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 β 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

Date: This Version: August 10, 2012.
Keywords: Inverse problems, Discrete Choice Models.

We thank Whitney Newey and two anonymous referees for comments that greatly improved this paper. We also thank seminar participants at Chicago, CREST, Harvard/MIT, the Henri Poincaré Institute, Hitotsubashi, LSE, Mannheim, Minnesota, Northwestern, NYU, Paris 6, Princeton, Rochester, Simon Fraser, Tilburg, Toulouse 1 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 and 2009 CIRM Rencontres de Statistiques Mathématiques 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, SES-0551271 and SES-0851759. Gautier is grateful for support from the Cowles Foundation as this research was initiated during his visit as a postdoctoral associate.

unobserved. The vectors $\tilde{X}$ and β correspond to observed and unobserved heterogeneity across agents, respectively. Note that the first element of β in this formulation absorbs the usual scalar stochastic shock term as well as a constant in 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. *β is independent of $\tilde{X}$,*

Section 5.3 considers ways to relax this assumption. Under (1.1) and Assumption 1.1, the choice probability function is given by

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

Discrete choice models with random coefficients 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 β 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 β. They present sufficient conditions for identification and prove the consistency of the NPMLE. The NPMLE requires high dimensional numerical maximization and can be computationally intensive even for a moderate sample size. Berry and Haile (2008) explore nonparametric identification problems in 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 β 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 β is not identified in the binary choice model, we normalize it so that β is a vector of Euclidean norm 1 in $\mathbb{R}^d$. The vector β then belongs to the $d-1$ dimensional sphere $\mathbb{S}^{d-1}$. This is not a restriction as long as the probability that β is equal to 0 is 0. Also, since only the angle between X and β matters 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_β of β with respect to the uniform spherical measure σ 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_β from r. Even under a set of assumptions that guarantees identification, however, the inverse of $\mathcal{H}$ is not a continuous mapping, making the problem ill-posed. 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 A.1.1 of Supplemental Appendix. 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 identifiability 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. Section 2 provides a practical guide for our procedure, which is easy to implement. Section 3 deals with identification while introducing basic notions used throughout the paper. We derive the convergence rates of the estimators in all the L^q spaces for $q \in [1,\infty]$ and also prove a pointwise CLT in Section 4. Some extensions, such as estimation of marginals, treatments of models with non-random coefficients, and the case with endogenous regressors are presented in Section 5. Simulation results are reported in Section 6. Section 7 concludes. Supplemental Appendix presents analysis of a toy model, technical tools used in the main text, estimators for choice probabilities that are used to construct our density estimators, and the proofs of the main results.

2. A Brief Guide for Practical Implementation

This section presents our basic estimation procedure when a random sample $\{(y_i, \tilde{x}_i)\}$ generated from the model (1.1) is available. As noted in Section 1, normalize covariates data and define $x_i = (1, \tilde{x}_i')/\|(1, \tilde{x}_i')\| \in \mathbb{S}^{d-1}, i = 1, ..., N$. To estimate the joint density of the random vector β, use the following formula:

$$\hat{f}_\beta(b) = \max\left(\frac{2}{|\mathbb{S}^{d-1}|}\sum_{p=0}^{T_N-1}\frac{\chi(2p+1,2T_N)h(2p+1,d)}{\lambda(2p+1,d)C_{2p+1}^{\nu(d)}(1)}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{(2y_i-1)C_{2p+1}^{\nu(d)}(x_i'b)}{\max\left(\hat{f}_X(x_i), m_N\right)}\right), 0\right). \tag{2.1}$$

The factors $|\mathbb{S}^{d-1}|$, χ, h and λ are constants that do not depend on data and trivial to compute. The surface area $\left|\mathbb{S}^{d-1}\right|$ of $\mathbb{S}^{d-1}$ is given by $\left|\mathbb{S}^{d-1}\right| = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ where Γ denotes the Gamma function.

