left[\overline{f}_{\beta,T}^{\,-}\right] - \mathbb{E}\left[\overline{f}_\beta^{\,-}\right]\right) - \left(\mathbb{E}\left[\overline{f}_\beta^{\,-}\right] - f_\beta^-\right),$$

and denote the terms on the right hand side by PI_β (plug-in), F_β (fluctuations), $B_{1,\beta}$ (trimming bias) and $B_{2,\beta}$ (approximation bias), where each term is $\mathcal{H}^{-1}$ of the corresponding term for R.

Proof of Theorem 5.1. Take $q \in [1,\infty)$,

$$\begin{aligned}\|\hat{f}_\beta - f_\beta\|_q^q &= \int (\hat{f}_\beta(b) - f_\beta(b))^q d\sigma(b)\\ &= \int_{I(b)=1,\hat{I}(b)=1} (\hat{f}_\beta(b) - f_\beta(b))^q d\sigma(b) + \int_{I(b)=0,\hat{I}(b)=1} (\hat{f}_\beta(b) - f_\beta(b))^q d\sigma(b)\\ &\quad + \int_{I(b)=1,\hat{I}(b)=0} (\hat{f}_\beta(b) - f_\beta(b))^q d\sigma(b) + \int_{I(b)=0,\hat{I}(b)=0} (\hat{f}_\beta(b) - f_\beta(b))^q d\sigma(b)\\ &:= A_1 + A_2 + A_3 + A_4.\end{aligned}$$

Obviously

$$A_1 = \int_{I(b)=1,\hat{I}(b)=1} (2\hat{f}_\beta^-(b) - 2f_\beta^-(b))^q d\sigma(b)$$

and $A_4 = 0$. Also,

$$A_2 = \int_{I(b)=0,\hat{I}(b)=1} (2\hat{f}_\beta^-(b) - f_\beta(b))^q d\sigma(b).$$

But given $I(b)=0$ and $\hat{I}(b)=1$, $2\hat{f}_\beta^-(b)>0$, $f_\beta(b)=0$ and $2f_\beta^-(b)\le 0$, so replacing f_β with $2f_\beta^-$ in the bracket,

$$A_2 \le \int_{I(b)=0,\hat{I}(b)=1} (2\hat{f}_\beta^-(b) - 2f_\beta^-(b))^q d\sigma(b).$$

Similarly,

$$A_3 = \int_{I(b)=1,\hat{I}(b)=0} (\hat{f}_\beta(b) - 2f_\beta^-(b))^q d\sigma(b).$$

and given $I(b)=1$ and $\hat{I}(b)=0$, $2f_\beta^-(b)>0$, $\hat{f}_\beta(b)=0$ and $2\hat{f}_\beta^-(b)\le 0$, so replacing f_β with $2f_\beta^-$ in the bracket,

$$A_3 \le \int_{I(b)=0,\hat{I}(b)=1} (2\hat{f}_\beta^-(b) - 2f_\beta^-(b))^q d\sigma(b).$$

Overall,

$$\|\hat{f}_\beta - f_\beta\|_q^q \le 4\|\hat{f}_\beta^- - f_\beta^-\|_q^q.$$

A similar proof could be carried out replacing $\mathrm{L}^q(\mathbb{S}^{d-1})$ by $\mathrm{L}^\infty(\mathbb{S}^{d-1})$. We can now focus on upper bounds for the estimation of f_β^- in $\mathrm{L}^2(\mathbb{S}^{d-1})$ or $\mathrm{L}^\infty(\mathbb{S}^{d-1})$.

We now denote by V_N the speed of convergence and T_N the smoothing parameter.

Let us start with PI_β. We have for $q\in[1,\infty]$

$$\begin{aligned}
\|PI_\beta\|_q &= \left\|\mathcal{H}^{-1}\left(\frac{1}{N}\sum_{i=1}^N \frac{(2y_i-1)S^{d-1\,-}_{2T_N+1}(x_i,\cdot)}{\max(f_X(x_i),(\log N)^{-r})}\left(\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right)\right)\right\|_q \\
&\le B(d,q)T_N^{d/2}\left\|\frac{1}{N}\sum_{i=1}^N \frac{(2y_i-1)S^{d-1\,-}_{2T_N+1}(x_i,\cdot)}{\max(f_X(x_i),(\log N)^{-r})}\left(\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right)\right\|_q \quad \text{(using Theorem 2.3)} \\
&\le CT_N^{d/2}\left(\left\|\frac{2}{N}\sum_{i=1}^N \frac{y_iS^{d-1\,-}_{2T_N+1}(x_i,\cdot)}{\max(f_X(x_i),(\log N)^{-r})}\left(\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right)\right\|_q \right. \quad \text{(by the triangular inequality)} \\
&\qquad \left. + \left\|\frac{1}{N}\sum_{i=1}^N \frac{S^{d-1\,-}_{2T_N+1}(x_i,\cdot)}{\max(f_X(x_i),(\log N)^{-r})}\left(\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right)\right\|_q\right)
\end{aligned}$$

