# NONPARAMETRIC ESTIMATION IN
# RANDOM COEFFICIENTS BINARY CHOICE MODELS

**By**

**Eric Gautier and Yuichi Kitamura**

**August 2009**

**COWLES FOUNDATION DISCUSSION PAPER NO. 1721**

**COWLES FOUNDATION FOR RESEARCH IN ECONOMICS**
**YALE UNIVERSITY**
**Box 208281**
**New Haven, Connecticut 06520-8281**

**http://cowles.econ.yale.edu/**

# NONPARAMETRIC ESTIMATION IN RANDOM COEFFICIENTS BINARY CHOICE MODELS

ERIC GAUTIER AND YUICHI KITAMURA

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 on spheres. This approach offers a clear insight on the identification problem. More importantly, it leads to a closed form estimator formula that yields a simple plug-in procedure requiring 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, therefore the vector $X$ is of the form $X = (1, \tilde{X}')'$. The vector $\beta$ is random. The random element $(Y, \tilde{X}, \beta)$ 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. The vectors $\tilde{X}$ and $\beta$ correspond to observed and unobserved heterogeneity across agents, respectively. Note that the first element of $\beta$ in this formulation absorbs the usual scalar stochastic

*Date*: This Version: June 26, 2009.
*Keywords*: Inverse problems, Discrete Choice Models.

We thank seminar participants at Chicago, CREST, Harvard/MIT, the Henri Poincaré Institute, Hitotsubashi, LSE, Mannheim, Minnesota, Northwestern, NYU, Princeton, Rochester, Simon Fraser, Tilburg University, UBC, UCL, UCLA, UCSD, the Tinbergen Institute and the University of Tokyo, and participants of the 2008 Cowles summer econometrics conference, EEA/ESEM, FEMES, Journées STAR and SETA for helpful comments. Yuhan Fang and Xiaoxia Xi provided excellent research assistance. Kitamura acknowledges financial support from the National Science Foundation via grants SES-0241770 and SES-0551271. Gautier is grateful for support from the Cowles Foundation as this research was initiated during his visit as a postdoctoral associate.

shock term as well as a constant in a standard binary choice model with non-random coefficients. This formulation is used in Ichimura and Thompson (1998), and is convenient for the subsequent development in this paper. Our basic model maintains exogeneity of the covariates $\tilde{X}$:

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

Section 6.3 considers ways to relax this assumption. Under (1.1) and Assumption 1.1, the choice probability function is given by

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

Discrete choice models with random coefficients are useful in applied research since it is often crucial to incorporate unobserved heterogeneity in modeling 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 impose parametric distributional assumptions. 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 $\beta$ itself is the structural parameter of interest. The goal for the econometrician is then to recover it nonparametrically from the information about $r(x)$ obtained from the data.

Nonparametric treatments for unobserved heterogeneity distributions have been considered in the literature for other models. 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 these models is relatively underdeveloped. In their important paper, Ichimura and Thompson (1998) propose an NPMLE for the CDF of $\beta$. 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 a random coefficients multinomial choice model that often arises in empirical IO.

This paper considers nonparametric estimation of the random coefficients distribution, using a novel approach that shares some similarities with standard deconvolution techniques. This allows us to reconsider the identifiability of the model and obtain a constructive identification result. Moreover, we develop a simple plug-in estimator for the density of $\beta$ that requires no numerical optimization or integration. It is easy to implement in empirical applications, while being flexible about the treatment of unobserved heterogeneity.

Since the scale of $\beta$ is not identified in the binary choice model, we normalize it so that $\beta$ is a vector of Euclidean norm 1 in $\mathbb{R}^d$. The vector $\beta$ then belongs to the $d-1$ dimensional sphere $\mathbb{S}^{d-1}$. This is not a restriction as long as the probability that $\beta$ is equal to 0 is 0. Also, since only the angle between $X$ and $\beta$ matters in the binary decision $\mathbb{I}\{X'\beta \geq 0\}$, we can replace $X$ by $X/\|X\|$ without any loss of information. We therefore assume that $X$ is on the sphere $\mathbb{S}^{d-1}$ as well in the subsequent analysis. Results from the directional data literature are thus relevant to our analysis. We aim to recover the joint probability density function $f_\beta$ of $\beta$ with respect to the uniform spherical measure $\sigma$ over $\mathbb{S}^{d-1}$ from the random sample $(y_1, x_1), \ldots, (y_N, x_N)$ 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 discusses some of its properties. $\mathcal{H}$ is not injective without further restrictions and conditions need to be imposed to ensure identification of $f_\beta$ from $r$. 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 $f_\beta$ to be in $\mathrm{L}^2(\mathbb{S}^{d-1})$. Since the kernel of $\mathcal{H}$ is square integrable by compactness of the sphere, it 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 its kernel that involves the scalar product $x'b$, the operator $\mathcal{H}$ is an analogue of convolution in $\mathbb{R}^d$, as illustrated in a simple example in Section 2. This analogy provides a clear insight into the identification issue. In particular, our problem is closely related to the so-called boxcar deconvolution (see, e.g. Groeneboom and Jongbloed (2003) and Johnstone and Raimondo (2004)), where identifiablity is often a significant problem. The connection with deconvolution is also useful in deriving an estimator based on a series expansion on the Fourier basis on $\mathbb{S}^1$ 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 kernel of the integral operator $\mathcal{H}$, however, does not satisfy the assumptions made by Healy and Kim. Unlike Healy and Kim (1996), we make use of so-called "condensed" harmonic expansions. 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 enables 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 identification and presents a general estimation procedure for the random coefficients density. In Section 4 we study a nonparametric estimator of the choice probability function and derive its asymptotic properties. This is important since it yields a nonparametric estimator for the random coefficients density with a simple closed form, which is the main proposal of the paper. We derive the convergence rates of the estimators in all the $\mathrm{L}^q$ spaces for $q \in [1, \infty]$ and also prove a pointwise CLT in Section 5. Some extensions, such as estimation of marginals, treatments of models with non-random coefficients, and the case with endogenous regressors are presented in Section 6. Simulation results are reported in Section 7. Section 8 concludes.

## 2. Preliminaries

This section introduces tools for making connections between the estimation of the density of $\beta$ and a deconvolution problem, and presents some results on the hemispherical transform.

2.1. **A Toy Model.** As noted above, the key insight for our estimation procedure lies in the fact the estimation of $f_\beta$ in (1.3) is mathematically equivalent to a statistical deconvolution problem. To see this, it is useful to first consider the case with $d = 2$. We parameterize the vectors $b = (b_1, b_2)'$ and $x = (x_1, x_2)'$ on $\mathbb{S}^1$ by their angles $\phi = \arccos(b_1)$ and $\theta = \arccos(x_1)$ in $[0, 2\pi)$. As is often the case when Fourier series techniques are used, we consider spaces of complex valued functions. Let $\mathrm{L}^p(\mathbb{S}^1)$ denote the Banach space 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 $\int_0^{2\pi} f(\theta)\overline{g(\theta)}d\theta$. Let $r_\theta$ and $f_\phi$ denote $r$ and $f_\beta$ after the reparameterization. Our task is then to obtain $f_\phi$ from the knowledge of $r_\theta$. Rewrite (1.3) using these definitions, then divide both sides by $\pi$, to get:

$$\frac{r_\theta}{\pi}(\theta) = \frac{\mathcal{H}(f_\beta)}{\pi}(x) = \int_0^{2\pi} \left(\frac{1}{\pi}\mathbb{I}\left\{|\theta - \phi| < \pi/2\right\}\right) f_\phi(\phi)d\phi. \tag{2.1}$$

If we further define $f_\theta := r_\theta/\pi$ and $f_\eta(\eta) := \frac{1}{\pi}\mathbb{I}\{|\eta| < \pi/2\}$, then using the standard notation for convolution, (2.1) can be written as $f_\theta = f_\eta * f_\phi$. It is now obvious that the estimation of $f_\phi$ (thus $f_\beta$) is linked to the following statistical deconvolution problem: unobservable random variables $\phi$ and $\eta$ with densities $f_\phi$ and $f_\eta$ are related to an observable random variable $\theta$ according to $\theta = \eta + \phi$, and one wishes to recover $f_\phi$ from $f_\theta$, the density of $\theta$, when $f_\eta$ is known (and it is Uniform$[-\pi/2, \pi/2]$ in this case).[1]

The problem of deconvolution on the unit circle can be conveniently solved using Fourier series. The set of functions $\left(\exp(-int)/\sqrt{2\pi}\right)_{n\in\mathbb{Z}}$ is the orthonormal basis of $\mathrm{L}^2(\mathbb{S}^1)$ used to define Fourier series. This system is also complete in $\mathrm{L}^1(\mathbb{S}^1)$. Reparameterize a function $f \in \mathrm{L}^1(\mathbb{S}^1)$ it using angles as above, and denote it by $f_t$. Denoting the Fourier coefficients of $f \in \mathrm{L}^1(\mathbb{S}^1)$ by $c_n(f_t) = \int_0^{2\pi} f_t(t)\exp(-int)dt/(2\pi)$,

$$f_t(\theta) = \sum_{n\in\mathbb{Z}} c_n(f_t)\exp(int) \tag{2.2}$$

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

$$c_n(f_t * g_t) = 2\pi c_n(f_t)c_n(g_t). \tag{2.3}$$

Using equation (2.3) we obtain the following proposition.

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

[1]It is also useful to note that the inversion of $\mathcal{H}$ is closely related to differentiation. Differentiating the right hand-side of expression (2.1) with respect to $\theta$ identifies $f_\phi(\theta + \pi/2) - f_\phi(\theta - \pi/2)$ where $f_\phi$ is defined on the line by periodicity. If $f_\phi$ is supported on a semicircle, with an assumption that is elaborated further in Section 3.1, $f_\phi$ (which is positive) is identified. Thus if the model is identified the inverse of $\mathcal{H}$ is a differential operator and as such unbounded.

As in classical deconvolution problems on the real line, our aim is to obtain $f_t$ (thus $f_\beta$) using equation (2.2) and Proposition 2.1. Proposition 2.1 shows that $c_{2p}(r_\theta) = 0$ holds for all non-zero $p$'s, regardless of the values of $c_{2p}(f_\phi), p \in \mathbb{Z}\backslash\{0\}$. Thus from $r(x) = r_\theta(\theta)$ one can only recover the Fourier coefficients $c_n(f_\phi)$ for $n = 0$ (which is easily seen to be $1/2\pi$, by integrating both sides of (2.1) and noting that $f_\beta$ is a probability density function) and $n = 2p+1, p \in \mathbb{Z}$. The same phenomenon occurs in higher dimensions, as explained in Section 2.2.

**Remark 2.1.** The vector spaces $H^{2p+1,2} = \text{span}\left\{\exp(i(2p+1)t)/\sqrt{2\pi}, \exp(-i(2p+1)t)/\sqrt{2\pi}\right\}, p \in \mathbb{N}$ are eigenspaces of the compact self-adjoint operator $\mathcal{H}$ on $\mathrm{L}^2(\mathbb{S}^1)$. These eigenspaces are associated with the eigenvalues $\frac{2(-1)^p}{2p+1}$. Also, $\bigoplus_{p\in\mathbb{N}\backslash\{0\}} H^{2p,2}$ is the null space $\ker \mathcal{H}$.

2.2. **Tools for Higher Dimensional Spheres.** Let us introduce some concepts used for the treatment of the general case $d \geq 2$. We consider functions defined on the sphere $\mathbb{S}^{d-1}$, which is a $d-1$ dimensional smooth submanifold in $\mathbb{R}^d$. The canonical measure on $\mathbb{S}^{d-1}$ (or the spherical measure) is denoted by $\sigma$. It is a uniform measure on $\mathbb{S}^{d-1}$ satisfying $\int_{\mathbb{S}^{d-1}} d\sigma = |\mathbb{S}^{d-1}|$, where $|\mathbb{S}^{d-1}|$ signifies the surface area of the unit sphere. The latter is given by $\left|\mathbb{S}^{d-1}\right| = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ where $\Gamma$ is the usual Gamma function. $\mathrm{L}^p(\mathbb{S}^{d-1})$ with norm $\|\cdot\|_p$ is the usual space of $p$-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 positive $n$.

Recall that the basis functions $\exp(\pm int)/\sqrt{2\pi}$ are eigenfunctions of $-\frac{\mathrm{d}}{\mathrm{d}t^2}$ associated with eigenvalue $n^2$. In a similar way, the Laplacian on the sphere $\mathbb{S}^{d-1}, d \geq 2$, denoted by $\Delta^S$, can be used to obtain an orthonormal basis for higher dimensional spheres. It can be defined by the formula

$$\Delta^S f = (\Delta \breve{f})\check{} \tag{2.4}$$

where $\Delta$ is the Laplacian in $\mathbb{R}^d$, $\breve{f}$ the radial extension of $f$, that is $\breve{f}(x) = f(x/\|x\|)$, and $\check{f}$ the restriction of $f$ to $\mathbb{S}^{d-1}$. Likewise the gradient on the sphere is given by:

$$\nabla^S f = (\nabla \breve{f})\check{} \tag{2.5}$$

where $\nabla$ is the gradient in $\mathbb{R}^d$.