The constants h, ν and λ are obtained via the numerical formulas $h(n,d) = \frac{(2n+d-2)(n+d-2)!}{n!(d-2)!(n+d-2)}$, $\nu(d) = (d-2)/2$ and $\lambda(2p+1,d) = \frac{(-1)^p|\mathbb{S}^{d-2}|1\cdot 3\cdots(2p-1)}{(d-1)(d+1)\cdots(d+2p-1)}$, respectively. The function χ is defined on $\mathbb{N}\times\mathbb{N}$ and used for smoothing. This is to be chosen by the user: see Proposition A.3 as well as the numerical example reported in Section 6 for examples of χ. The truncation parameter T_N needs to be chosen so that it grows with the sample size with a sufficiently slow rate. The trimming factor m_N is also user-defined, and it is chosen so that it goes to zero as the sample size increases. The notation $C_n^{\nu}(\cdot)$ signifies the Gegenbauer polynomial[1]; They, for example, correspond to the Chebychev polynomials of the first kind in the case of one random slope (i.e. the case with $d=2$)[2]. The only remaining factor which needs to be calculated in the above formula is the nonparametric density estimator $\hat{f}_X$ for f_X on $\mathbb{S}^{d-1}$. For example, the following nonparametric estimator can be used:

$$\hat{f}_X(x) = \max\left(\frac{1}{|\mathbb{S}^{d-1}|}\sum_{n=0}^{T'_N}\frac{\chi(n,T'_N)h(n,d)}{C_n^{\nu(d)}(1)}\left(\frac{1}{N}\sum_{i=1}^{N}C_n^{\nu(d)}(x_i'x)\right),0\right) \tag{2.2}$$

where T'_N is an another truncation parameter, playing a role similar to T_N.

Our estimator $\hat{f}_\beta$ requires neither numerical integration nor optimization. This is a clear advantage over existing estimators for random coefficient binary choice models, including many parametric estimators. This is our main proposal, on which the rest of the paper focuses. In Section 4 we explain how the formula (2.1) is derived, and investigate its asymptotic properties.

3. Identification Analysis

In this section we address the following two questions:

(Q1) Under what conditions is f_β identified?

(Q2) Does the random coefficients model impose restrictions?

---

[1]The Gegenbauer polynomials are given by

$$C_n^{\nu}(t) = \sum_{l=0}^{[n/2]}\frac{(-1)^l(\nu)_{n-l}}{l!(n-2l)!}(2t)^{n-2l}, \quad \nu > -1/2, n\in\mathbb{N}$$

where $(a)_0 = 1$ and for n in $\mathbb{N}\setminus\{0\}$, $(a)_n = a(a+1)\cdots(a+n-1) = \Gamma(a+n)/\Gamma(a)$. See Section A.1.2 for further properties of the Gegenbauer polynomials.

[2]When $d=2$, the following relations can be used in (2.1) and (2.2)

$$\forall p\geq 0,\ \frac{1}{|\mathbb{S}^{d-1}|}\frac{h(2p+1,2)C_{2p+1}^0(x_i'b)}{\lambda(2p+1,2)C_{2p+1}^0(1)} = \frac{(-1)^p(2p+1)}{\pi}\cos\left((2p+1)\arccos(x_i'b)\right),$$

$$\forall n\geq 0,\ \frac{1}{|\mathbb{S}^{d-1}|}\frac{h(n,2)C_n^0(x_i'b)}{C_n^0(1)} = \frac{1}{\pi}\cos\left(n\arccos(x_i'b)\right).$$

To answer these questions it is useful to introduce the notion of the odd and even part of a function defined on the sphere.

Definition 3.1. 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 $\mathbb{S}^{d-1}$.

Let us start with the question (Q1). As noted in Section A.1.4, operating $\mathcal{H}$ reduces the even part of a function to a constant 1 and therefore it is impossible to recover f_β^+ 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_β^-; then (Step 2) Show that f_β is uniquely determined from f_β^- 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 5 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 σ 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_β^-. The problem at hand calls for solving $r = \mathcal{H}f_\beta = \frac{1}{2} + \mathcal{H}f_\beta^-$ for f_β^-, and the inversion formula derived in (4.1) 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_β 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_β^- in terms of $R(x), x \in \mathbb{S}^{d-1}$, which shows the identifiability of f_β^- 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_β^-.