$$\begin{aligned}
\|PI_\beta\|_q &\le CT_N^{d/2}\left(2\left\|\frac{1}{N}\sum_{i=1}^N \frac{y_iS^{d-1}_{2T_N+1}(x_i,\cdot)}{\max(f_X(x_i),(\log N)^{-r})}\left(\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right)\right\|_q \right. \quad \text{(by the triangular inequality)} \\
&\qquad \left. + \left\|\frac{1}{N}\sum_{i=1}^N \frac{S^{d-1}_{2T_N+1}(x_i,\cdot)}{\max(f_X(x_i),(\log N)^{-r})}\left(\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right)\right\|_q\right) \\
&\le CT_N^{d/2}(\log N)^r\left\|\frac{1}{N}\sum_{i=1}^N S^{d-1}_{2T_N+1}(x_i,\cdot)\right\|_q \max_{i=1,\dots,N}\left|\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right| \quad \text{(by positivness)}
\end{aligned}$$

and can bound the norm from above by

$$\left\|\frac{1}{N}\sum_{i=1}^{N} S_{2T_N+1}^{d-1}(x_i,\cdot)-\mathbb{E}\left[S_{2T_N+1}^{d-1}(X,\cdot)\right]\right\|_q+\left\|\mathbb{E}\left[S_{2T_N+1}^{d-1}(X,\cdot)\right]\right\|_q := \|T_1\|_q+\|T_2\|_q. \tag{9.14}$$

Let us start with the term $\|T_1\|_q$. We begin with the case where $q\in[1,2]$. Using the Hölder inequality we obtain that

$$\begin{aligned}\mathbb{E}\left[\|T_1\|_q^q\right] &= \int_{\mathbb{S}^{d-1}}\mathbb{E}\left[T_1(x)^q\right]d\sigma(x)\\ &\le \int_{\mathbb{S}^{d-1}}\mathbb{E}\left[T_1(x)^2\right]^{q/2}d\sigma(x)\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}\left[T_1(x)^2\right] &\le \frac{1}{N}\mathbb{E}\left[\left(S_{2T_N+1}^{d-1}(X,x)\right)^2\right]\\ &\le \frac{C}{N}\left\|S_{2T_N+1}^{d-1}(\star_2,\cdot)\right\|_2^2 \quad \text{(using that } f_X \text{ is bounded)}\\ &\le \frac{CT_N^{d-1}}{N} \quad \text{(using (2.15))}\end{aligned}$$

which implies

$$T_N^{d/2}(\log N)^r\|T_1\|_q = O_p\left((\log N)^r N^{-1/2}T_N^{(2d-1)/2}\right).$$

When $q\in(2,\infty]$, all $\mathrm{L}^q(\mathbb{S}^{d-1})$ norms can be interpolated between $\mathrm{L}^2(\mathbb{S}^{d-1})$ and $\mathrm{L}^\infty(\mathbb{S}^{d-1})$ norms using the Hölder inequality as follows

$$\forall f\in \mathrm{L}^\infty(\mathbb{S}^{d-1}),\ \|f\|_q = \|f\|_2^{2/q}\|f\|_\infty^{1-2/q}.$$

We can thus focus on the $\mathrm{L}^\infty(\mathbb{S}^{d-1})$ case. We cover the sphere by $\mathfrak{N}(N,r,d)$ geodesic balls (caps) $(B_i)_{i=1}^{\mathfrak{N}(N,r,d)}$ of centers $(\tilde{x}_i)_{i=1}^{\mathfrak{N}(N,r,d)}$ and radius $R(N,r,d)$, i.e. $\{x\in\mathbb{S}^{d-1}:\ \cos(\tilde{x}_i'x)\le R_N\}$. We know that $\mathfrak{N}(N,r,d)\asymp R(N,r,d)^{-(d-1)}$. For some speed $\tilde{V}_N$ and using Theorem 2.3 it is enough to show that for every ϵ positive, there exists a positive M such that

$$\mathbb{P}\left(\tilde{V}_N^{-1}B(d,\infty)T_N^{d/2}(\log N)^r\sup_{x\in\mathbb{S}^{d-1}}|T_1(x)|\ge M\right)\le\epsilon.$$