**Definition 2.1.** A surface harmonic of degree $n$ is the restriction of a homogeneous harmonic polynomial (a homogeneous polynomial $p$ whose Laplacian $\Delta p$ is zero) of degree $n$ in $\mathbb{R}^d$ to $\mathbb{S}^{d-1}$.

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. Here are some useful results:

**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 degree* $n$ *are eigenfunctions of* $-\Delta^S$ *for the eigenvalue* $\zeta_{n,d} := n(n+d-2)$;

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

$$h(n,d) := \frac{(2n+d-2)(n+d-2)!}{n!(d-2)!(n+d-2)}; \tag{2.6}$$

*(iv) A system formed of orthonormal bases* $(Y_{n,l})_{l=1}^{h(n,d)}$ *of* $H^{n,d}$ *for each degree* $n = 0, \ldots, \infty$ *is complete in* $\mathrm{L}^1(\mathbb{S}^{d-1})$, *that is, for every* $f \in \mathrm{L}^1(\mathbb{S}^{d-1})$ *the following equality holds in the* $\mathrm{L}^1(\mathbb{S}^{d-1})$ *sense:*

$$f = \sum_{n=0}^{\infty} \sum_{l=1}^{h(n,d)} (f, Y_{n,l})_{\mathrm{L}^2(\mathbb{S}^{d-1})} Y_{n,l}.$$

Thus $h(n,d)$ is the multiplicity of the eigenvalue $\zeta_{n,d}$, and $H^{n,d}$ is the corresponding eigenspace. Lemma 2.1 (i), (ii) and (iv) give the decomposition

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

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 $d > 3$. Instead, here we work with the decomposition of a function on the spaces $H^{n,d}$ as presented in the next definition so that we avoid specific expressions of basis functions.

**Definition 2.2.** The condensed harmonic expansion of a function $f$ in $\mathrm{L}^1(\mathbb{S}^{d-1})$ is the series $\sum_{n=0}^{\infty} Q_{n,d} f$, where $Q_{n,d}$ is the projector from $\mathrm{L}^2(\mathbb{S}^{d-1})$ to $H^{n,d}$.

This leads to a simple method both in terms of theoretical developments and practical implementations. The projector $Q_{n,d}$ can be expressed as an integral operator with kernel

$$q_{n,d}(x,y)=\sum_{l=1}^{h(n,d)}\overline{Y_{n,l}(x)}Y_{n,l}(y), \tag{2.7}$$

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)**.** *For every $x$ and $y\in\mathbb{S}^{d-1}$, we have*

$$q_{n,d}(x,y)={}^{\flat}q_{n,d}(x'y),\quad {}^{\flat}q_{n,d}(t):=\frac{h(n,d)C_n^{\nu(d)}(t)}{|\mathbb{S}^{d-1}|C_n^{\nu(d)}(1)} \tag{2.8}$$

*where $C_n^{\nu}$ are Gegenbauer polynomials and $\nu(d)=(d-2)/2$.*

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 $[-1,1]$. 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, the following recursion relation holds

$$(n+2)C_{n+2}^{\nu}(t)=2(\nu+n+1)tC_{n+1}^{\nu}(t)-(2\nu+n)C_n^{\nu}(t). \tag{2.9}$$

Implementation of our estimator requires evaluation of the Gegenbauer polynomials for a series of successive values of $n$. The recursion relation (2.9) 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.** The Sobolev space $\mathrm{W}_p^s(\mathbb{S}^{d-1})$ for $p\in[0,\infty]$ and $s\geq 0$ is the space of functions $f$ in $\mathrm{L}^p(\mathbb{S}^{d-1})$ for which the distribution

$$\left(-\Delta^S\right)^{s/2}f:=\sum_{n=0}^{\infty}\zeta_{n,d}^{s/2}Q_{n,d}f$$

belongs to $\mathrm{L}^p(\mathbb{S}^{d-1})$. It is equipped with the norm

$$\|f\|_{p,s}=\|f\|_p+\left\|\left(-\Delta^S\right)^{s/2}f\right\|_p.$$

For the case where $p=2$, 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

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

Note that the following integration by parts holds for functions $f$ in $\mathrm{H}^1(\mathbb{S}^{d-1})$

$$-\int_{\mathbb{S}^{d-1}} f(x)\Delta^S f(x)d\sigma(x) = \int_{\mathbb{S}^{d-1}} \nabla_x^S f' \nabla_x^S f d\sigma(x) \tag{2.10}$$

and as a consequence for the second definition of the norm of $\mathrm{H}^1(\mathbb{S}^{d-1})$ we have

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

We use these Sobolev spaces to make smoothness assumptions.

In Section 2.1 we observed the close relationship between the random coefficient binary choice model and convolution for $d = 2$. This connection remains valid in higher dimensions. Suppose a function $f(x, y)$ defined on $\mathbb{S}^{d-1} \otimes \mathbb{S}^{d-1}$ depends on $x$ and $y$ only through the spherical distance $d(x, y) = \arccos(x'y)$ (that is, $f$ is a zonal function). Consider the following integral:

$$h(x) = \int_{\mathbb{S}^{d-1}} f(x, y)g(y)d\sigma(y) := f * g(x),$$

then the function $h$ is a convolution on the sphere. We now see that the choice probability function $r(x) = \mathcal{H}(f_\beta)(x) = \int_{\mathbb{S}^{d-1}} \mathbb{I}\{x'b \geq 0\} f_\beta(b)d\sigma(b)$ is a special case of $h$ and therefore can also be regarded as convolution. Obtaining $f_\beta$ from $r$ (or, inverting $\mathcal{H}$) is therefore a deconvolution problem.

In what follows we often write $f(x, \star)$ when a function $f$ on $\mathbb{S}^{d-1} \otimes \mathbb{S}^{d-1}$ is regarded as a function of $\star$. Also, the notation $\|f(x, \star)\|_p$ is used for the $\mathrm{L}^p$ norm of $f(x, \star)$, that is, $\|f(x, \star)\|_p = \int_{\mathbb{S}^{d-1}} |f(x, y)|^p d\sigma(y)$. Note that if $f$ is a zonal function as in the above definition of spherical convolution, its $\mathrm{L}^p$ norm $\|f(x, \star)\|_p$ does not depend on $x$. The following Young inequalities for convolution on the sphere (see, for example, Kamzolov, 1983) are useful:

**Proposition 2.2** (Young inequalities)**.** *Suppose $f(x, \star)$ and $g$ belong to $\mathrm{L}^r(\mathbb{S}^{d-1})$ and $\mathrm{L}^p(\mathbb{S}^{d-1})$, respectively. Then $h(x) = f * g(x)$ is well-defined in $\mathrm{L}^q(\mathbb{S}^{d-1})$ and*

$$\|h\|_q \leq \|f\|_r \|g\|_p,$$

*where $1 \leq p, q, r \leq \infty$ and $\frac{1}{q} = \frac{1}{p} + \frac{1}{r} - 1$.*

Let $P_T$ denote the projection operator onto $\bigoplus_{n=0}^T H^{n,d}$, i.e.

$$P_T f(x) = \sum_{n=0}^{T} Q_{n,d} f(x) = \int_{\mathbb{S}^{d-1}} D_T(x, y)f(y)d\sigma(y)$$

where

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

The kernel $D_T$ extends the classical Dirichlet kernel on the circle to the sphere $\mathbb{S}^{d-1}$. The sum over $T$ in the definition of $D_T$ also has the simple closed form in terms of derivatives of Gegenbauer polynomials; see Equation (52) in Müller (1966). The linear form $f \to \int_{\mathbb{S}^{d-1}} D_T(x,y)f(y)d\sigma(y)$ converges to $\int_{\mathbb{S}^{d-1}} f(y)d\delta_x(y) = f(x)$ as $T$ goes to infinity, where $\delta_x$ denotes the Dirac measure. The Dirichlet kernel yields the best approximation $P_T f$ of $f$ in $\mathrm{L}^2(\mathbb{S}^{d-1})$ by polynomials that belong to $\bigoplus_{n=0}^T H^{n,d}$, but is known to have flaws. For example, $D_T$ does not satisfy

$$\forall f \in \mathrm{L}^1(\mathbb{S}^{d-1}), \lim_{T\to\infty} \|D_T * f - f\|_{\mathrm{L}^1(\mathbb{S}^{d-1})} = 0,$$

that is, the sequence $D_T, T = 0, 1, ...$ is not an approximate identity (see, e.g., Devroye and Gyorfi 1985) in $\mathrm{L}^1(\mathbb{S}^{d-1})$. Indeed, the $\mathrm{L}^1(\mathbb{S}^{d-1})$ norm of the kernel is not uniformly bounded; more precisely, we have

$$\|D_T(\cdot, x)\|_1 \asymp T^{(d-2)/2} \tag{2.11}$$

when $d \geq 3$ and

$$\|D_T(\cdot, x)\|_1 \asymp \log T \tag{2.12}$$

when $d = 2$ (as noted above, these norms do not depend on the value of $x \in \mathbb{S}^{d-1}$). These bounds can be found in Gronwall (1914) for $d = 3$ and Ragozin (1972) and Colzani and Traveglini (1991) for higher dimensions. Also, $D_T$ does not have good approximation properties in $\mathrm{L}^\infty(\mathbb{S}^{d-1})$; in particular, we do not have

$$\forall f \in \mathrm{L}^\infty(\mathbb{S}^{d-1}), \lim_{T\to\infty} \|D_T * f - f\|_{\mathrm{L}^\infty(\mathbb{S}^{d-1})} = 0.$$

Near the points of discontinuity of $f$, $D_T * f$ has oscillations which do not decay to zero as $T$ grows to infinity, known as the Gibbs oscillations. This phenomenon deteriorates as the dimension increases. These problems can be addressed by using kernels that involves extra smoothing instead of the Dirichlet kernel $D_T$. To this end, define a general class of kernel

$$K_T(x,y) = \sum_{n=0}^{T} \chi(n,T) q_{n,d}(x,y) \tag{2.13}$$

for some sequence $\chi(n,T)$. These are called smoothed projection kernels. Typically the function $\chi$ is chosen so that it puts more weight on lower frequencies. In particular we impose the following conditions:

**Assumption 2.1.** *(i)* $\|K_T(x,\star)\|_1$ *is uniformly bounded in* $T$.

(*ii*) *There exists constants* $C$ *and* $\alpha$ *such that for all* $x, y, z \in \mathbb{S}^{d-1}$,

$$|K_T(z,x) - K_T(z,y)| \leq C\|x-y\|T^{\alpha},$$

*where* $\|\cdot\|$ *denotes the Euclidean norm.*

(*iii*) *For* $p \in [1,\infty]$ *and* $s > 0$, *there exists a constant* $C$ *such that for every* $f$ *in* $\mathrm{W}_p^s(\mathbb{S}^{d-1})$,

$$\left\| f(\cdot) - \int_{\mathbb{S}^{d-1}} K_T(\cdot,y)f(y)d\sigma(y)\right\|_p \leq CT^{-s}\|f\|_{p,s}.$$

(*iv*) $\chi(\cdot,T)$ *takes values in* $[0,1]$ *and is such that there exists* $c > 0$ *such that for all* $0 \leq n \leq \lfloor T/2 \rfloor$, $\chi(n,T) \geq c$.

The smoothed projection kernel $K_T(x,y)$ depends on $x$ and $y$ only through $d(x,y)$, thus the value of the norm $\|K_T(x,\star)\|_1$ in Assumption (i) does not depend on $x \in \mathbb{S}^{d-1}$. Assumption (i) could be relaxed, but imposing this on $K_T$ allows us to make relatively weak assumptions on the smoothness of the density of the covariates later in this paper. Assumption (ii) is used to establish the $\mathrm{L}^{\infty}$-rates of convergence of our estimators. Assumption (iii) provides bounds for approximation errors. Under this condition, $K_T * f$ approximates $f \in \mathrm{L}^p(\mathbb{S}^{d-1})$ with an error of the same order as that of the best $n$-th degree spherical harmonic approximation of a function $f \in \mathrm{L}^p(\mathbb{S}^{d-1})$ in $\mathrm{W}_p^s(\mathbb{S}^{d-1})$ (see e.g. Kamzolov 1983 and Ditzian 1998). This is useful in our treatment of the bias terms in our estimators. As concrete examples, the following two choices for the weight function $\chi$ in (2.13) satisfy Assumption 2.1, as shown in the appendix. The first and the second choices of $\chi$ correspond to the *Riesz kernel* and the *delayed means kernel*, respectively.