Now we turn to Step 2 in our identification argument. Obviously f_β^- does not uniquely determine f_β without further assumptions. This is a fundamental identification problem in our model. We need to identify f_β 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_β 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 β is a subset of some hemisphere.*

As noted by Ichimura and Thompson (1998), Assumption 3.2 does not seem too stringent in many economic applications. It is often reasonable to assume that an element of the random coefficients vector, such as a price coefficient, has a known sign. If the j-th element of β has a known sign (and positive), then Assumption 3.2 holds with c being a unit vector with its j-th element being 1. This is a case in which the location of the hemisphere in Assumption 3.2 is known *a priori*, though the knowledge about its location is not necessary for identification. Assumption 3.2 implies the following mapping from f_β^- to f_β developed in (A.24):

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

This is useful because it shows that Assumption 3.2 guarantees identification if f_β^- 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_β that is guaranteed to be non-negative.

Remark 3.1. Assumption 3.2 is testable since it imposes restrictions on f_β^-, 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_β^- 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 subsequent result, Proposition 3.1, answers question (Q2), and a proof is given in Supplemental Appendix.

Notation. We use the notation $\mathrm{L}^2(\mathbb{S}^{d-1})$ for the space of square integrable complex valued functions equipped with the hermitian product $(f,g)_{\mathrm{L}^2(\mathbb{S}^{d-1})} = \int_{\mathbb{S}^{d-1}} f(x)\overline{g}(x)d\sigma(x)$, and more generally use $\mathrm{L}^p(\mathbb{S}^{d-1})$ for $p \in [1,\infty]$ the Banach space of p-integrable functions and $\|\cdot\|_p$ the corresponding norm. We also use the notation $\mathrm{W}_p^s(\mathbb{S}^{d-1})$ (and $\mathrm{H}^s(\mathbb{S}^{d-1})$ for $p=2$) to signify the corresponding Sobolev spaces with norm $\|\cdot\|_{p,s}$ defined as

$$\|f\|_{p,s} = \|f\|_p + \left\|\left(-\Delta^S\right)^{s/2} f\right\|_p \tag{3.6}$$

where Δ^S denotes the Laplacian on the sphere $\mathbb{S}^{d-1}$: See Section A.1.3 for further discussions.

Proposition 3.1. *A $[0,1]$-valued function r is compatible with the random coefficients model (1.1) with f_β 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 coefficients 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 $-x$.

4. Nonparametric Estimation of f_β

4.1. **Derivation of the closed form estimation formula.** This section discusses how the closed form estimation formula (2.1) is derived. Suppose an odd function f^- defined on $\mathbb{S}^{d-1}$ satisfies an integral equation $f^- = Hg$ with g square integrable with respect to the spherical measure. In Section A.1.4 we show that the solution to this equation is given by:

$$\mathcal{H}^{-1}(f^-)(y) = \sum_{p=0}^{\infty} \frac{1}{\lambda(2p+1,d)} \int_{\mathbb{S}^{d-1}} q_{2p+1,d}(x,y) f^-(x) d\sigma(x) \tag{4.1}$$

where expressions for λ and q are provided in Proposition A.4 and Theorem A.1, respectively. If an appropriate estimator $\hat{R}^-$ of R^- is available, an application of the inversion formula (4.1) to (3.4) suggests the following estimator for f_β^-:

$$\begin{aligned} \hat{f}_\beta^- &= \mathcal{H}^{-1}\left(\hat{R}^-\right) \\ &= \sum_{p=0}^{\infty} \frac{1}{\lambda(2p+1,d)} \int_{\mathbb{S}^{d-1}} q_{2p+1,d}(\cdot,x) \hat{R}^-(x) d\sigma(x). \end{aligned} \tag{4.2}$$

Then use the mapping (3.5) to define

$$\hat{f}_\beta(b) = 2\hat{f}_\beta^-(b)\mathbb{I}\left\{\hat{f}_\beta^-(b) > 0\right\} \tag{4.3}$$

as an estimator for f_β.