We write

$$\begin{aligned}
&\mathbb{P}\left(\tilde{V}_N^{-1}B(d,\infty)T_N^{d/2}(\log N)^r \sup_{x\in\mathbb{S}^{d-1}}|T_1(x)|\geq M\right)\\
&\leq \mathbb{P}\left(\bigcup_{i=1,\dots,\mathfrak{N}(N,r,d)}\left\{\tilde{V}_N^{-1}B(d,\infty)T_N^{d/2}(\log N)^r|T_1(\tilde{x}_i)|\geq M/2\right\}\right)\\
&\quad+\mathbb{P}\left(\exists i\in\{1,\dots,\mathfrak{N}(N,r,d)\}:\ \tilde{V}_N^{-1}B(d,\infty)T_N^{d/2}(\log N)^r\sup_{x\in B_i}|T_1(x)-T_1(\tilde{x}_i)|\geq M/2\right)\\
&\leq \mathfrak{N}(N,r,d)\sup_{i=1,\dots,\mathfrak{N}_N}\mathbb{P}\left(\tilde{V}_N^{-1}B(d,\infty)T_N^{d/2}(\log N)^r|T_1(\tilde{x}_i)|\geq M/2\right)
\end{aligned}$$

where the last inequality is obtained taking R_N small enough and such that $R_N \asymp (\log N)^{-r}\tilde{V}_N T_N^{-(3d/2+1)}M$ and using Proposition 2.3. For the remaining probabilities we write

$$\begin{aligned}
&\mathbb{P}\left(\tilde{V}_N^{-1}B(d,\infty)T_N^{d/2}(\log N)^r|T_1(\tilde{x}_i)|\geq M/2\right)\\
&=\mathbb{P}\left(\left|\sum_{j=1}^N\frac{S^{d-1}_{2T_N+1}(x_j,\tilde{x}_i)}{T_N^{d-1}}-\mathbb{E}\left[\frac{S^{d-1}_{2T_N+1}(X,\tilde{x}_i)}{T_N^{d-1}}\right]\right|\geq T_N^{-(d-1)}\tilde{V}_N B(d,\infty)^{-1}T_N^{-d/2}(\log N)^{-r}NM/2\right)\\
&\leq 2\exp\left\{-\frac{1}{2}\left(\frac{t^2}{v+Lt/3}\right)\right\}\quad\text{(Bernstein inequality)}
\end{aligned}$$

where

$$\begin{aligned}
&t=T_N^{-(d-1)}\tilde{V}_N B(d,\infty)^{-1}T_N^{-d/2}(\log N)^{-r}NM/2\\
&v\geq\sum_{j=1}^N\operatorname{var}\left(\frac{S^{d-1}_{T,d}(X_j,\tilde{x}_i)}{T_N^{d-1}}\right)\\
&\forall j=1,\dots,N,\ \left|\frac{S^{d-1}_{T,d}(X_j,\tilde{x}_i)}{T_N^{d-1}}\right|\leq L\quad\text{(for some constant L using(2.15)).}
\end{aligned}$$

We can take $v=CN\|S^{d-1}_{T,d}(\star_2,\cdot)\|_2^2\|f_X\|_\infty T_N^{-2(d-1)}$, i.e. from (2.15) $v=CN$. v is the leading term in the denominator. Thus we have for positive constants C and C_2 and N large enough

$$\begin{aligned}
&\mathbb{P}\left(\tilde{V}_N^{-1}B(d,\infty)T_N^{d/2}(\log N)^r\sup_{x\in\mathbb{S}^{d-1}}|T_1(x)|\geq M\right)\\
&\leq C\exp\left\{-(d-1)\log\left((\log N)^{-r}\tilde{V}_N T_N^{-(3d/2+1)}\right)-(d-1)(\log M)-C_2\tilde{V}_N^2T_N^{-(2d-1)}(\log N)^{-2r}NM^2\right\}\\
&\leq C\exp\left\{-(d-1)\log\left((\log N)^{-r}\tilde{V}_N T_N^{-(3d/2+1)}\right)-C_2\tilde{V}_N^2T_N^{-(2d-1)}(\log N)^{-2r}NM^2\right\}
\end{aligned}$$

and if we take $\tilde{V}_N = (\log N)^r V_N = (\log N)^{r+1/2} N^{-1/2} T_N^{(2d-1)/2}$ we obtain for some positive constants C_1 and C_2