**Proposition 2.3.** *In the definition of the smoothed kernel* (2.13), *let*

$$\chi(n,T) = \left(1 - \left(\frac{\zeta_{n,d}}{\zeta_{T,d}+1}\right)^{s/2}\right)^l,$$

*where* $l$ *is an integer satisfying* $l > (d-2)/2$, *or*

$$\chi(n,T) = \psi(n/T), \quad T = 2^j \text{ for some } j$$

*where* $\psi: [0,\infty) \to [0,\infty)$ *is infinitely differentiable, nonincreasing, such that* $\psi(x) = 1$ *if* $x \leq 1/2$ *and* $\psi(x) = 0$ *if* $x \geq 1$. *Then* $K_T$ *satisfies Assumption 2.1.*

The delayed means kernel has the nice property that it does not require prior knowledge of the regularity $s$ in Assumption 2.1. The Dirichlet kernel satisfies (ii), (iii) (for $p = 2$) and (iv) of Assumption

2.1. Like the delayed means kernel, it achieves the optimal rate of approximation without the prior knowledge of $s$.

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

**Definition 2.4.** We denote the odd part and the even part of a function $f$ by

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

and

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

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

If the function $f$ is in $\mathrm{L}^2(\mathbb{S}^{d-1})$ then Equations (2.8) and (9.8) imply that $Q_{2p,d}f(x) = Q_{2p,d}f(-x)$ and $Q_{2p+1,d}f(x) = -Q_{2p+1,d}f(-x)$ for $p \in \mathbb{N}$. Consequently, the odd order terms in the condensed harmonic expansions of $f$, $f^{+}$ and $f^{-}$ satisfy $Q_{2p+1}f^{-} = Q_{2p+1}f$ and $Q_{2p+1}f^{+} = 0$. Likewise, for the even order terms in the condensed harmonic expansions of these functions $Q_{2p}f^{+} = Q_{2p}f$ and $Q_{2p}f^{-} = 0$ hold. We conclude that the sum of the odd order terms in the condensed harmonic expansion corresponds to $f^{-}$ and that of the even order terms to $f^{+}$. As anticipated from the analysis of the $d = 2$ case, the operator $\mathcal{H}$ reduces the even part of $f_\beta$ to a constant $\frac{1}{2}$, therefore Fourier-Laplace series expansions for $f_\beta$ derived later involve only odd order terms.

We now provide a formula that is later used to obtain our estimator for $f_\beta$. If a non-negative function $f$ has its support included in some hemisphere of $\mathbb{S}^{d-1}$ then

$$f(x) = 2f^{-}(x)\mathbb{I}\left\{f^{-}(x) > 0\right\}. \tag{2.14}$$

Denote the support of $f$ by $\mathrm{supp}f$ and let $-\mathrm{supp}f = \{x| -x \in \mathrm{supp}f\}$, then this formula follows from the fact that $f^{-}(x) = f^{+}(x) \geq 0$ on $\mathrm{supp}f$ while $f^{-}(x) = -f^{+}(x) \leq 0$ on $-\mathrm{supp}f$ and both $f^{-}$ and $f^{+}$ are 0 on $\mathbb{S}^{d-1} \setminus (\mathrm{supp}f \bigcup -\mathrm{supp}f)$.

**Remark 2.2.** If $f$ is a probability density function, the coefficient of degree 0 in the expansion of $f$ on surface harmonics is $1/|\mathbb{S}^{d-1}|$. Conversely, any harmonic polynomial or series such that its degree 0 coefficient is $1/|\mathbb{S}^{d-1}|$ integrates to one.

The next theorem shows that Fourier-Laplace series on the sphere is a natural tool for the study of the operator $\mathcal{H}$.

**Theorem 2.2** (Funk-Hecke Theorem)**.** *If $g$ belongs to $H^{n,d}$ for some $n$, and a function $F$ on $(-1,1)$ satisfies*

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

*then*

$$\int_{\mathbb{S}^{d-1}} F(x'y)g(y)d\sigma(y) = \lambda_n(F)g(x) \tag{2.15}$$

*where*

$$\lambda_n(F) = |\mathbb{S}^{d-2}|C_n^{\nu(d)}(1)^{-1}\int_{-1}^{1} F(t)C_n^{\nu(d)}(t)(1-t^2)^{\frac{d-3}{2}}dt.$$

In other words, the kernel operator 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, in the subspace $H^{n,d}$, equivalent to the multiplication by $\lambda_n(F)$. Thus a basis of surface harmonics diagonalizes an integral operator if its kernel is a function of the scalar product $x'y$.

**Remark 2.3.** Healy and Kim (1996) use Fourier-Laplace expansions to analyze a deconvolution problem on $\mathbb{S}^2$. As we shall see below, the Addition Formula along with condensed harmonic expansions provide a general treatment that works for arbitrary dimensions.

2.3. **The Hemispherical Transform.** The hemispherical transform $\mathcal{H}$, defined by $\mathcal{H}f(x) = \int_{\mathbb{S}^{d-1}} \mathbb{I}\{x'y \geq 0\}f(y)d\sigma(y)$, plays a central role in our analysis. It is a special case of the operator considered in the Funk-Hecke theorem above, with $F(t) = \mathbb{I}\{t \in [0,1]\}$, therefore the next proposition follows.

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

**Proposition 2.4.** *When $d \geq 2$, the coefficients $\lambda(n,d)$ have the following expressions*

(*i*) $\lambda(0,d) = \frac{2}{|\mathbb{S}^{d-1}|}$

(*ii*) $\lambda(1,d) = \frac{|\mathbb{S}^{d-2}|}{d-1}$

(*iii*) $\forall p \in \mathbb{N},\ \lambda(2p,d) = 0$

(*iv*) $\forall p \in \mathbb{N},\ \lambda(2p+1,d) = \frac{(-1)^p|\mathbb{S}^{d-2}|1\cdot3\cdots(2p-1)}{(d-1)(d+1)\cdots(d+2p-1)}$.

For the sake of completeness we give a simple proof of this result in the appendix (see also Groemer (1996) and Rubin (1999)). 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 and similarly $\mathrm{L}^2_{\mathrm{even}}(\mathbb{S}^{d-1})$ and $\mathrm{H}^s_{\mathrm{even}}(\mathbb{S}^{d-1})$ for even functions. The following corollary is a direct consequence of the Funk-Hecke Theorem and Proposition 2.4, and corresponds to an observation made in Remark 2.1 for the $d = 2$ case.

**Corollary 2.1.** *The null space of the hemispherical transform $\mathcal{H}$ is given by*

$$\ker \mathcal{H} = \bigoplus_{p=1}^{\infty} H^{2p,d} = \left\{ f \in \mathrm{L}^2_{\text{even}}(\mathbb{S}^{d-1}) : \int_{\mathbb{S}^{d-1}} f(x) d\sigma(x) = 0 \right\},$$

*when $\mathcal{H}$ is viewed as an operator on $\mathrm{L}^2(\mathbb{S}^{d-1})$. The spaces $H^{0,d}$ and $H^{2p+1,d}$ for $p \in \mathbb{N}$ are the eigenspaces associated with the non-zero eigenvalues of $\mathcal{H}$.*

As a consequence of Proposition 2.4, $\mathcal{H}$ is not injective and restrictions have to be imposed in order to ensure identification of $f_\beta$. Section 3 presents sufficient conditions that allows us to reconstruct $f_\beta$ from $f_\beta^-$.

The following proposition can be found in Rubin (1999).

**Proposition 2.5.** *$\mathcal{H}$ is a bijection from $\mathrm{L}^2_{\text{odd}}(\mathbb{S}^{d-1})$ to $\mathrm{H}^{d/2}_{\text{odd}}(\mathbb{S}^{d-1})$.*

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

**Proposition 2.6.** *For all $s > 0$, there exists positive constants $C_l$ and $C_u$ such that for all $f$ in $\mathrm{H}^s(\mathbb{S}^{d-1})$*

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

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

$$\mathcal{H}^{-1}(f^-)(y) = \sum_{p=0}^{\infty} \frac{1}{\lambda(2p+1,d)} \int_{\mathbb{S}^{d-1}} q_{2p+1,d}(x,y) f^-(x) d\sigma(x). \tag{2.16}$$

This is straightforward given our results at hand: for example, operate $\mathcal{H}$ on the RHS to see:

$$\begin{aligned} \mathcal{H}\left( \sum_{p=0}^{\infty} \frac{1}{\lambda_{2p+1}} \int_{\mathbb{S}^{d-1}} q_{2p+1,d}(x,y) f^-(x) d\sigma(x) \right) &= \sum_{p=0}^{\infty} \frac{1}{\lambda_{2p+1}} \mathcal{H} Q_{2p+1} f^- \\ &= \sum_{p=0}^{\infty} \frac{\lambda_{2p+1}}{\lambda_{2p+1}} Q_{2p+1} f^- \quad \text{(by the Funk-Hecke Theorem)} \\ &= f^-. \end{aligned}$$

If $f^-$ belongs to $\mathrm{H}^{d/2}(\mathbb{S}^{d-1})$, then $\mathcal{H}^{-1}(f^-)(b)$ is a well-defined $\mathrm{L}^2(\mathbb{S}^{d-1})$ function. Otherwise it should be understood as a distribution and is only defined in a Sobolev space with negative exponent. Moreover, if $d$ is a multiple of 4, it is possible to relate the inverse of the operator $\mathcal{H}$ with differentiation as in the case of $d = 2$:

**Proposition 2.7.** *If $d$ is a multiple of 4,*

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

See the appendix for the proof. This connection between the inverse of $\mathcal{H}$ and differentiation suggests that a Bernstein-type inequality might hold for $\mathcal{H}^{-1}$. Indeed, even though the above inversion formula is concerned with $d$'s that are multiples of 4, the following Bernstein inequality holds for every dimension.

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

$$\|\mathcal{H}^{-1}P\|_q \leq B(d,q)T^{d/2}\|P\|_q. \tag{2.17}$$

This result is proved in the appendix. It is important for our subsequent analysis of the estimation of the random coefficients density.

Rubin (1999) gives other inversion formulas for the Hemispherical transform in terms of differential operators. The fact that the inversion roughly corresponds to differentiation is another manifestation of the ill-posedness of our problem at hand. The inverse operator $\mathcal{H}^{-1}$ is indeed unbounded. We call the factor $d/2$ in (2.17) the degree of ill-posedness of the inverse problem. For the case $q = 2$, there exists a lower bound for $\|\mathcal{H}^{-1}P\|_q$ in (2.17) of order $T^{d/2}$ as well, implying that the upper bound $T^{d/2}$ in the order of $T$ obtained in Theorem 2.3 is tight.

## 3. General Results

3.1. **Identification in the Random Coefficient Model.** In this section we address the following two questions:

- (Q1) Under what conditions is $f_\beta$ identified?
- (Q2) Does the random coefficients model impose restrictions?

Let us start with the question (Q1). As noted in Section 2.3, operating $\mathcal{H}$ reduces the even part of a function to a constant 1 and therefore it is impossible to recover $f_\beta^+$ from the knowledge of $r$, which is what observations offer. Our identification strategy is therefore as follows: (Step 1) Assume conditions that guarantee the identification of $f_\beta^-$; then (Step 2) Show that $f_\beta$ is uniquely determined from $f_\beta^-$ under a reasonable assumption. We first consider Step 1. Define $H^+ = H(\mathbf{n}) = \{x \in \mathbb{S}^{d-1} : x'\mathbf{n} \geq 0\}$, where $\mathbf{n} = (1,,0,...,0)'$, that is, the northern hemisphere of $\mathbb{S}^{d-1}$. For later use, also define its southern hemisphere $H^- = H(-\mathbf{n})$. Since the model we consider has a constant as the first element of the covariate vector before normalization, the same vector after normalization is necessarily an element of $H^+$. We make the following assumption, which also appears in Ichimura and Thompson (1998), and show that it achieves Step 1.

**Assumption 3.1.** *The support of $X$ is $H^+$.*

This assumption demands that $\tilde{X}$, the vector of non-constant covariates in the original scale, is supported on the whole space $\mathbb{R}^{d-1}$. It rules out discrete or bounded covariates; see Section 6 for a potential approach to deal with regressors with limited support. In what follows we assume that the law of $X$ is absolutely continuous with respect to $\sigma$ and denote its density by $f_X$.

Step 1 of our identification argument is to show that the knowledge of $r(x)$ on $H^+$, which is available under Assumption 3.1, identifies $f_\beta^-$. The problem at hand calls for solving $r = \mathcal{H}f_\beta = \frac{1}{2} + \mathcal{H}f_\beta^-$ for $f_\beta^-$, and the inversion formula derived in (2.16) is potentially useful for the purpose. A direct application of the formula to $r$ is inappropriate, however, since it requires integration of $r$ on the whole sphere $\mathbb{S}^{d-1}$, but $r$ is defined only on $H^+$ even when $\tilde{X}$ has full support on $\mathbb{R}^{d-1}$. An appropriate extension of $r(x), x \in H^+$ to the entire $\mathbb{S}^{d-1}$ is in order. Using the random coefficients model (1.1) and Assumption 1.1, then noting that $f_\beta$ is a probability density function, conclude

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

for $x$ in $H^+$. This suggests an extension $R$ of $r$ to $\mathbb{S}^{d-1}$ as follows:

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

The function $R$ is well-defined on the whole sphere under Assumption 3.1. Later we derive a formula for $f_\beta^-$ in terms of $R(x), x \in \mathbb{S}^{d-1}$, which shows the identifiablity of $f_\beta^-$ under Assumption 3.1.

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 completely determined by its odd part and therefore,

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

or

$$R^- = \mathcal{H} f_\beta^-. \tag{3.4}$$

We can invert this equation to obtain $f_\beta^-$.