We use the following notation in the rest of 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 $M^{-1}b_n \le a_n \le M b_n$ for every positive n.

Proposition A.6 implies that if $\hat{f}_\beta^- - f_\beta^- \in \mathrm{H}^s(\mathbb{S}^{d-1})$ then $\hat{R}^- - R^- \in \mathrm{H}^\sigma(\mathbb{S}^{d-1})$, $\sigma = s + \frac{d}{2}$ and for $v \in [0, s]$,

$$\|\hat{f}_\beta^- - f_\beta^-\|_{2,v} \asymp \|\hat{R}^- - R^-\|_{2,v+d/2}. \tag{4.4}$$

As discussed earlier, the estimation of f_β 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 absolute values of the eigenvalues of $\mathcal{H}$ (c.f. Proposition A.4) $\lambda(n,d), n=2p+1, p\in\mathbb{N}$ converges to zero as p grows, as shown in (A.27). 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_β 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 (4.4), 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 (4.2) 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 (4.2), for example, by truncating the series. This results in a general estimator that can be written in the following two equivalent forms

$$\begin{aligned}\hat{f}_\beta^- &= \mathcal{H}^{-1}\left(P_{\tilde{T}_N}\hat{R}^-\right)\\ &= \sum_{p=0}^{T_N}\frac{1}{\lambda(2p+1,d)}\int_{\mathbb{S}^{d-1}} q_{2p+1,d}(\cdot,x)\hat{R}^-(x)d\sigma(x)\end{aligned} \tag{4.5}$$

for suitably chosen $\tilde{T}_N$ that goes to infinity with N and $P_{\tilde{T}_N}$ defined in (A.20). The sequence $\mathcal{H}^{-1}\circ P_{\tilde{T}_N}, N=1,2,...$ can be interpreted as regularized inverses of $\mathcal{H}$, with the spectral cut-off method often used in statistical inverse problems.

We now discuss how to obtain $\hat{R}^-$ in the calculation of (4.5). The following choice is particularly convenient:

$$\hat{R}^-(x) = \frac{1}{N}\sum_{i=1}^N \frac{(2y_i-1)K_{2T_N}^-(x_i,x)}{\max\left(\hat{f}_X(x_i), m_N\right)} \tag{4.6}$$

where m_N is a trimming factor going to 0 with the sample size, $K^-(x_i,\cdot)$ denotes the odd part (of the second argument) of the kernel function $K(x_i,\cdot)$ defined in (A.23) and $\hat{f}_X$ is a nonparametric density estimator for f_X. See Section A.1.5 of Supplemental Appendix for the derivation of the above formula. Various nonparametric estimators for f_X can be used in (4.6), 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).

Note also that Baldi et al. (2009) develops an adaptive density estimator on the sphere using needlet thresholding. In the simulation experiment we use

$$\hat{f}_X(x) = \max\left(\frac{1}{N}\sum_{i=1}^{N} K_{T'_N}(x_i, x), 0\right) \tag{4.7}$$

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 (A.23) satisfying Assumption A.1. Note that its rate of convergence in sup-norm can be obtained in the same manner as the proof of Theorem 4.1. 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 χ in (A.23)), 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.