$$\mathbb{P}\left(\tilde{V}_N^{-1} B(d,\infty) T_N^{d/2} (\log N)^r \sup_{x\in\mathbb{S}^{d-1}} |T_1(x)| \geq M\right) \leq C \exp\left\{(\log N)(C_1 - C_2 M^2)\right\} \tag{9.15}$$

for M large enough this could be made as small as we wish, thus

$$B(d,\infty) T_N^{d/2} (\log N)^r \|T_1\|_\infty = O_p\left((\log N)^{r+1/2} N^{-1/2} T_N^{(2d-1)/2}\right)$$
$$B(d,\infty) T_N^{d/2} (\log N)^r \|T_1\|_q = O_p\left((\log N)^{r+1/2-1/q} N^{-1/2} T_N^{(2d-1)/2}\right).$$

Concerning $\|T_2\|_q$, since f_X is bounded there exists a positive C such that the second term in the right hand side of (9.14) is less than

$$C \left\|\left\|S_{2T_N+1}^{d-1}(\star_1, \star_q)\right\|_1\right\|_q$$

where integration in $\|\cdot\|_1$ is with respect to argument $\star_1$ and integration in $\|\cdot\|_q$ is with respect to $\star_q$. $\left\|S_{2T_N+1}^{d-1}(\star_1, \star_q)\right\|_1$ is a constant and does not depend on $\star_q$. This is because we integrate over the whole sphere and $S_{2T_N+1}^{d-1}(\star_1, \star_q)$ is indeed a function of $\star_1' \star_q$. Thus

$$\left\|\left\|S_{2T_N+1}^{d-1}(\star_1, \star_q)\right\|_1\right\|_q = |\mathbb{S}^{d-1}|^{1/q} \left\|S_{2T_N+1}^{d-1}(\star_1, \star_q)\right\|_1$$

and we can use the fact that for $d-1 > (d-2)/2$ the Cesàro kernels are uniformly integrable to conclude that this term is $O(1)$ thus

$$T_N^{d/2} (\log N)^r \|T_2\|_q = O\left((\log N)^r T_N^{d/2}\right).$$

For the choice made later for T_N this term is of higher order than the first term.
Analogously to our treatment of $\|T_2\|_q$, we can prove that when $q \in [1,2]$,

$$\|F_\beta\|_q = O_p\left((\log N)^r N^{-1/2} T_N^{(2d-1)/2}\right),$$

while for $q \in (2,\infty]$

$$\|F_\beta\|_q = O_p\left((\log N)^{r+1/2-1/q} N^{-1/2} T_N^{(2d-1)/2}\right).$$

Let us now study the bias term induced by trimming

$$\begin{aligned} B_{1,\beta}(b) &= \mathbb{E}\left[\frac{(2Y-1)\mathcal{H}^{-1}\left(S_{2T_N+1}^{d-1\,-}(X,\cdot)\right)(b)}{f_X(X)}\left(\frac{f_X(X)}{\max(f_X(X),(\log N)^{-r})} - 1\right)\right] \\ &= \int_{\{z\in\mathbb{S}^{d-1}:\ 0<f_X(z)<(\log N)^{-r}\}} \mathbb{E}[(2Y-1)|X=z]\mathcal{H}^{-1}\left(S_{2T_N+1}^{d-1\,-}(X,\cdot)\right)(b)\left(f_X(z)(\log N)^r - 1\right) d\sigma(z). \end{aligned}$$

Using Proposition 2.2, (2.15) along with Theorem 2.3 we have

$$\|B_{1,\beta}\|_q \le T_N^{d/2+(d-1)(1-1/q)}\sigma(0 < f_X < (\log N)^{-r}).$$

Under the assumptions of the theorem this term is negligible compared to the variance term.
We finally treat $B_{2,\beta}$ using Proposition 2.4 with the assumption that $f_\beta^- \in \mathrm{W}_q^s(\mathbb{S}^{d-1})$,

$$\|B_{2,\beta}\|_q \le C T_N^{-s}.$$

We now need to choose V_N and T_N such that

$$V_N^{-1}(\log N)^r T_N^{d/2} \max_{i=1,\dots,N}\left|\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right| = O_p(1) \tag{9.16}$$

$$V_N^{-1}(\log N)^{r+(1/2-1/q)\mathbb{I}\{q\ge 2\}}N^{-1/2}T_N^{(2d-1)/2} = O(1) \tag{9.17}$$

$$V_N^{-1}T_N^{3d/2-1-(d-1)/q}\sigma(0 < f_X < (\log N)^{-r}) = O(1) \tag{9.18}$$