Now we turn to Step 2 in our identification argument. Obviously $f_\beta^-$ does not uniquely determine $f_\beta$ without further assumptions. This is a fundamental identification problem in our model. We need to identify $f_\beta$ from the choice probability function $r$, but we can choose an appropriate even function $g$ so that $f_\beta + g$ is a legitimate density function (see the proof of Proposition 3.1 for such a construction). Then $r = \mathcal{H}(f_\beta + g)$, and the knowledge of $r$ identifies $f_\beta$ only up to such a function $g$. Ichimura and Thompson (1998, Theorem 1) give a set of conditions that imply the identification of the model (1.1). One of their 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 $\beta$ 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 $\beta$ 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 $f_\beta^-$ to $f_\beta$ developed in (2.14):

$$f_\beta(b) = 2f_\beta^-(b)\mathbb{I}\left\{f_\beta^-(b) > 0\right\}. \tag{3.5}$$

This is useful because it shows that Assumption 3.2 guarantees identification if $f_\beta^-$ is identified. Moreover, it will be used in the next section to develop a key formula that leads to a simple and practical estimator for $f_\beta$ that is guaranteed to be non-negative.

**Remark 3.1.** Assumption 3.2 is testable since it imposes restrictions on $f_\beta^-$, which is identified under weak conditions. For example, for values of $b$ with $f_\beta^-(b) > 0$, $f_\beta^-(-b) < 0$ must hold. Or, it implies that $f_\beta^-$ integrates to $1/(2|\mathbb{S}^{d-1}|)$ on a hemisphere $H(x)$ for some $x$, and $-1/(2|\mathbb{S}^{d-1}|)$ on the other $H(-x)$.

The following proposition answers question (Q2), and a proof is given in the appendix.

**Proposition 3.1.** *A* $[0,1]$*-valued function* $r$ *is compatible with the random coefficient model* (1.1) *with* $f_\beta$ *in* $\mathrm{L}^2(\mathbb{S}^{d-1})$ *and Assumption 1.1 if and only if* $r$ *is homogeneous of degree 0 and its extension* $R$ *according to* (3.2) *belongs to* $\mathrm{H}^{d/2}(\mathbb{S}^{d-1})$.

The global smoothness assumption that $R$ belongs to $\mathrm{H}^{d/2}(\mathbb{S}^{d-1})$ imposes substantial restriction on the property of observables, that is, the behavior of the choice probability function $r$. Note that the smoothness condition in this proposition is stated in terms of $R$, and even if the choice probability function $r$ is sufficiently smooth on the support of $X$, which is $H^+$, it is not necessarily consistent with the random coefficient binary choice model (1.1) unless its extension is smooth globally on $\mathbb{S}^{d-1}$. In particular, the Sobolev embedding of $\mathrm{H}^s(\mathbb{S}^{d-1})$ into the space of continuous functions for $s > (d-1)/2$ implies that if the extension $R$ is in $\mathrm{H}^{d/2}(\mathbb{S}^{d-1})$, it has to be continuous on $\mathbb{S}^{d-1}$. This, in turn, means that the corresponding $r$ has to satisfy certain matching conditions at a boundary point $x$ of $H^+$ (i.e. $x'\mathbf{n} = 0$) and its opposite point $-x$.

3.2. **Nonparametric Estimation of $f_\beta$.** If an appropriate estimator $\tilde{R}^-$ of $R^-$ is available, an application of the inversion formula (2.16) to (3.4) suggests the following estimator for $f_\beta^-$:

$$\tilde{f}_\beta^- = \mathcal{H}^{-1}\left(\tilde{R}^-\right) = \sum_{p=0}^{\infty} \frac{1}{\lambda(2p+1,d)} \int_{\mathbb{S}^{d-1}} q_{2p+1,d}(\cdot,x)\tilde{R}^-(x)d\sigma(x). \tag{3.6}$$

Then use the mapping (3.5) to define

$$\tilde{f}_\beta(b) = 2\tilde{f}_\beta^-(b)\mathbb{I}\left\{\tilde{f}_\beta^-(b) > 0\right\} \tag{3.7}$$

as an estimator for $f_\beta$. Proposition 2.6 implies that if $\tilde{f}_\beta^- - f_\beta^- \in \mathrm{H}^s(\mathbb{S}^{d-1})$ then $\tilde{R}^- - R^- \in \mathrm{H}^\sigma(\mathbb{S}^{d-1})$, $\sigma = s + \frac{d}{2}$ and for $v \in [0,s]$,

$$\|\tilde{f}_\beta^- - f_\beta^-\|_{2,v} \asymp \|\tilde{R}^- - R^-\|_{2,v+d/2}. \tag{3.8}$$

As discussed earlier, the estimation of $f_\beta$ is related to deconvolution in $\mathbb{S}^{d-1}$, and the degree of ill-posedness in our model is $d/2$, which is indeed the rate at which the eigenvalues $\lambda(n,d), n = 2p+1, p \in \mathbb{N}$ converges to zero as $n$ grows, as shown in (9.11). Existing results for deconvolution problems (see, for example, Fan, 1991 and Kim and Koo, 2000) then suggest that we should be able to estimate $f_\beta$ at the rate $N^{\frac{s}{2s+2d-1}}$ in the $\mathrm{L}^2(\mathbb{S}^{d-1})$ provided that $f_\beta \in \mathrm{H}^s(\mathbb{S}^{d-1})$. The relationship (3.8), evaluated at $v = 0$, implies that this can be achieved if we can estimate $R^-$ at the rate $N^{\frac{\sigma-\frac{d}{2}}{2\sigma+d-1}}$ in the $\|\cdot\|_{2,d/2}$

norm. The latter is the usual nonparametric rate for estimation of densities on $d-1$ dimensional smooth submanifolds of $\mathbb{R}^d$ (see, for example, Hendriks, 1990).

The estimation formula given in (3.6) is natural and reasonable, though it typically requires numerical evaluation of integrals to implement it. Moreover, in practice one needs to evaluate the infinite sum in (3.6), for example, by truncating the series. This results in a general estimator that can be written in the following two equivalent forms

$$\tilde{f}_\beta^- = \mathcal{H}^{-1}\left(P_{T_N}\tilde{R}^-\right) = \sum_{p=0}^{T_N} \frac{1}{\lambda(2p+1,d)} \int_{\mathbb{S}^{d-1}} q_{2p+1,d}(\cdot,x)\tilde{R}^-(x)d\sigma(x) \tag{3.9}$$

for suitably chosen $T_N$ that goes to infinity with $N$. The sequence $\mathcal{H}^{-1}\circ P_{T_N}, N=1,2,...$ can be interpreted as regularized inverses of $\mathcal{H}$, with the spectral cut-off method often used in statistical inverse problems. The next section gives an example of an estimator $\tilde{R}^-$ that implies a very simple closed form expression for $\tilde{f}_\beta$ that avoids numerical evaluation of the integrals in (3.6).

## 4. Estimators for the Choice Probability Function

This section considers estimation of the choice probability function $r$ and its extension $R$. We propose an estimator for $r$, which, in turn, yields a computationally simple estimator for $f_\beta$. Also the asymptotic results presented here are useful for the next section where we study the limiting properties of our estimator for the random coefficients density $f_\beta$.

Since $R$ is square integrable on $\mathbb{S}^{d-1}$, it has a condensed harmonic expansion which enables us to obtain the expressions in the next theorem.

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

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

This suggests an estimator of the form $\hat{R}_1(x) = \frac{1}{2} + \hat{R}_1^-$ with

$$\hat{R}_1^-(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{(2y_i-1)}{\hat{f}_X(x_i)} \sum_{p=0}^{T_N} q_{2p+1,d}(x_i,x)$$

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

$$\hat{R}_2^-(x) = \frac{1}{N}\sum_{i=1}^N \frac{(2y_i-1)}{\hat{f}_X(x_i)} K_{2T_N}^-(x_i,x) \tag{4.2}$$

where $K_{2T_N}^-$ is the odd part of a kernel of the form (2.13) satisfying Assumption 2.1, such as the two kernels in Proposition 2.3.

The estimator (4.2) is convenient, though the plug-in term $\hat{f}_X$ has to be treated with care. We avoid restrictive assumptions on the distributions of covariates and allow $f_X(x)$ to decay to zero as $x$ approaches the boundary of its support $H^+$. To deal with the latter problem, we modify (4.2) using a trimming factor to define

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

where $a_N$ is a sequence of the form

$$a_N = \log(N)^{-r} \tag{4.4}$$

for some positive $r$. Our estimator for $R$ is then

$$\widehat{R} = \frac{1}{2} + \hat{R}^-. \tag{4.5}$$

**Remark 4.1.** Alternative estimators of $R^-$ are available. For example, one may use kernel regression on the sphere to estimate $r$ in order to obtain an estimator for $R^-$. As noted before, however, we then need to use numerical integration to evaluate (3.9) to calculate $\hat{f}_\beta^-$.

Various nonparametric estimators for $f_X$ can be used in (4.3), since 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, and Hall et al. (1987) and Klemelä (2000) for higher dimensional spheres). We now assume that the following holds for $f_X$ and its estimator $\hat{f}_X$.

**Assumption 4.1.** *Suppose for $q$ and $\sigma$ that will be specified later*

(*i*)

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

*holds for some $r>0$, and $f_X$ and $\hat{f}_X$ satisfy either*

(*ii*)

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

*or*

(*iii*) *for some constant* $C$,

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

Assumption 4.1 (ii) (or (iii)) can be met easily when $f_X$ is smooth enough. In the simulation experiment we use

$$\hat{f}_X(x)=\max\left(\frac{1}{N}\sum_{i=1}^{N}K_{T_N}(x_i,x),0\right)\tag{4.6}$$

for a suitably chosen $T_N$ that depends on the sample size and the smoothness of $f_X$ and $K_{T_N}$ is a kernel of the form (2.13) satisfying Assumption 2.1. Theoretical details of this estimator will appear elsewhere but note that its rate of convergence in sup-norm can be obtained in the same manner as 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 derive a closed form using the condensed harmonic expansions together with the Addition Formula. Note also that $K_{T_N}$ is a smoothed projection kernel (note the factor $\chi$ in (2.13)), which is used here in order to have good approximation properties in the $\mathrm{L}^q(\mathbb{S}^{d-1})$ norms with arbitrary $q\in[1,\infty]$, in particular in the $\mathrm{L}^\infty(\mathbb{S}^{d-1})$ norm.

We now present asymptotic properties of the estimators for $R$. 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 rate of convergence, including the strong uniform convergence rate. Apart from 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 2.1, 3.1, 4.1*(i) *and 4.1*(ii) *hold. If $R$ belongs to $\mathrm{W}^\sigma_q(\mathbb{S}^{d-1})$ with $q$ in $[1,\infty]$ and $\sigma$ positive, and $T_N$ satisfies*

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

*then*

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

*Moreover, if Assumptions 4.1* (i) *and 4.1* (iii) *hold then there exists a constant $C$ such that*

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

Assumption 4.1(i) is used to achieve a rate of convergence logarithmically close to the desired nonparametric rate $N^{\frac{1}{2\sigma+d-1}}$. Relaxing it while still keeping the exponent $\frac{1}{2\sigma+d-1}$ (up to a logarithmic term) seems difficult.

Next we consider asymptotic normality:

**Theorem 4.3** (Asymptotic normality of $\hat{R}$)**.** *Suppose $R$ belongs $\mathrm{W}^{\sigma}_{\infty}(\mathbb{S}^{d-1})$ with $\sigma$ positive and Assumptions 2.1 and 3.1 hold. If $f_X$, $\hat{f}_X$, $f_X$, $T_N$ and $r$ satisfy*

$$N^{1/2}T_N^{-(d-1)/2}(\log N)^r\max_{i=1,\dots,N}\left|\frac{\max\left(f_X(x_i),\log(N)^{-r}\right)}{\max\left(\hat{f}_X(x_i),\log(N)^{-r}\right)}-1\right|=o_p(1),\tag{4.7}$$

$$N^{-1/2}T_N^{(d-1)/2}(\log N)^{r+\epsilon}=o(1)\quad\text{for some arbitrary }\epsilon>0,\tag{4.8}$$

$$N^{1/2}T_N^{-\frac{2\sigma+d-1}{2}}=o(1),\tag{4.9}$$

$$N^{1/2}T_N^{(d-1)/2}\sigma\left(\left\{0<f_X<(\log N)^{-r}\right\}\right)=o(1),\ \sup_{x\in H^+}f_X(x)<\infty,\tag{4.10}$$

*then*

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

*where*

$$s_{1N}^2(x):=\operatorname{var}\left(\frac{(2Y-1)K^{-}_{2T_N}(X,x)}{\max\left(f_X(X),(\log N)^{-r}\right)}\right).$$

The lower bound for the rate of $T_N$ implied by (4.9) is faster than the optimal rate (undersmoothing). This ensures that the approximation bias vanishes asymptotically. Condition (4.7) guarantees that the effect of replacing $f_X$ with $\hat{f}_X$ is also asymptotically negligible. Viewed as a condition on $f_X$ and $\hat{f}_X$, it becomes more stringent as the rate for $T_N$ gets slower, but as far as

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

holds, every $T_N$ that satisfies the lower bound (4.9) automatically fulfills (4.7). On the other hand, (4.8) imposes an upper bound for the growth rate of the parameter $T_N$. It is a technical condition under which the Lyapounov condition for asymptotic normality holds. Also, we impose (4.10) under which the bias due to trimming is asymptotically negligible. It becomes increasingly more restrictive as the growth rate for $T_N$ rises.