Using (4.5) and (4.6) with $\tilde{T}_N = 2T_N$, define

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

Computing $\hat{f}_\beta^-$ is straightforward. First, note that the estimator (4.6) for R^- resides in a finite dimensional space $\bigoplus_{p=0}^{T_N} H^{2p+1,d}$, therefore $P_{2T_N}\hat{R}^- = \hat{R}^-$ holds. Consequently, unlike in (4.5) 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

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

which, in turn, can be seen by the definition of K_T in (A.23), the fact that the integral operators with q as kernels are projections and (A.16). Thus

$$\hat{f}_\beta^-(b) = \frac{1}{N}\sum_{i=1}^{N}\frac{2y_i - 1}{\max\left(\hat{f}_X(x_i), m_N\right)}\sum_{p=0}^{T_N - 1}\frac{\chi(2p+1, 2T_N)}{\lambda(2p+1, d)} q_{2p+1,d}(x_i, b).$$

Using (4.3) and the Addition formula (Theorem A.1), we arrive at an estimator for f_β with the following explicit form:

$$\hat{f}_\beta(b) = 2\hat{f}_\beta^-(b)\mathbb{I}\{\hat{f}_\beta^-(b) > 0\}, \tag{4.8}$$

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

This is equivalent to the formula (2.1) previously presented in Section 2. Likewise, using the definition of the smoothing kernel (A.23) and the Addition Theorem in the above definition (4.7) of $\hat{f}_X$, we obtain the formula (2.2) as well.

4.2. **Rates of Convergence in $\mathrm{L}^q(\mathbb{S}^{d-1})$-norms.** Now we analyze the rate of our estimator $\hat{f}_\beta$. The following assumption is weak and reasonable.

Assumption 4.1. $f_X \in \mathrm{L}^\infty$.

The proofs of the following theorems and corollaries in the rest of this section are given in Section A.1.6 of Supplemental Appendix.

Theorem 4.1 (Upper bounds in $\mathrm{L}^q(\mathbb{S}^{d-1})$)**.** *Suppose Assumptions A.1, 3.1 and 4.1 hold, and choose T_N that does not grow more than polynomially fast in N. If f_β^- belongs to $\mathrm{W}_q^s(\mathbb{S}^{d-1})$ with q in $[1,\infty]$ and $s>0$, and*

$$\max_{i=1,\dots,N}\left|f_X(x_i)-\hat{f}_X(x_i)\right| = O_p\left(m_N\right), \tag{4.9}$$

then, for any $1\le r\le q$,

$$\begin{aligned}\left\|\hat{f}_\beta - f_\beta\right\|_q = O_p\Big(&m_N^{-1}N^{-1/2}T_N^{(2d-1)/2}(\log N)^{(1/2-1/q)\mathbb{I}\{q\ge 2\}}\\ &+T_N^{-s}+T_N^{d/2}m_N^{-2}\max_{i=1,\dots,N}\left|f_X(x_i)-\hat{f}_X(x_i)\right|\\ &T_n^{d/2+(d-1)(1-1/r)}\sigma\left(f_X<m_N\right)^{1/q-1/r+1}\Big).\end{aligned} \tag{4.10}$$

When there exists $m>0$ such that $f_X\ge m\ \sigma$ a.e. on H^+, the following holds for the estimator without the trimming factor (i.e. $m_N=0$) when the estimator $\hat{f}_X$ which is consistent in sup norm:

$$\begin{aligned}\left\|\hat{f}_\beta - f_\beta\right\|_q = O_p\Big(&N^{-1/2}T_N^{(2d-1)/2}(\log N)^{(1/2-1/q)\mathbb{I}\{q\ge 2\}}\\ &+T_N^{-s}+T_N^{d/2}\max_{i=1,\dots,N}\left|f_X(x_i)-\hat{f}_X(x_i)\right|\Big).\end{aligned}$$

The first term in (4.10) is the stochastic error, the second term is the approximation bias, the third the plug-in error and the fourth the trimming bias. Note that Theorem 4.1 imposes the mild assumption (4.9); otherwise, we need to replace $T_N^{d/2} m_N^{-2} \max_{i=1,\dots,N} \left| f_X(x_i) - \hat{f}_X(x_i) \right|$ in (4.10) with $T_N^{d/2} m_N^{-2} \max_{i=1,\dots,N} \left| f_X(x_i) - \hat{f}_X(x_i) \right| \left(1 + (\log N)^{(1/2-1/q)\mathbb{I}\{q \geq 2\}} N^{-1/2} T_N^{(d-1)/2}\right)$. Since