$$V_N^{-1}T_N^{-s} = O(1) \tag{9.19}$$

and look for solutions of the form

$$V_N = \left(\frac{N}{(\log N)^{2(r+(1/2-1/q)\mathbb{I}\{q\ge 2\})}}\right)^{-\alpha}, \qquad T_N = \left(\frac{N}{(\log N)^{2(r+(1/2-1/q)\mathbb{I}\{q\ge 2\})}}\right)^{\gamma}$$

where α and γ are non-negative. The optimal upper bound on V_N is obtained by setting

$$2\alpha + \gamma(2d-1) = 1 \text{ (from (9.17))} \tag{9.20}$$

$$\alpha = \gamma s \text{ (from (9.19))} \tag{9.21}$$

indeed the left hand side of (9.17) is

$$N^{\alpha-1/2+\gamma(2d-1)/2}(\log N)^{-(\alpha+\gamma(2d-1)/2-1)2(r+(1/2-1/q)\mathbb{I}\{q\ge 2\})}$$

which is equal to 1. Condition (5.1) and Assumption 4.1 have been taken so that (9.16) and (9.18) are then satisfied as well.

In order to prove the strong uniform consistency, noticing that the bias terms are not stochastic are bounded after proper scaling, we just have to focus on the fluctuations and plug-in. Concerning the plug-in note that taking M large enough so that $C_1 - C_2M^2 < -1$ implies summability of the left hand side in (9.15) and we conclude from the first Borel-Cantelli lemma that the probability that the events occur infinitely often is zero thus with probability one

$$\overline{\lim}_{N\to\infty}\tilde{V}_N^{-1}B(d,\infty)T_N^{d/2}(\log N)^r \sup_{x\in\mathbb{S}^{d-1}}|T_1(x)| < M.$$

For the term involving T_2 we use the same non stochastic upper bound. We then use Assumption 4.1 (ii) instead of Assumption 4.1 (i) to show that almost sure uniform boundedness of the plug-in term after proper rescaling. The treatment of the fluctuation term is analogous to that of T_1. □

Proof of Theorem 5.2. We first prove that the Lyapounov condition holds: there exists $\delta > 0$ such that for N going to infinity,

$$\frac{\mathbb{E}\left[\left|Z_{N,1}-\mathbb{E}\left[Z_{N,1}\right]\right|^{2+\delta}\right]}{N^{\delta/2}\left(\mathrm{var}\left(Z_{N,1}\right)\right)^{1+\delta/2}} \to 0 \tag{9.22}$$

(see, e.g. Billingsley, 1995) and impose assumptions so that the plug-in and bias terms properly rescaled are $o_p(1)$.

We need a lower bound on $\mathrm{var}\left(Z_{N,1}\right)$. Since $\mathbb{E}[Z_{N,1}]$ converges to $f_{\beta}^{-}(x)$ while the variance blows-up, it is enough to obtain a lower bound on

$$\begin{aligned}
&\mathbb{E}[Z_{N,1}^2](b)\\
&=\sum_{p=0}^{T_N}\left(\frac{2A_{2(T_N-p)}^{d-1}}{A_{2T_N+1}^{d-1}}\right)^2\int_{H^+}\left(\frac{q_{2p+1,d}(z,b)}{\max\left(f_X(z),(\log N)^{-r}\right)\lambda(2p+1,d)}\right)^2 f_X(z)d\sigma(z)\\
&=\sum_{p=0}^{T_N}\left(\frac{2A_{2(T_N-p)}^{d-1}}{A_{2T_N+1}^{d-1}}\right)^2\int_{H^+}\left(\frac{q_{2p+1,d}(z,b)}{\lambda(2p+1,d)}\right)^2\left(\frac{1}{f_X(z)}\mathbb{I}\{f_X\geq(\log N)^{-r}\}+f_X(z)(\log N)^{2r}\mathbb{I}\{f_X<(\log N)^{-r}\}\right)d\sigma(z)\\
&\geq\frac{1}{\|f_X\|_\infty}\sum_{p=0}^{T_N}\left(\frac{2A_{2(T_N-p)}^{d-1}}{A_{2T_N+1}^{d-1}}\right)^2\left(\int_{H^+}\frac{q_{2p+1,d}(z,b)^2}{\lambda(2p+1,d)^2}d\sigma(z)-\int_{\{0<f_X<(\log N)^{-r}\}}\frac{q_{2p+1,d}(z,b)^2}{\lambda(2p+1,d)^2}d\sigma(z)\right)d\sigma(z)
\end{aligned}$$

where, using Proposition 2.2 and (9.7)