## 5. A Closed Form Estimator of $f_\beta$

This section presents a computationally convenient estimator for $f_\beta$, and shows that it has desirable asymptotic properties. It is based on an estimator for $f_\beta^-$ of the form

$$\hat{f}_\beta^- = \mathcal{H}^{-1}\left(\hat{R}^-\right) = \mathcal{H}^{-1}\left(\frac{1}{N}\sum_{i=1}^N \frac{(2y_i-1)K_{2T_N}^-(x_i,\cdot)}{\max\left(\hat{f}_X(x_i),(\log N)^{-r}\right)}\right).$$

Computing $\hat{f}_\beta^-$ is straightforward. First, note that the estimator (4.3) for $R^-$ resides in a finite dimensional space $\bigoplus_{p=0}^{T_N} H^{2p+1,d}$, therefore $P_{T_N}\hat{R}^- = \hat{R}^-$ holds. Consequently, unlike in (3.9) where a general estimator for $R^-$ is considered, we do not need to apply any additional series truncation to $\hat{R}^-$ prior to the inversion of $\mathcal{H}$. Second, the estimator requires no numerical integration. To see this, note the formula

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

which follows from

$$\begin{aligned}\int_{\mathbb{S}^{d-1}} q_{2p+1,d}(x,b)K_{2T_N}^{-1}(x,x_i)d\sigma(x) &= \int_{\mathbb{S}^{d-1}} q_{2p+1}(x,b)\sum_{p'=1}^{T_N-1}\chi(2p'+1,2T_N)q_{2p'+1,d}(x,x_i)d\sigma(x)\\ &= \chi(2p+1,2T_N)q_{2p+1,d}(b,x_i).\end{aligned}$$

Thus

$$\begin{aligned}\hat{f}_\beta^-(b) &= \frac{1}{N}\sum_{i=1}^N \frac{(2y_i-1)\mathcal{H}^{-1}\left(K_{2T_N}^-(x_i,\cdot)\right)(b)}{\max\left(\hat{f}_X(x_i),(\log N)^{-r}\right)}\\ &= \frac{1}{N}\sum_{i=1}^N \frac{(2y_i-1)\sum_{p=0}^{T_N-1}\frac{\chi(2p+1,2T_N)}{\lambda(2p+1,d)}q_{2p+1,d}(x_i,b)}{\max\left(\hat{f}_X(x_i),(\log N)^{-r}\right)}.\end{aligned}$$

Using (3.7) and the Addition formula, we arrive at an estimator for $f_\beta$ with the following explicit form:

$$\hat{f}_\beta(b) = 2\hat{f}_\beta^-(b)\mathbb{I}\{\hat{f}_\beta^-(b)>0\}, \tag{5.1}$$

$$\text{where } \hat{f}_\beta^-(b) = \frac{1}{|\mathbb{S}^{d-1}|}\sum_{p=0}^{T_N-1}\frac{\chi(2p+1,2T_N)h(2p+1,d)}{\lambda(2p+1,d)C_{2p+1}^{\nu(d)}(1)}\left(\frac{1}{N}\sum_{i=1}^N\frac{(2y_i-1)C_{2p+1}^{\nu(d)}(x_i'b)}{\max\left(\hat{f}_X(x_i),(\log N)^{-r}\right)}\right).$$

This is our main proposal, on which the rest of the paper focuses.

**Remark 5.1.** Our estimator $\hat{f}_\beta$ requires neither numerical integration nor optimization. Recall that $h(n,d)=\frac{(2n+d-2)(n+d-2)!}{n!(d-2)!(n+d-2)}$, $\nu(d)=(d-2)/2$ and $\lambda(2p+1,d)=\frac{(-1)^p|\mathbb{S}^{d-2}|1\cdot 3\cdots(2p-1)}{(d-1)(d+1)\cdots(d+2p-1)}$ by (2.6), Theorem 2.1 and Proposition 2.4(ii)(iv), respectively, so these are trivial to calculate. As discussed in Section 2.2 the polynomial $C_{2p+1}^{\nu(d)}$ can be evaluated recursively using (2.9). Examples of the specification of $\chi$ are given in Proposition 2.3.

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 Assumptions 2.1, 3.1, 4.1*(i) *and 4.1*(ii) *with* $\sigma=s+\frac{d}{2}$ *hold. If* $f_\beta^-$ *belongs to* $\mathrm{W}_q^s(\mathbb{S}^{d-1})$ *with* $q$ *in* $[1,\infty]$ *and* $s>0$, *and* $T_N$ *satisfies*

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

*then*

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

*Moreover, if Assumptions 4.1*(i) *and 4.1* (iii) *hold then there exists a constant* $C$ *such that*

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

The rate $N^{-\frac{s}{2s+2d-1}}$ is in accordance with the $\mathrm{L}^2$ 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, however, involves neither a plug-in method nor 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 $\mathcal{H}$. In a recent important paper, Hoderlein et al. (2007) study a linear model of the form $W=X'\beta$ where $\beta$ is a $d$-vector of random coefficients. They obtain a nonparametric random coefficients density estimator that has the rate $N^{-\frac{s}{2s+2d-1}}$ without the log correction,[2] when $f_X$ is assumed to be bounded from below and thus no trimming is required. Our log correction is closely related to the speed at which the density $f_X$ decays to zero as $x$ approaches the boundary of $H^+$. Also, our result covers $\mathrm{L}^q$ loss for all $q\in[1,\infty]$.

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

[2]Note that the dimension of their estimator is $d$, whereas that of ours is $d-1$. On the other hand, in their problem $W$ is observable, and it is obviously more informative than our binary outcome $Y$, which causes difficulties both in identification and estimation.

**Theorem 5.2** (Asymptotic normality)**.** *Suppose $f_\beta^-$ belongs to $\mathrm{W}^s_\infty(\mathbb{S}^{d-1})$ with $s>0$, and Assumptions 2.1 and 3.1 hold. If $\hat{f}_X$, $f_X$, $T_N$ and $r$ satisfy*

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

$$N^{-1/2}T_N^{(d-1)/2}(\log N)^{r+\epsilon}=o(1) \quad \text{for some arbitrary } \epsilon>0, \tag{5.5}$$

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

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

*then*

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

*holds for $b$ such that $f_\beta(b)\neq 0$, where* $s_N^2(b):=4\mathrm{var}(Z_N(b)), Z_N(b)=\frac{(2Y-1)\mathcal{H}^{-1}\left(K^-_{2T_N}(X,\cdot)\right)(b)}{\max(f_X(X),(\log N)^{-r})}$.

Note that the conditions (5.4), (5.5), (5.6) and (5.7) are the same as conditions (4.7), (4.8), (4.9) and (4.10) in the case of estimation of $R$. To see this for (5.6) it is enough to set $\sigma=s+\frac{d}{2}$. The standard error $s_N(b)$ is 2 times the standard deviation of

$$Z_N(b)=\frac{1}{|\mathbb{S}^{d-1}|}\sum_{p=0}^{T_N-1}\frac{\chi(2p+1,2T_N)h(2p+1,d)}{\lambda(2p+1,d)C_{2p+1}^{\nu(d)}(1)}\left(\frac{(2Y-1)C_{2p+1}^{\nu(d)}(X'b)}{\max\left(f_X(X),(\log N)^{-r}\right)}\right)$$

(see equation (5.1)), which can be estimated by replacing $f_X$ with $\hat{f}_X$.

## 6. Discussion

6.1. **Estimation of Marginals.** In Section 3 we have provided an expression for the estimator of the full joint density of $\beta$, from which an estimator for a marginal density can be obtained. Let $\sigma_k$ denote the surface measure and $\underline{\sigma}_k=\sigma_k/|\mathbb{S}^k|$ the uniform probability 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 $d-k$. 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 $\overline{P}_*\underline{\sigma}_{d-1}$ and $\overline{\overline{P}}_*\underline{\sigma}_{d-1}$ the direct image probability measures. One possibility is to define the marginal law of $\overline{\overline{\beta}}$ as the measure $\overline{\overline{P}}_*P_\beta$, where $dP_\beta=f_\beta d\sigma$. This may not be convenient, however, since the uniform distribution over $\mathbb{S}^{d-1}$ would have U-shaped marginals. The U-shape becomes more pronounced as the dimension of $\beta$ 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 $\overline{\overline{P}}_*\underline{\sigma}_{d-1}$. 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 $\overline{\overline{P}}_*\underline{\sigma}_{d-1}$ is given by

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

One can use deterministic methods to compute the integral (e.g., Hesse et al. (2007) for quadrature methods on the sphere) or for example one may use a Monte-Carlo method, by forming

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

where $u_j, j=1,...,M$ are draws from independent uniform random variables on $\mathbb{S}^{d-1-k}$.

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.[3] First, the convergence rate of our estimator of the joint density of $\beta$ slows down as the dimension $d$ of $\beta$ grows, which is a manifestation of the curse of dimensionality. Treating some coefficients 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. 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

[3] Hoderlein et al. (2007) suggest a method to deal with non-random coefficients in their treatment of random coefficient linear regression models.

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 $\beta_1^*(Z_2) := \beta_1 + \alpha_2' Z_2$. 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:

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

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 conditions on $f_{W|Z_2=z_2}$, the conditional density of $W$ given $Z_2 = z_2$. More specifically, suppose there exists a set $\mathcal{Z}_2 \subset \mathbb{R}^{d_Z-1}$, 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 (2.16) we obtain

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

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

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

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

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

for all $z_2 \in \mathcal{Z}_2$ as well. If we are only interested in the joint distribution of $\beta_2$ 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

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

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

$$f\left(\frac{(\beta_2', \alpha_1, \alpha_2')'}{\|\beta_2\|}\right) \tag{6.7}$$

denote the joint density of all the coefficient (except for $\beta_1$, which corresponds to the conventional disturbance term in the original model (6.4), normalized by the length of $\beta_2$). Then

$$f\left(\frac{(\beta_2', \alpha_1, \alpha_2')'}{\|\beta_2\|}\right) = f\left(\begin{bmatrix} I_{d_X-1} & 0 \\ 0 & 1 \\ \vdots & \frac{c\alpha_2}{c\alpha_1} \end{bmatrix} \begin{bmatrix} \frac{\beta_2}{\|\beta_2\|} \\ \frac{\alpha_1}{\|\beta_2\|} \end{bmatrix}\right).$$

In the expression on the right hand side, $f\left((\beta_2', \alpha_1)'/\|\beta_2\|\right)$ is 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 $(\beta_2', \alpha_1, \alpha_2')'$ 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 $f_{W|Z_2=z_2}$ 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 $(\beta, V)$. Note that Hoderlein et al. (2007) utilize a linear structure of the form (6.9) in estimating a random coefficient linear model. 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 $\Gamma Z$ 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 $\Gamma$ 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. Numerical Examples

The purpose of this section is to illustrate the performance of our new estimator in finite samples using simulated data. We consider the model of the form (1.1) with $d = 3$. The covariates are specified to be $X = (1, X_1, X_2)$ where $(X_2, X_3)' \sim N(\binom{0}{0}, 2 \cdot I_2)$. The coefficients vector $\beta = (\beta_1, \beta_2, 1)'$ is set random except for the last element. Fixing the last component constant fulfills Assumption 3.2 for identification. Two specifications for the random elements $(\beta_1, \beta_2)$ are considered. In the first specification (Model 1) we let $(\beta_1, \beta_2)' \sim N(\binom{0}{0}, 0.3 \cdot I_2)$. In the second (Model 2) we consider a two point mixture of normals

$$\binom{\beta_1}{\beta_2} \sim \lambda N\left(\binom{\mu}{-\mu}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}\right) + (1-\lambda) N\left(\binom{-\mu}{\mu}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}\right),$$

where $\mu = 0.7, \sigma^2 = 0.3, \rho = 0.5$ and $\lambda = 0.5$. Random samples of size 500 from each of the two specifications are generated, then the new estimator (5.1) is computed. It is implemented using the Riesz kernel with $s = 2$ and $l = 3$ (see Proposition 2.3). The truncation parameter $T_N$ is set at 3, and the trimming parameter $r$ is 2. It also requires a nonparametric estimator for $f_X$, and we use the projection estimator (4.6) based on the same Riesz kernel (i.e. $s = 2$, $l = 3$) and $T_N = 10$.