$$\max_{i=1,\dots,N} \left| f_X(x_i) - \hat{f}_X(x_i) \right| \leq \left| f_X - \hat{f}_X \right|_\infty,$$

this term can be made of order $O_P\left(\left(\frac{N}{\log N}\right)^{-v/(2v+d-1)}\right)$ when $f_X \in \mathrm{W}_\infty^v$ with a suitably chosen parameter T_N' if we take (4.7) as an estimator. The proof of the latter statement is classical and can be obtained simplifying the proof of Theorem 4.1 and Corollary 4.1. Equation (4.10) yields that, for proper choices of m_N going to zero and T_N to infinity, $\hat{f}_\beta$ is consistent given that f_X has some smoothness in the Sobolev scales.

Though the additional condition of f_X being bounded away from 0 in the last statement of Theorem 4.1 is convenient, it is restrictive. To see this, consider the $d = 2$ case. In polar coordinates, $f_X(\cos(\theta), \sin(\theta)) = f_{\tilde{X}}(\tan(\theta))(1 + \tan^2(\theta)$, thus, assuming $f_X \geq m$ on H^+, which does not require trimming, yields

$$\forall x \in \mathbb{R}, f_{\tilde{X}}(x) \geq \frac{m}{1+x^2}.$$

It implies that $\tilde{X}$ has tails larger than Cauchy tails and all moments are infinite. The introduction of the trimming factor m_N allows us to relax the assumption $f_X \geq m$, though it introduces bias. As is clear from (4.10), the condition for the trimming bias to go to zero with N depends both on T_N and m_N. The quantity $\sigma(f_X < m_N)$ should decay to zero with N sufficiently fast. We can check, for example, that when $\tilde{X}$ is standard Gaussian then $\sigma(f_X < m_N) = O\left((-\log m_N)^{-1/2}\right)$, when it is Laplace then $\sigma(f_X < m_N) = O\left((-\log m_N)^{-1}\right)$ and when $f_{\tilde{X}}$ is proportional to $(1+x^2)^{-k}$ with $k > 1$ we obtain that $\sigma(f_X < m_N) = O\left(m_N^{1/(2(k-1))}\right)$. In all these cases, it is possible to adjust adequately T_N and m_N and to obtain rates of convergence. The upper bound on the rates become slower as the tail of f_X becomes thinner.

Nonparametric estimation of the regression function with random degenerate design, in the sense that the density of regressors can be low on its support, is a difficult issue. It has been studied for the pointwise risk in Hall et al. (1997), Gaïffas (2005), Gaïffas (2009) and Guerre (1999). Extension to inverse problems setting is a widely open problem. We tackle this problem for our specific inverse problem. Future research includes the study of lower bounds from the minimax point of view that account for the degeneracy of the design.

Let us now return to the general case of $d-1$ regressors. The assumptions below allow us to obtain rates that differ slightly from the rates that we would obtain in the ideal case where $f_X \geq m\ \sigma$ *a.e.* for positive m on H^+.

Assumption 4.2. *Suppose for q in $[1,\infty]$, there exist positive τ and r_X such that*

(i) $\sigma(f_X < h) = O(h^\tau)$ *and* $f_X \in \mathrm{L}^\infty$,

and either

(ii)

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

or,

(iii) for some constant C,

$$\overline{\lim}_{N\to\infty}\left(\frac{N}{(\log N)^{(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{r_X}\max_{i=1,\dots,N}\left|f_X(x_i) - \hat{f}_X(x_i)\right| \leq C \quad a.s.$$

holds.

As seen before, Assumption 4.2 (ii) or (iii) are very mild. (i) holds for a reasonable class of distributions for f_X. In the above example where $f_{\tilde{X}}$ is proportional to $(1+x^2)^{-k}$ with $k>1$, we have the relation $\tau = \rho/(2(k-1))$. This allows for a higher order moment to exist for a large k.