$$\begin{aligned}
\int_{\{0<f_X<(\log N)^{-r}\}}\frac{q_{2p+1,d}(z,b)^2}{\lambda(2p+1,d)^2}d\sigma(z)&\leq C\left(\frac{h(2p+1,d)}{\lambda(2p+1,d)}\right)^2\sigma\left(0<f_X<(\log N)^{-r}\right)\\
&\leq Cp^{3d-4}\sigma\left(0<f_X<(\log N)^{-r}\right)
\end{aligned}$$

thus

$$\mathbb{E}[Z_{N,1}^2](b)\geq\frac{1}{\|f_X\|_\infty}\sum_{p=0}^{\lfloor T_N/2\rfloor}\left(\frac{2A_{2(T_N-p)}^{d-1}}{A_{2T_N+1}^{d-1}}\right)^2\int_{H^+}\frac{q_{2p+1,d}(z,b)^2}{\lambda(2p+1,d)^2}d\sigma(z)d\sigma(z)-CT_N^{3(d-1)}\sigma\left(0<f_X<(\log N)^{-r}\right).$$

The first term on the right hand side can be bounded from below by

$$\frac{C}{2^{2(d-1)}}\sum_{p=0}^{\lfloor T_N/2\rfloor}\left\|\frac{q_{2p+1,d}(z,b)}{\lambda(2p+1,d)}\right\|_2^2$$

i.e. by CT_N^{2d-1}. Thus as $\sigma\left(0<f_X<(\log N)^{-r}\right)$ decays fast enough to zero, here it is enough that $\sigma\left(0<f_X<(\log N)^{-r}\right)=O\left(T_N^{-d+2}\right)$,

$$\mathbb{E}[Z_{N,1}^2](b)\geq CT_N^{2d-1}$$

and the denominator of (9.22) is greater than $CN^{\delta/2}N^{\alpha(2d-1)(1+\delta/2)}$.

We now obtain an upper bound of $\mathbb{E}\left[|Z_{N,1}|^{2+\delta}\right]$ using Theorem 2.3 and (2.15)

$$\begin{aligned}\mathbb{E}\left[|Z_{N,1}|^{2+\delta}\right]&\leq\|f_X\|_\infty(\log N)^{r(2+\delta)}\left\|\mathcal{H}^{-1}\left(S_{2T_N+1}^{d-1\,-}(z,\cdot)\right)\right\|_{2+\delta}^{2+\delta}\\&\leq\|f_X\|_\infty(\log N)^{r(2+\delta)}B(d,2+\delta)^{2+\delta}T_N^{d(2+\delta)/2}\left\|S_{2T_N+1}^{d-1\,-}(z,\cdot)\right\|_{2+\delta}^{2+\delta}\\&\leq C(\log N)^{r(2+\delta)}T_N^{d(2+\delta)/2}T_N^{(d-1)(1+\delta)}.\end{aligned}$$

An upper bound for the ratio appearing in (9.22) is given by

$$(\log N)^{r(2+\delta)}\left(\frac{T_N^{d-1}}{N}\right)^{\delta/2}$$

as a consequence the Lyapounov condition is satisfied as soon as (5.5) holds. We now need to prove that the remaining terms multiplied by $N^{1/2}s_N^{-1}$ are $o_p(1)$.

The plug in is treated in a similar manner as in the proof of Theorem 5.1.

$$|PI_\beta(b)|\leq2\left(\frac{1}{N}\sum_{i=1}^N\frac{\left|\mathcal{H}^{-1}\left(S_{2T_N+1}^{d-1\,-}(x_i,\cdot)\right)(b)\right|}{\max(f_X(x_i),(\log N)^{-r})}\right)\max_{i=1,\dots,N}\left|\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right|.$$

Using the Markov inequality the term in parenthesis is an O_p of

$$(\log N)^r\left\|\mathcal{H}^{-1}\left(S_{2T_N+1}^{d-1\,-}(\star_1,\cdot)\right)\right\|_1$$

and from Theorem 2.3 of

$$B(d,1)T_N^{d/2}(\log N)^r\left\|S_{2T_N+1}^{d-1\,-}(\star_1,\cdot)\right\|_1$$

where using the definition of the odd part is an O_p of

$$B(d,1)T_N^{d/2}(\log N)^r\left\|S_{2T_N+1}^{d-1}(\star_1,\cdot)\right\|_1$$

where the last quantity does not depend on $\cdot$ and is uniformly bounded. We obtained

$$N^{1/2}B(d,1)T_N^{-(d-1/2)}|PI_\beta(b)|\leq\left(N^{1/2}T_N^{-(d-1)/2}(\log N)^r\right)\max_{i=1,\dots,N}\left|\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right|$$