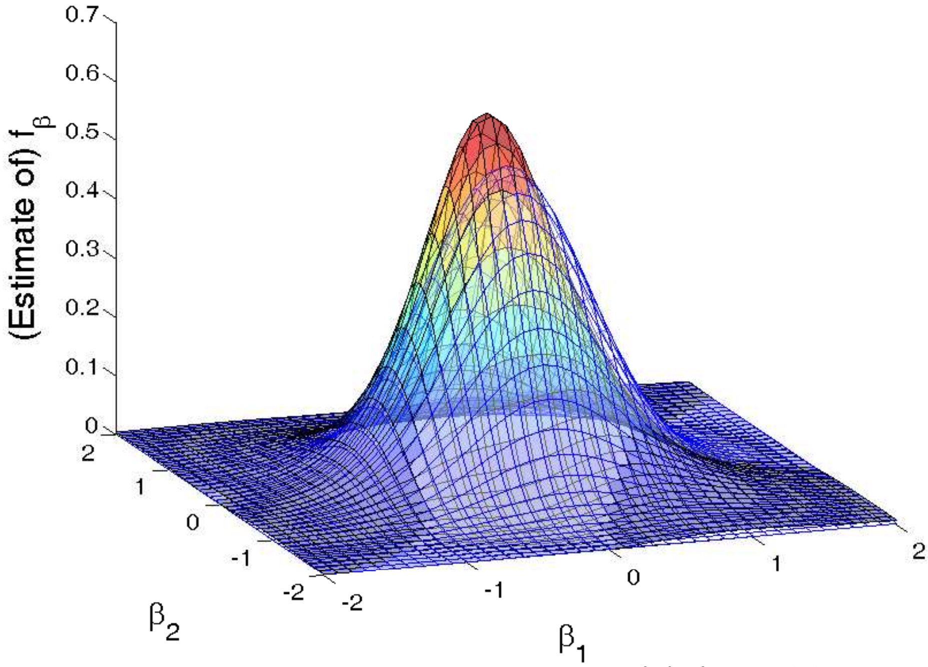


FIGURE 1. Nonparametric estimator of $f_\beta$ for Model 1

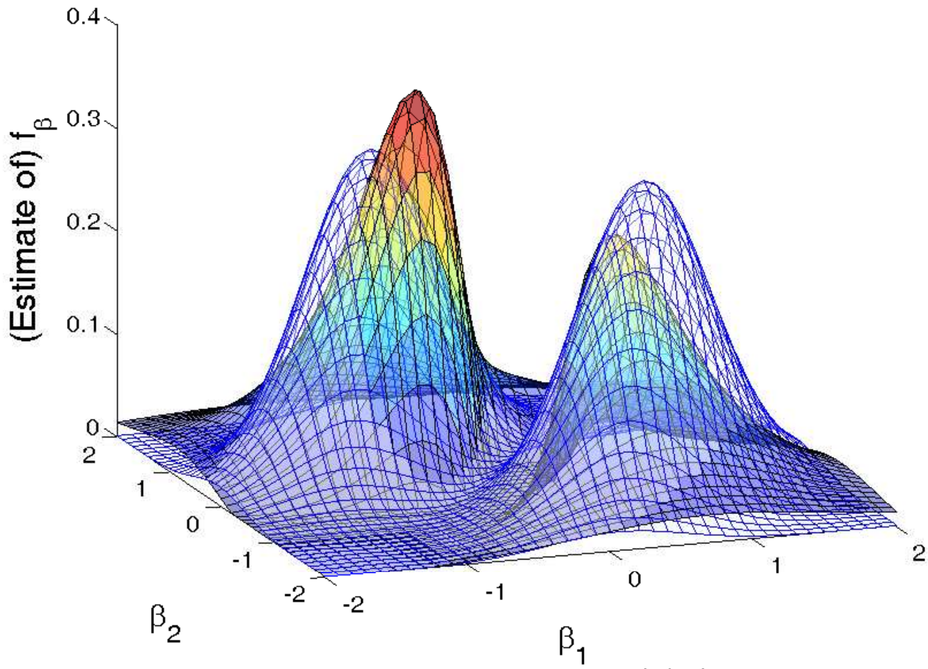


FIGURE 2. Nonparametric estimator of $f_\beta$ for Model 2

Figure 1 presents the surface plot of the true density (blue mesh) and our estimate (multi-colored surface) for Model 1. Our estimator (5.1) is defined on $\mathbb{S}^2$ in this case, and we performed an appropriate transformation to plot it as a density on $\mathbb{R}^2$. With the reasonable sample size, the location of the peak of the density, as well as its shape, are successfully recovered by our procedure. Next, Figure 2 shows the estimation results for Model 2. Again, our procedure works well: the estimated surface plot nicely captures the locations of the two peaks and their shapes of the true density, thereby exhibiting the underlying mixture structure. While further experimentations are necessary, these results seem to indicate our estimator's good performance in practical settings.

## 8. Conclusion

In this paper we have considered nonparametric estimation of a random coefficients binary choice model. By exploiting (previously unnoticed) connections between the model and statistical deconvolution problems and applying results of integral transformation on the sphere, we have developed a new estimator that is practical and possesses desirable statistical properties. It requires neither numerical optimization nor numerical integration, and as such its computational cost is trivial and local maxima and other difficulties in optimization need not be of concern. Its rate of convergence in the $\mathrm{L}^q$ norm for all $q \in [1, \infty]$ is derived. Our numerical example suggests that the new procedure works well in finite samples, consistent with its good theoretical properties. It is of great theoretical interest to examine rigorously whether the rate is optimal in a minimax sense, though it is a task we defer to subsequent investigations. With appropriate under-smoothing, the estimator is shown to be asymptotically normal, providing a theoretical basis for nonparametric statistical inference for the random coefficient distribution.

## 9. Appendix

We first summarize some results on the Gegenbauer polynomials, which are used in various parts of the paper. These can be found in Erdélyi et al. (1953) and Groemer (1996). The Gegenbauer polynomials have the following explicit representation

$$C_n^{\nu}(t) = \sum_{l=0}^{[n/2]} \frac{(-1)^l (\nu)_{n-l}}{l!(n-2l)!} (2t)^{n-2l} \tag{9.1}$$

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 $\nu = 0$ and $d = 2$, it is related to the Chebychev polynomials of the first kind, as

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

and

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

hold for

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

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

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

The Gegenbauer polynomials are related to each other through differentiation, that is, they satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t} C_n^{\nu}(t) = 2\nu C_{n-1}^{\nu+1}(t) \tag{9.2}$$

for $\nu > 0$ and

$$\frac{\mathrm{d}}{\mathrm{d}t} C_n^{0}(t) = 2 C_{n-1}^{1}(t). \tag{9.3}$$

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

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

The following results are also used in the paper:

$$\sup_{t\in[-1,1]} \left| \frac{C_n^{\nu}(t)}{C_n^{\nu}(1)} \right| \leq 1, \tag{9.5}$$

$$\forall\ \nu > 0,\ \forall n \in \mathbb{N},\ C_n^{\nu}(1) = \begin{pmatrix} n+2\nu-1 \\ n \end{pmatrix} \tag{9.6}$$

$$C_0^0(1) = 1 \text{ and } \forall n \in \mathbb{N}\setminus\{0\},\ C_n^0(1) = \frac{2}{n}, \tag{9.7}$$

$$C_n^{\nu}(-t) = (-1)^n C_n^{\nu}(t). \tag{9.8}$$

These orthogonal polynomials are normalized such that

$$\|C_n^{\nu(d)}\left(x'\cdot\right)\|_2 = \int_{-1}^{1} (C_n^{\nu(d)}(t))^2 (1-t^2)^{(d-3)/2} dt = \frac{|\mathbb{S}^{d-1}|(C_n^{\nu(d)}(1))^2}{|\mathbb{S}^{d-2}| h(n,d)}. \tag{9.9}$$

In the proofs we often denote a constant that depends only on the dimension $d$ by $C$, thus its value is determined by the context it is used.

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

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

*Proof.* Using (2.8), (9.2), (9.3), (9.6) and (2.6)

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

The desired result follows, since, using again (9.6) and (2.6),

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

□

**Proof of Proposition 2.3**. First consider the Riesz kernel. (i) follows from (2.4) in Ditzian (1998) and by the fact that Cesàro kernels $C_h^l$ are uniformly bounded in $\mathrm{L}^1(\mathbb{S}^{d-1})$ for $l > \frac{d-2}{2}$ (see, e.g. Bonami and Clerc 1973, p. 225). To show (iii) we use Theorem 4.1 in Ditzian (1998), by letting $P(D) = \Delta^S$, $\lambda = \zeta_{T,d} + 1 = T(T+d-2)+1$, $\alpha = s/2$ and $m = 1$. Then it implies an approximation error upper bound $CK_{s/2}(f, \Delta^S, (\zeta_{T,d}+1)^{-\frac{s}{2}})$, which, in turn, is bounded by $CT^{-s}\|(-\Delta^S)^{s/2}f\|_p$ (see equations (4.2) and (4.1) therein). By the definition of the norm of the Sobolev space $W_p^s(\mathbb{S}^{d-1})$ (see Definition 2.3) the result follows. Concerning the delayed means, (i) corresponds to the inequality (A16) of Hesse et al. (2007). To see (iii), use Proposition 15 in Hesse et al. (2007) to obtain an upper bound $C\inf_{g\in\bigoplus_{n=0}^{T/2}H^{n,d}}\|f-g\|_p$. Let $\lambda = \zeta_{T/2,d}+1 = \frac{T}{2}(\frac{T}{2}+d-2)+1$, $\alpha = s/2, m = 1, P(D) = \Delta^S$ in Ditzian's (1998) Theorem 6.1, which gives an upper bound on the best spherical harmonic approximation in $\mathrm{L}^p(\mathbb{S}^{d-1})$ to functions in $\mathrm{W}_p^s(\mathbb{S}^{d-1})$ (see also Kamzolov, 1983), then apply equation (4.1) in Ditzian (1998) again to obtain the desired result. The proof of (ii) for both Riesz and delayed means kernels is as follows. Write

$$|K_T(z,x) - K_T(z,y)| \leq \sum_{n=0}^{T}\chi(n,T)\left|{}^{\flat}q_{n,d}(z'x) - {}^{\flat}q_{n,d}(z'y)\right|$$

where

$$\left|{}^\flat q_{n,d}(z'x) - {}^\flat q_{n,d}(z'y)\right| = \left|\int_{z'x}^{z'y} \left(\frac{\mathrm{d}}{\mathrm{dt}}{}^\flat q_{n,d}\right)(t)dt\right|$$
$$\leq \frac{d|\mathbb{S}^{d+1}|}{|\mathbb{S}^{d-1}|}\left\|{}^\flat q_{n-1,d+2}\right\|_\infty |x-y| \quad \text{(by lemma 9.1)}$$
$$\leq \frac{d|\mathbb{S}^{d+1}|}{|\mathbb{S}^{d-1}|^2} h(n-1,d+2)|x-y| \quad \text{(by (2.8) and (9.5))}$$

and conclude using that $\chi(n,T) \in [0,1]$ and (9.10) below. (iv) holds by setting $c$ to $(1/2)^l$ in the case of the Riesz kernel and to 1 in the case of the delayed means. □

The following results are useful.

**Lemma 9.2.**

$$h(n,d) \asymp n^{d-2}, \tag{9.10}$$

$$|\lambda(2p+1,d)| \asymp p^{-d/2}. \tag{9.11}$$

*Proof.* Estimate (9.10) 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

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

and the results follow. Next we turn to (9.11). 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

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

and (9.11) follows. Sterling's double inequality (see Feller (1968) p.50-53), that is,

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

implies that

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

and therefore

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

Thus 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 (9.11) holds for both even and odd $d$. $\square$

**Proof of Proposition 2.4**. Define $\alpha(n,d) := C_n^{\nu(d)}(1)|\mathbb{S}^{d-2}|^{-1}\lambda_n\left(\mathbb{I}\left\{t\in[0,1]\right\}\right)$. By the Funk-Hecke theorem

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

thus using (9.4),

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

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

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

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 the result follows since $k_1$ in the sum cannot be equal to 0.