Corollary 4.1. *Assume that f_β^- belongs to $\mathrm{W}_q^s(\mathbb{S}^{d-1})$ with q in $[1,\infty]$ and $s>0$. Let assumptions A.1, 3.1, 4.1 and 4.2 (i) and (ii) hold, and take*

$$m_N \asymp \left(\frac{N}{(\log N)^{(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{-\rho}, \quad T_N \asymp \left(\frac{N}{(\log N)^{(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{\gamma(\rho)}$$

where ρ yields a maximum γ of

$$\gamma(\rho) = \min\left(\frac{1-2\rho}{2s+2d-1}, \frac{2\rho\tau}{2s+d+2(d-1)(1-1/q)}, \frac{2r_X - 4\rho}{2s+d}, \frac{1}{d-1}\right).$$

We then have

$$\left\|\hat{f}_\beta - f_\beta\right\|_q = O_p\left(\left(\frac{N}{(\log N)^{(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{-\gamma s}\right). \tag{4.11}$$

Moreover, if, instead of Assumption 4.2 (ii), Assumption 4.2 (iii) holds with $q=\infty$, then there exists a constant C such that

$$\overline{\lim}_{N\to\infty}\left(\frac{N}{\log N}\right)^{\gamma s}\left\|\hat{f}_\beta - f_\beta\right\|_\infty \leq C \quad a.s. \tag{4.12}$$

The rate γs in Corollary 4.1 accounts for the dimension $d-1$, the degree of smoothing $d/2$ of the operator and features of the density of the covariates (i.e. its smoothness and tail behavior).

We now make stronger assumptions on f_X and its estimate that yield, up to a logarithmic term, the convergence rate $N^{-\frac{s}{2s+2d-1}}$. We need to be able to trim the estimate of f_X with a term which is logarithmic in N: $m_N = (\log N)^{-\rho}$ for some positive ρ.

Assumption 4.3. *Suppose for q in $[1,\infty]$, and positive r_σ and r_X,*

(i) $\sigma(f_X < (\log N)^{-\rho}) = O\left(\left(\frac{N}{(\log N)^{2\rho+(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{-r_\sigma}\right)$,

and either

(ii)

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

or,

(iii) for some constant C,

$$\overline{\lim}_{N\to\infty}(\log N)^{2\rho}\left(\frac{N}{(\log N)^{2\rho+(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{r_X}\max_{i=1,\dots,N}\left|f_X(x_i)-\hat{f}_X(x_i)\right| \leq C \quad a.s.$$

Corollary 4.2. *Assume that f_β^- belongs to $\mathrm{W}_q^s(\mathbb{S}^{d-1})$ with q in $[1,\infty]$ and $s>0$. Let assumptions A.1, 3.1, 4.1 and 4.3 (i)-(ii), hold, and take*

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

where

$$\gamma = \min\left(\frac{1}{2s+2d-1}, \frac{2r_\sigma}{2s+d+2(d-1)(1-1/q)}, \frac{2r_X}{2s+d}\right)$$

then we have

$$\left\|\hat{f}_\beta - f_\beta\right\|_q = O_p\left(\left(\frac{N}{(\log N)^{2\rho+(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{-\gamma s}\right). \tag{4.13}$$

Moreover, if, instead of Assumption 4.3 (ii), Assumption (iii) holds with $q=\infty$, then there exists a constant C such that

$$\overline{\lim}_{N\to\infty}\left(\frac{N}{(\log N)^{2\rho+1}}\right)^{\gamma s}\left\|\hat{f}_\beta - f_\beta\right\|_\infty \leq C \quad a.s. \tag{4.14}$$

When $f_X \in \mathrm{W}_\infty^{s+d/2+\epsilon}(\mathbb{S}^{d-1})$ for any positive ϵ then $\frac{2r_X}{2s+d} > \frac{1}{2s+2d-1}$ and γ in Corollary 4.2 is simply $\min\left(\frac{1}{2s+2d-1}, \frac{2r_\sigma}{2s+d+2(d-1)(1-1/q)}\right)$. Recall that the smoothness $s+d/2$ is related to the smoothness of R. Indeed, we have seen in Section 3 that $R\in \mathrm{W}_2^{s+d/2}(\mathbb{S}^{d-1})$ if and only if $f_\beta \in \mathrm{W}_2^s(\mathbb{S}^{d-1})$.