thus $N^{1/2}B(d,1)T_N^{-(d-1/2)}|PI_\beta(b)| = o_p(1)$ when

$$\max_{i=1,\dots,N}\left|\frac{\max\left(f_X(x_i),(\log N)^{-r}\right)}{\max\left(\hat{f}_X^N(x_i),(\log N)^{-r}\right)}-1\right| = o_p\left(N^{-1/2}T_N^{(d-1)/2}(\log N)^{-r}\right)$$

and when condition (5.5) it is enough to assume (5.4). Let us now consider the bias term induced by the trimming procedure.

In the proof of Theorem 5.1 we have obtained an upper bound for $\|B_{1,\beta}\|_\infty$ and we deduce that

$$N^{1/2}T_N^{-(d-1/2)}\|B_{1,\beta}\|_\infty = o(1)$$

when condition (5.6) is satisfied.

Finally, $N^{1/2}T_N^{-(d-1/2)}\|B_{1,\beta}\|_\infty$ is an $o(1)$ as soon as condition (5.7) is satisfied.

We conclude using that the asymptotic normality holds for b such that $f_\beta(b) > 0$ and the factor 4 in the variance comes from the fact that $\hat{f}_\beta = 2\hat{f}_\beta^-\hat{\mathbb{I}}$. □

References

[1] Athey, S., and Imbens, G. W. (2007): "Discrete Choice Models with Multiple Unobserved Choice Characteristics". Preprint.

[2] Bajari, P., Fox, J. and Ryan, S. (2007): "Linear Regression Estimation of Discrete choice Models with Nonparametric Distribution of Random Coefficients". *American Economic Review, Papers and Proceedings*, 97, 459-463.

[3] Beran, R. and Hall, P. (1992): "Estimating Coefficient Distributions in Random Coefficient Regression". *Annals of Statistics*, 20, 1970-1984.

[4] Berry, S., and Haile, P. (2008): "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers". Preprint.

[5] Billingsley, P. (1995): *Probability and Measure - Third Edition*. New York: Wiley.

[6] Bonami, A., and Clerc, J. L. (1973): "Sommes de Cesàro et Multiplicateurs des Développements en Harmoniques Sphériques", *Transactions of the American Mathematical Society*, 183, 223-263.

[7] Briech, R., Chintagunta, P. and Matzkin, R. (1996): "Nonparametric Discrete Choice Models with Unobserved Heterogeneity". Preprint.

[8] Brownstone, D., and Train, K. (1999): "Forecasting New Product Penetration with Flexible Substitution Patterns". *Journal of Econometrics*, 89, 109-129.

[9] Carrasco, M., Florens, J. P., and Renault, E. (2007): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization", *Handbook of Econometrics,* J. J. Heckman and E. E. Leamer (eds.), vol. 6B, North Holland, chapter 77, 5633-5751.

[10] Chesher, A., and Silva Santos, J. M. C.. (2002): "Taste Variation in Discrete Choice Models". *Review of Economic Studies*, 69, 147-168.

[11] Colzani, L., and Traveglini, G. (1991): "Estimates for Riesz Kernels of Eigenfunction Expansions of Elliptic Differential Operators on Compact Manifolds". *Journal of Functional Analysis*, 96, 1-30.

[12] Devroye, L., and Gyorfi, L. (1985): *Nonparametric Density Estimation. The L1-View*. Wiley, New York.

[13] do Carmo, M. P. (1976): *Differential Geometry of Curves and Surfaces*. Prentice-Hall, Englewood Cliffs.

[14] Ditzian, Z. (1998): "Fractional Derivatives and Best Approximation". *Acta Mathematica Hungarica*, 81, 323-348.

[15] Elbers, C. and Ridder, G. (1982): "True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Models". *Review of Economics Studies*, 49, 403-410.

[16] Erdélyi, A. et al. ed. (1953): *Higher Transcendental Functions*, vol. 1,2 of the *Bateman Manuscript Project*. Mc Graw-Hill: New-York.

[17] Evans, L. C. (1998): *Partial Differential Equations*. American Mathematical Society: Providence.

[18] Fan, J. (1991): "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems". *Annals of Statistics*, 19, 1257-1272.

[19] Feller, W. (1968): *An Introduction to Probability Theory and Its Applications - Third edition*, vol. 1. Wiley: New York.

[20] Fisher, N. I., Lewis, T., and Embleton, B. J. J. (1987): *Statistical Analysis of Spherical Data*. Cambridge University Press: Cambridge.