When $n = 2p+1$ for $p \in \mathbb{N}$ we obtain, again using the Faá di Bruno formula, that the derivative at $t=0$ is equal to

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

Together with (9.6), the desired result follows. For the case $d=2$ we use Proposition 2.1. $\square$

**Proof of Proposition 2.6**. By definition we have

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

where according to the Funk-Hecke Theorem

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

The result follows since Lemma 9.2 implies that $(1+\zeta_{2p+1,d})^{s+d/2}\lambda^2(2p+1,d) \asymp (1+\zeta_{2p+1,d})^s$. □

**Proof of Proposition 2.7**. If we consider the case where $d$ is even, we know from Proposition 2.4, that

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

Thus if $d$ is a multiple of 4,

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

Using this and (2.16),

$$\begin{aligned}
\mathcal{H}^{-1} &= \sum_{p=0}^{\infty}\frac{1}{\lambda(2p+1,d)}Q_{2p+1,d}\\
&= \sum_{p=0}^{\infty}|\mathbb{S}^{d-2}|\left(\prod_{k=1}^{d/4}[-\zeta_{2p+1,d}+2(k-1)(d-2k)]\right)Q_{2p+1,d}.
\end{aligned}$$

Recall Definition 2.3 and the proposition is proved. □

**Proof of Theorem 2.3**. We can write

$$\mathcal{H}^{-1} = P_1(D) - P_2(D)$$

where $P_1(D)$ and $P_2(D)$ are defined for all odd function $f^-$ by

$$\begin{aligned}
P_1(D)f^- &= \sum_{p=0}^{\infty}\frac{1}{\lambda(4p+3)}\int_{\mathbb{S}^{d-1}} q_{4p+3}(x,y)f^-(x)d\sigma(x)\\
P_2(D)f^- &= -\sum_{p=0}^{\infty}\frac{1}{\lambda(4p+1)}\int_{\mathbb{S}^{d-1}} q_{4p+1}(x,y)f^-(x)d\sigma(x).
\end{aligned}$$

$P_1(D)$ and $P_2(D)$ are two unbounded operators on $B = \mathrm{L}^q_{\mathrm{odd}}(\mathbb{S}^{d-1})$ with non-positive eigenvalues. We apply Theorem 3.2. of Ditzian (1998) to $-P_1(D)$ and $-P_2(D)$ choosing $\alpha = 1$. Condition (1.6) of

Ditzian (1998) can be verified using Proposition 2.2 with $r = 1$ and $p = q$ and the fact that for the Cesaro kernels $C_h^l$ are uniformly bounded in $\mathrm{L}^1(\mathbb{S}^{d-1})$ for $l > \frac{d-2}{2}$ (see, e.g. Bonami and Clerc, 1973). We see, using the triangle inequality, that for all $P$ in $\bigoplus_{p=0}^{T} H^{2p+1,d}$,

$$\begin{aligned}\|\mathcal{H}^{-1}P\|_q &\leq C\frac{1}{\lambda^2(2T+1,d)}\|P\|_q \\ &\leq CT^d\|P\|_q.\end{aligned}$$

The last inequality follows from (9.11). □

**Proof of Proposition 3.1**. It is straightforward that the model (1.1) and Assumption 1.1 imply that the choice probability function $r$ given by (1.2) is homogeneous of degree 0. Proposition 2.5 along with the fact that $R = \frac{1}{2} + \mathcal{H}\left(f_\beta^-\right)$ with $f_\beta^- \in \mathrm{L}^2_{\mathrm{odd}}(\mathbb{S}^{d-1})$ implies that $R$ belongs to $\mathrm{H}^{d/2}(\mathbb{S}^{d-1})$. We now turn to the proof of sufficiency. If the extension $R$ given by (3.2) belongs to $\mathrm{H}^{d/2}(\mathbb{S}^{d-1})$ then so does $R^-$ and Proposition 2.5 shows that there exists a unique odd function $f^-$ in $\mathrm{L}^2(\mathbb{S}^{d-1})$ such that

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

Moreover, since $0 \leq R(x) \leq 1$ holds for every $x \in \mathbb{S}^{d-1}$, the above relationship implies that $\frac{1}{2} \geq \mathcal{H}f^-(x), \forall x \in \mathbb{S}^{d-1}$. But $\mathcal{H}f^-(x) \geq \int_{\{f^-(b)\geq 0\}} f^-(b)d\sigma(b)$ holds for some $x$. Therefore we conclude that $\frac{1}{2} \geq \int_{\{f^-(b)\geq 0\}} f^-(b)d\sigma(x) = -\int_{\{f^-(b)\leq 0\}} f^-(b)d\sigma(b)$, thus $\int_{\mathbb{S}^{d-1}} |f^-(b)|d\sigma(b) \leq 1$. Also, following the discussion in Section 2.2, $\frac{1}{|\mathbb{S}^{d-1}|} + f^-$ integrates to 1. We have seen in Corollary 2.1 that for even function $g$ that has 0 as the coefficient of degree 0 in its expansion on the surface harmonics (i.e. an even function that integrates to zero over the sphere),

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

holds. Now consider

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

then this certainly is even and integrates to zero. Using this, define

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

Obviously $f_\beta^{*-} = f^-$. This function $f_\beta^*$ is non-negative and integrates to one, and thus it is a proper probability density function (pdf). It is indeed bounded from below by $\frac{1}{|\mathbb{S}^{d-1}|}\left(1 - \int_{\mathbb{S}^{d-1}} |f^-(b)|d\sigma(b)\right)$. As a consequence, there exists a pdf $f_\beta^*$ such that

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

and for all $x$ in $H^+$, $r(x)=\mathcal{H}\left(f_\beta^*\right)(x)$. □

**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 (9.8),

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

□

**Proof of Theorems 4.2 and 4.3**. The proofs concerning the estimation of $R$ is the same as that of $f_\beta$ below (though the latter requires a step that uses Theorem 2.3, which is not necessary for the former). □

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

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

where

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

We use the decomposition

$$\hat{f}_\beta^--f_\beta^-=\left(\hat{f}_\beta^--\overline{f}_{\beta,T}^{\,-}\right)-\left(\overline{f}_{\beta,T}^{\,-}-\mathbb{E}\left[\overline{f}_{\beta,T}^{\,-}\right]\right)-\left(\mathbb{E}\left[\overline{f}_{\beta,T}^{\,-}\right]-\mathbb{E}\left[\overline{f}_\beta^{\,-}\right]\right)-\left(\mathbb{E}\left[\overline{f}_\beta^{\,-}\right]-f_\beta^-\right),\tag{9.12}$$

and denote the terms on the right hand side by $S_{\mathrm{p}}$ (stochastic component due to plug-in), $S_{\mathrm{e}}$ (stochastic component of the infeasible estimator $\overline{f}^{-}_{\beta,T}$), $B_{\mathrm{t}}$ (trimming bias) and $B_{\mathrm{a}}$ (approximation bias). Note that the same decomposition, with $\mathcal{H}$ operated on each term, can be used to show Theorems 4.2 and 4.3.

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

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

Obviously

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

and $A_4 = 0$. Also,

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

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

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

Similarly,

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

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

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

Overall,

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

A similar proof can be carried out replacing $\mathrm{L}^q(\mathbb{S}^{d-1})$ by $\mathrm{L}^\infty(\mathbb{S}^{d-1})$. Thus it is enough to consider the behavior of $\hat{f}_\beta^- - f_\beta^-$ instead of $\hat{f}_\beta - f_\beta$. A noted above, the former can be decomposed into four terms, $S_{\mathrm{p}}$, $S_{\mathrm{e}}$, $B_{\mathrm{t}}$ and $B_{\mathrm{a}}$.

Let the sequence of smoothing parameters satisfy:

$$T_N = \left( \frac{N}{(\log N)^{2(r+(1/2-1/q)\mathbb{I}\{q\geq 2\})}} \right)^{\gamma} \tag{9.13}$$

for some $\gamma > 0$. We later show that the above form with the choice $\gamma = \frac{1}{2s+2d-1}$ leads to the optimal rate of convergence for $\hat{f}_\beta$.

We start with the analysis of $S_{\mathrm{p}}$. Note that for $q \in [1, \infty]$

$$\begin{aligned}
\|S_{\mathrm{p}}\|_q &= \left\| \mathcal{H}^{-1} \left( \frac{1}{N} \sum_{i=1}^N \frac{(2y_i - 1) K_{2T_N}^-(x_i, \cdot)}{\max(f_X(x_i), (\log N)^{-r})} \left( \frac{\max(f_X(x_i), (\log N)^{-r})}{\max\left(\hat{f}_X(x_i), (\log N)^{-r}\right)} - 1 \right) \right) \right\|_q \\
&\leq B(d,q) T_N^{d/2} \left\| \frac{1}{N} \sum_{i=1}^N \frac{(2y_i - 1) K_{2T_N}^-(x_i, \cdot)}{\max(f_X(x_i), (\log N)^{-r})} \left( \frac{\max(f_X(x_i), (\log N)^{-r})}{\max\left(\hat{f}_X(x_i), (\log N)^{-r}\right)} - 1 \right) \right\|_q \quad \text{(by Theorem 2.3)} \\
&\leq B(d,q) T_N^{d/2} (\log N)^r \left\| \frac{1}{N} \sum_{i=1}^N |K_{2T_N}(x_i, \cdot)| \right\|_q \max_{i=1,\dots,N} \left| \frac{\max(f_X(x_i), (\log N)^{-r})}{\max\left(\hat{f}_X(x_i), (\log N)^{-r}\right)} - 1 \right|
\end{aligned}$$

holds, where we have used the triangle inequality. The $\mathrm{L}^q$-norm on the right hand side is bounded from above by

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

First consider the term $\|T_1\|_q$. We begin with the case of $q \in [1, 2]$. By the Hölder inequality,

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

where

$$\begin{aligned}
\mathbb{E}\left[T_1(x)^2\right] &\le \frac{1}{N}\mathbb{E}\left[(K_{2T_N}(X,x))^2\right] \\
&\le \frac{C}{N}\left\|K_{2T_N}(\star_2,x)\right\|_2^2 \quad \text{(boundedness assumption on } f_X) \\
&= \frac{C}{N}\left\|\sum_{n=0}^{2T_N}\chi(n,2T_N)q_{n,d}(\star_2,x)\right\|_2^2 \\
&\le \frac{C}{N}\sum_{n=0}^{2T_N}\left\|q_{n,d}(\star_2,x)\right\|_2^2 \quad \text{(by Assumption 2.1(iv))} \\
&\le \frac{C}{N}\sum_{n=0}^{2T_N}\frac{h^2(n,d)\left\|C_n^{\nu(d)}(\star_2'x)\right\|_2^2}{|\mathbb{S}^{d-1}|^2(C_n^{\nu(d)}(1))^2} \\
&\le \frac{C}{N}\sum_{n=0}^{2T_N}h(n,d) \quad \text{(by (9.9))} \\
&\le \frac{CT_N^{d-1}}{N} \quad \text{(by lemma 9.2).}
\end{aligned} \tag{9.15}$$

By the Markov inequality,

$$T_N^{d/2}(\log N)^r\|T_1\|_q = O_p\left((\log N)^r N^{-1/2}T_N^{(2d-1)/2}\right), \tag{9.16}$$

providing a convergence rate for $\|T_1\|_q, q \in [1,2]$. So if we can establish a similar rate for $\|T_1\|_\infty$, all $\mathrm{L}^q(\mathbb{S}^{d-1})$ convergence rates of $T_1$ for $q \in (2,\infty]$ can be interpolated between the $\mathrm{L}^2(\mathbb{S}^{d-1})$ and $\mathrm{L}^\infty(\mathbb{S}^{d-1})$ convergence rates using the following inequality:

$$\forall f \in \mathrm{L}^\infty(\mathbb{S}^{d-1}),\ \|f\|_q \le \|f\|_2^{2/q}\|f\|_\infty^{1-2/q}. \tag{9.17}$$

To see this, note

$$\begin{aligned}
\|f\|_q &= \|f^2|f|^{q-2}\|_1^{1/q} \\
&\le \left[\|f^2\|_1\||f|^{q-2}\|_\infty\right]^{1/q} \quad \text{(by Hölder)} \\
&= \|f\|_2^{2/q}\|f\|_\infty^{1-2/q}.
\end{aligned}$$

We can thus focus on $\|T_1\|_\infty$. We cover the sphere $\mathbb{S}^{d-1}$ 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 $R(N,r,d)$, that is, $B_i = \{x \in \mathbb{S}^{d-1} :\ \|x-\tilde{x}_i\| \le R(N,r,d)\}$. As the notation suggests, we let the radius of the balls depend on $N$, $r$ and $d$, as specified more precisely below. Note that $\mathfrak{N}(N,r,d) \asymp R(N,r,d)^{-(d-1)}$.

We now prove that for every $\epsilon > 0$ positive, there exists a positive $M$ such that

$$\mathbb{P}\left( v_N T_N^{d/2} (\log N)^r \sup_{x \in \mathbb{S}^{d-1}} |T_1(x)| \geq M \right) \leq \epsilon \tag{9.18}$$

holds for an appropriately chosen sequence $v_N \uparrow \infty$. Write

$$\begin{aligned}
&\mathbb{P}\left( v_N T_N^{d/2} (\log N)^r \sup_{x \in \mathbb{S}^{d-1}} |T_1(x)| \geq M \right) \\
&\leq \mathbb{P}\left( \bigcup_{i=1,\dots,\mathfrak{N}(N,r,d)} \left\{ v_N T_N^{d/2} (\log N)^r |T_1(\tilde{x}_i)| \geq M/2 \right\} \right) \\
&\quad + \mathbb{P}\left( \exists i \in \{1,\dots,\mathfrak{N}(N,r,d)\} : \; v_N T_N^{d/2} (\log N)^r \sup_{x \in B_i} |T_1(x) - T_1(\tilde{x}_i)| \geq M/2 \right) \\
&\leq \mathfrak{N}(N,r,d) \sup_{i=1,\dots,\mathfrak{N}_N} \mathbb{P}\left( v_N T_N^{d/2} (\log N)^r |T_1(\tilde{x}_i)| \geq M/2 \right)
\end{aligned} \tag{9.19}$$

where the last inequality is obtained using Assumption 2.1 (ii) on the kernel and letting $R(N,r,d) \asymp (\log N)^{-r} v_N^{-1} T_N^{-(d/2+\alpha)} M$ (where $\alpha$ is given in Assumption 2.1 (ii)). Notice

$$\begin{aligned}
&\mathbb{P}\left( v_N T_N^{d/2} (\log N)^r |T_1(\tilde{x}_i)| \geq M/2 \right) \\
&= \mathbb{P}\left( \left| \sum_{j=1}^N \frac{|K_{2T_N}(x_j, \tilde{x}_i)|}{T_N^{d-1}} - \mathbb{E}\left[ \frac{|K_{2T_N}(X, \tilde{x}_i)|}{T_N^{d-1}} \right] \right| \geq T_N^{-(d-1)} v_N^{-1} T_N^{-d/2} (\log N)^{-r} NM/2 \right) \\
&\leq 2 \exp\left\{ -\frac{1}{2} \left( \frac{t^2}{\omega + Lt/3} \right) \right\} \quad \text{(Bernstein inequality)}
\end{aligned} \tag{9.20}$$

where

$$\begin{aligned}
t &= T_N^{-(d-1)} v_N^{-1} T_N^{-d/2} (\log N)^{-r} NM/2 \\
\omega &\geq \sum_{j=1}^N \operatorname{var}\left( \frac{|K_{2T_N}(X_j, \tilde{x}_i)|}{T_N^{d-1}} \right) \\
&\forall j = 1,\dots,N, \; \left| \frac{K_{2T_N}(X_j, \tilde{x}_i)}{T_N^{d-1}} \right| \leq L \quad \text{(using (2.8) and (9.5))}.
\end{aligned}$$

The bound $L$ in the last line is obtained by noting that $|K_{2T_N}(X_j, \tilde{x}_i)| = \left| \sum_{n=0}^{2T_N} \chi(n, 2T_N) q_{n,d}(X_j, \tilde{x}_i) \right| \leq C \sum_{n=0}^{2T_N} |h(n,d)| \asymp T_N^{d-1}$, which follows from (2.8), (9.5) and (9.10). Here we can take $\omega = CN\mathbb{E}[K_{2T_N}(X, \tilde{x}_i)^2]$, then by the calculations in (9.15), we can write $\omega = CNT_N^{-(d-1)}$. $\omega$ is the leading term in the denominator of the exponent in the last inequality.