[21] Funk, P. (1916): "Uber Eine Geometrische Anwendung der Abelschen Integralgleichung". *Mathematische Annalen*, 77, 129-135.

[22] Gallot, S., Hulin, D., and Lafontaine, J. (2004): *Riemannian Geometry - Third edition.* Springer, Berlin.

[23] Groemer, H. (1996): *Geometric Applications of Fourier Series and Spherical Harmonics.* Cambridge University Press: Cambridge, *Encyclopedia of Mathematics and its Applications.*

[24] Groeneboom, P., and Jongbloed, G. (2003): "Density Estimation in the Uniform Deconvolution Model". *Statistica Neerlandica*, 57, 136-157.

[25] Gronwall, T. H. (1914): "On the Degree of Convergence of Laplace's Series". *Transactions of the American Mathematical Association*, 15, 1-30.

[26] Hall, P., Watson, G. S., and Cabrera, J. (1987): "Kernel Density Estimation with Spherical Data". *Biometrika*, 74, 751-62.

[27] Harding, M. C. and Hausman, J. (2006): "Using a Laplace Approximation to Estimate the Random Coefficients Logit Model by Non-linear Least Squares". Preprint.

[28] Heckman, J. J. and Singer, B. (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data". *Econometrica*, 52, 271-320.

[29] Hendriks, H. (1990): "Nonparametric Estimation of a Probability Density on a Riemannian Manifold Using Fourier Expansions". *Annals of Statistics*, 18, 832-849.

[30] Hess, S., Bolduc, D. and Polak, J. (2005): "Random Covariance Heterogeneity in Discrete Choice Models". Preprint.

[31] Healy, D. M., and Kim, P. T. (1996): "An Empirical Bayes Approach to Directional Data and Efficient Computation on the Sphere". *Annals of Statistics*, 24, 232-254.

[32] Hoderlein, S., Klemelä, J., and Mammen, E. (2007): "Reconsidering the Random Coefficient Model. *Preprint.*

[33] Ichimura, H., and Thompson, T. S. (1998): "Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution". *Journal of Econometrics*, 86, 269-295.

[34] Johnstone, I. M., and Raimondo, M. (2004): "Periodic Boxcar Deconvolution and Diophantine Approximation". *Annals of Statistics*, 32, 1781-1804.

[35] Kamzolov, A. I. (1983): "The Best Approximation of the Class of Functions $\mathbf{W}_p^{\alpha}(\mathbb{S}^n)$ by Polynomials in Spherical Harmonics". *Matematical Notes* 32 (1982), 285293.

[36] Katznelson, Y. (2004): *An Introduction to Harmonic Analysis - Third Edition.* Cambridge University Press: Cambridge.

[37] Kim, P. T., and Koo, J. Y. (2000): "Directional Mixture Models and Optimal Esimation of the Mixing Density". *The Canadian Journal of Statistics*, 28, 383-398.

[38] Klemelä, J. (2000): "Estimation of Densities and Derivatives with Directional Data". *Journal of Multivariate Analysis*, 73, 18-40.

[39] Kogbetliantz, E. (1924): "Recherches sur la Sommabilité des Séries Ultrasphériques par la Méthode des Moyennes Arithmétiques". *Journal de Mathématiques Pures et Appliquées*, 89, 107-187.

[40] Kushpel, A. K., Levesley, J., and Tas, K. (1997): "$\epsilon-$Entropy of Sobolev's Classes on $\mathbb{S}^d$ ". *Journal of Mathematics and Computer Science*, 4, 1-13.

[41] Müller, C. (1966): *Spherical Harmonics.* Lecture Notes in Mathematics, 17, Springer.

[42] Ragozin, D. (1972): "Uniform Convergence of Spherical Harmonic Expansions". *Mathematische Annalen*, 195, 87-94.

[43] Rubin, B. (1999): "Inversion and Characterization of the Hemispherical Transform". *Journal d'Analyse Mathématique*, 77, 105-128.

[44] Ruymgaart, F. H. (1989): "Strong Uniform Convergence of Density Estimators on Spheres". *Journal of Statistical Planning and Inference*, 23, 45-52.

[45] Train, K. E. (2003): *Discrete Choice Methods with Simulation.* Cambridge University Press: Cambridge.

ENSAE-CREST, 3 avenue Pierre Larousse, 92245 Malakoff Cedex, France.

E-mail address: gautier@ensae.fr

Cowles Foundation for Research in Economics, Yale University, New Haven, CT-06520.

E-mail address: yuichi.kitamura@yale.edu