If we take $v_N = (\log N)^{-r-1/2} N^{1/2} T_N^{-(2d-1)/2}$, then

$$\frac{t^2}{\omega + Lt/3} \asymp (\log N) M^2. \tag{9.21}$$

Also, use this $v_N$ and the form of $T_N$ as specified in (9.13) in our choice of $R(N,r,d)$ made above to get:

$$R(N,r,d) \asymp (\log N)^{-r} v_N^{-1} T_N^{-(d/2+\alpha)} M = (\log(N))^{1/2} N^{-1/2} T_N^{\frac{d-1}{2}-\alpha} M$$

Thus

$$\mathfrak{N}(N,r,d) \asymp R(N,r,d)^{-(d-1)} = \exp\left(\frac{1}{2}(d-1)\log N + o(\log N)\right) = \exp\left(C_1 \log N + o(\log N)\right) \tag{9.22}$$

with $C_1 = \frac{1}{2}(d-1)$. (9.19), (9.20), (9.21) and (9.22) imply that, for a positive constants $C$ and $C_2$,

$$\mathbb{P}\left(v_N T_N^{d/2} (\log N)^r \sup_{x \in \mathbb{S}^{d-1}} |T_1(x)| \geq M\right) \leq C \exp\left\{(\log N)(C_1 - C_2 M^2)\right\} \tag{9.23}$$

holds. For a large enough $M$, $C_1 - C_2 M^2 < 0$ and the right hand side of (9.23) converges to zero, so (9.18) follows. In summary, we have just shown that

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

and with (9.16) and (9.17) we also conclude that

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

Concerning $\|T_2\|_q$, $q \in [1,\infty]$, since $f_X$ is bounded by assumption, there exists a positive $C$ such that

$$\|T_2\|_q \leq C \left\| \|K_{2T_N}(\star_1, \star_q)\|_1 \right\|_q$$

where integration in $\|\cdot\|_1$ is with respect to argument $\star_1$ and integration in $\|\cdot\|_q$ is with respect to $\star_q$. But $\|K_{2T_N}(\star_1, \star_q)\|_1$ is a constant and does not depend on $\star_q$, as previously noted. Thus

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

and we conclude that this term is $O(1)$ using Assumption 2.1 (i) on the kernel, 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 smaller order than the first term $T_N^{d/2} (\log N)^r \|T_1\|_q$.

Analogously to our treatment of $\|T_1\|_q$, we can prove that when $q \in [1,2]$,

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

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

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

Let us now turn to the bias term induced by trimming

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

Using Theorem 2.3 along with Proposition 2.2 with $r = q$ and $p = 1$, where the $\mathrm{L}^q$-norm of the kernel is interpolated using Hölder's inequality between the uniformly bounded $\mathrm{L}^1$-norm and the upper bound on the sup norm of the order of $T_N^{d-1}$ seen previously, we have

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

We finally treat $B_{\mathrm{a}}$ using Assumption 2.1 (iii) with the condition that $f_\beta^- \in \mathrm{W}_q^s(\mathbb{S}^{d-1})$:

$$\|B_{\mathrm{a}}\|_q \le C T_N^{-s}.$$

We now choose $T_N$ to balance the bounds for the approximation bias $B_{\mathrm{a}}$ and the stochastic fluctuation $S_{\mathrm{e}}$ of the infeasible estimator $\overline{f}_{\beta,T}^{-}$. This can be achieved by setting

$$(\log N)^{r+(1/2-1/q)\mathbb{I}\{q\ge 2\}} N^{-1/2} T_N^{(2d-1)/2} \asymp T_N^{-s}.$$

Solve this to obtain (9.13) with $\gamma = 1/(2s+2d-1)$. For this choice of $T_N$ both terms are of the order

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

which is the desired rate of convergence. It is easy to check that Assumption 4.1 implies that

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

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

This proves the $\mathrm{L}^q$ convergence result.

In order to prove the strong uniform consistency, noticing that the bias terms $B_{\mathrm{t}}$ and $B_{\mathrm{a}}$ are not stochastic and bounded after proper scaling, we just have to focus on $S_{\mathrm{p}}$ and $S_{\mathrm{e}}$. Concerning $S_{\mathrm{p}}$, proceed as before and note that taking $M$ large enough so that $C_1 - C_2M^2 < -1$ implies summability of the left hand side in (9.23). We conclude from the first Borel-Cantelli lemma that the probability that the events occur infinitely often is zero thus with probability one

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

The term $T_2$ is non-stochastic and its treatment in our previous analysis remains valid, therefore we can use the same non-stochastic upper bound. We then use Assumption 4.1 (iii) instead of Assumption 4.1 (ii) to show that almost sure uniform boundedness of $S_{\mathrm{p}}$ after proper rescaling. The treatment of $S_{\mathrm{e}}$ 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,

$$\frac{\mathbb{E}\left[\left|Z_N(b) - \mathbb{E}\left[Z_N(b)\right]\right|^{2+\delta}\right]}{N^{\delta/2}\left(\mathrm{var}\left(Z_N(b)\right)\right)^{1+\delta/2}} \to 0 \tag{9.24}$$

(see, e.g. Billingsley, 1995). We start from deriving a lower bound on $\mathrm{var}\left(Z_N(b)\right)$. Since $\mathbb{E}[Z_N(b)]$ converges to $f_\beta^-(b)$, it is enough to obtain a lower bound on

$$\begin{aligned}
&\mathbb{E}[Z_{N,1}^2](b)\\
&= 4\int_{H^+}\left(\sum_{p=0}^{T_N-1}\chi(2p+1,2T_N)\frac{q_{2p+1,d}(z,b)}{\max\left(f_X(z),(\log N)^{-r}\right)\lambda(2p+1,d)}\right)^2 f_X(z)d\sigma(z)\\
&= 4\int_{H^+}\left(\sum_{p=0}^{T_N-1}\chi(2p+1,2T_N)\frac{q_{2p+1,d}(z,b)}{\lambda(2p+1,d)}\right)^2\left(\frac{1}{f_X(z)}\mathbb{I}\{f_X\ge(\log N)^{-r}\} + f_X(z)(\log N)^{2r}\mathbb{I}\{f_X<(\log N)^{-r}\}\right)d\sigma(z)\\
&\ge 4\frac{1}{\|f_X\|_\infty}\int_{H^+}\left(\sum_{p=0}^{T_N-1}\chi(2p+1,2T_N)\frac{q_{2p+1,d}(z,b)}{\lambda(2p+1,d)}\right)^2 d\sigma(z)\\
&\qquad - 4\frac{1}{\|f_X\|_\infty}\int_{\{0<f_X<(\log N)^{-r}\}}\left(\sum_{p=0}^{T_N-1}\chi(2p+1,2T_N)\frac{q_{2p+1,d}(z,b)}{\lambda(2p+1,d)}\right)^2 d\sigma(z)\\
&\ge 4\frac{1}{\|f_X\|_\infty}\sum_{p=0}^{T_N-1}\chi(2p+1,2T_N)^2\int_{H^+}\frac{q_{2p+1,d}(z,b)^2}{\lambda(2p+1,d)^2}d\sigma(z)\\
&\qquad - 4\frac{1}{\|f_X\|_\infty}\int_{\{0<f_X<(\log N)^{-r}\}}\left(\sum_{p=0}^{T_N-1}\chi(2p+1,2T_N)\frac{q_{2p+1,d}(z,b)}{\lambda(2p+1,d)}\right)^2 d\sigma(z).
\end{aligned}$$

Using (9.5) and Lemma 9.2, we see that there exists a constant $C$ such that

$$\left\|\sum_{p=0}^{T_N-1}\chi(2p+1,2T_N)\frac{q_{2p+1,d}(z,\star)}{\lambda(2p+1,d)}\right\|_\infty \le CT^{3d/2-1},$$

therefore using Proposition 2.2 we obtain

$$\mathbb{E}[Z_{N,1}^2](b) \geq \frac{4}{\|f_X\|_\infty} \sum_{p=0}^{T_N-1} \chi(2p+1, 2T_N)^2 \int_{H^+} \frac{q_{2p+1,d}(z,b)^2}{\lambda(2p+1,d)^2} d\sigma(z) - CT_N^{3d-2}\sigma\left(0 < f_X < (\log N)^{-r}\right).$$

Using Assumption 2.1 (iv), the first term on the right hand side can be bounded from below by

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

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

$$\mathbb{E}[Z_{N,1}^2](b) \geq CT_N^{2d-1}. \tag{9.25}$$

We now derive an upper bound of $\mathbb{E}\left[|Z_N(b)|^{2+\delta}\right]$ using Theorem 2.3 and interpolation between $\mathrm{L}^\infty(\mathbb{S}^{d-1})$ and $\mathrm{L}^1(\mathbb{S}^{d-1})$ norms of the kernels using the Hölder inequality:

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

By this and (9.25) an upper bound for the ratio appearing in (9.24) is given by

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

Therefore the Lyapounov condition is satisfied if (5.5) holds, and it follows that $N^{1/2}s_N^{-1}(b)S_{\mathrm{e}} \xrightarrow{d} N(0,1)$.

We now need to prove that the remaining terms $S_{\mathrm{p}}$, $B_{\mathrm{t}}$ and $B_{\mathrm{a}}$, multiplied by $N^{1/2}s_N^{-1}$, are $o_p(1)$. The term $S_{\mathrm{p}}$ is treated in a similar manner as in the proof of Theorem 5.1.

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

Using the Markov inequality, the empirical average in the parenthesis is of the stochastic order of

$$(\log N)^r \left\|\mathcal{H}^{-1}\left(K_{2T_N}^-(\star,\cdot)\right)\right\|_1.$$

But

$$(\log N)^r \left\| \mathcal{H}^{-1}\left(K_{2T_N}^{-}(\star,\cdot)\right)\right\|_1 \le B(d,1)T_N^{d/2}(\log N)^r \left\|K_{2T_N}^{-}(\star,\cdot)\right\|_1$$
$$\le B(d,1)T_N^{d/2}(\log N)^r \left\|K_{2T_N}(\star,\cdot)\right\|_1$$

where the first inequality follows from Theorem 2.3 and the second is obtained using the definition of the odd part and the triangle inequality. Note that the term $\|K_{2T_N}(\star,\cdot)\|_1$ in the last line does not depend on $\cdot$ and is uniformly bounded. By the lower bound (9.25) it is enough to show $N^{1/2}B(d,1)T_N^{-(d-1/2)}|S_{\mathrm{p}}(b)| = o_p(1)$. From the inequality above,

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

Its right hand side is of $o_p(1)$ if

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

which is met under (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 $\|B_{\mathrm{t}}\|_\infty$ and we deduce that

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

when condition (5.7) is satisfied. Finally, $N^{1/2}T_N^{-(d-1/2)}\|B_{\mathrm{a}}\|_\infty = o(1)$ if condition (5.6) is satisfied. We conclude that the asymptotic normality holds for $b$ such that $f_\beta(b) > 0$. The factor 4 in the variance comes from the fact that $\hat{f}_\beta = 2\hat{f}_\beta^{-}\mathbb{I}$. □

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

[14] 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.

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

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

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

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

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

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

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

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

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

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

[25] 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.

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

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

[28] Hesse, K., Mhaskar, H. N. and Sloan, I. H. (2007): "Quadrature in Besov Spaces on the Euclidian Sphere". *Journal of Complexity*, 23, 528-552.

[29] 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.

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

[31] 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.

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

[33] 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.

[34] 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.

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

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

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

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

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

[40] 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