Consider now the most restrictive case where $f_X \geq m\ \sigma$ *a.e.*, then the estimator without the trimming factor (i.e. $m_N = 0$) satisfies the following:

Corollary 4.3. *Assume that f_β^- belongs to $\mathrm{W}_q^s(\mathbb{S}^{d-1})$ with q in $[1,\infty]$ and $s > 0$. Let assumptions A.1, 3.1 and 4.1 hold, and suppose, for positive r_X,*

$$\max_{i=1,\dots,N}\left|f_X(x_i) - \hat{f}_X(x_i)\right| = O_p\left(\left(\frac{N}{(\log N)^{(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{-r_X}\right). \tag{4.15}$$

Take

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

where

$$\gamma = \min\left(\frac{1}{2s+2d-1}, \frac{2r_X}{2s+d}\right)$$

then we have

$$\left\|\hat{f}_\beta - f_\beta\right\|_q = O_p\left(\left(\frac{N}{(\log N)^{(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{-\gamma s}\right). \tag{4.16}$$

Moreover, if we replace (4.15) by for some positive C

$$\left(\frac{N}{(\log N)^{(1-2/q)\mathbb{I}\{q\geq 2\}}}\right)^{r_X}\max_{i=1,\dots,N}\left|f_X(x_i) - \hat{f}_X(x_i)\right| \leq C \quad a.s. \tag{4.17}$$

then

$$\overline{\lim}_{N\to\infty}\left(\frac{N}{\log N}\right)^{\gamma s}\left\|\hat{f}_\beta - f_\beta\right\|_\infty \leq C \quad a.s. \tag{4.18}$$

When f_X belongs to $\mathrm{W}_\infty^{s-d/2+\epsilon}$, for arbitrary positive ϵ, $\gamma = \frac{1}{2s+2d-1}$ in Corollary 4.3, and we recover the L^2 convergence rate of $N^{\frac{s}{2s+2d-1}}$, the rate mentioned in Section 4.1. It is in accordance with the L^2 rate in Healy and Kim (1996) who study deconvolution on $\mathbb{S}^2$ for non-degenerate kernels. Kim and Koo (2000) prove that the rate in Healy and Kim (1996) is optimal in the minimax sense. Their statistical problem, 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 $\mathcal{H}$. Hoderlein et al. (2010) study a linear model of the form $W = X'\beta$ where β is a d-vector of random coefficients. They obtain a nonparametric random coefficients density estimator that has the L^2-rate $N^{-\frac{s}{2s+2d-1}}$ when $f_X \geq m\sigma$ *a.e.* for positive m[3] when f_X is assumed to be bounded from below and thus no trimming is required. They also consider trimming but the

[3]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.

approach is slightly different and rates of convergence are not given. Unlike the previous results, we cover L^q loss for all $q \in [1, \infty]$.

4.3. **Pointwise Asymptotic Normality.** This section discusses the asymptotic normality property of our estimator.

Theorem 4.2 (Asymptotic normality). *Suppose f_β^- belongs to $\mathrm{W}^s_\infty(\mathbb{S}^{d-1})$ with $s > 0$, and Assumptions A.1, 3.1 and 4.1 hold. If $\hat{f}_X$, f_X, m_N and T_N satisfy*

$$N^{1/2} T_N^{-(d-1)/2} m_N^{-2} \max_{i=1,\dots,N} \left| f_X(x_i) - \hat{f}_X(x_i) \right| = o_p(1), \tag{4.19}$$

$$N^{-1/2} T_N^{(d-1)/2} m_N^{-(1+\epsilon)} = o(1) \quad \text{for some } \epsilon > 0, \tag{4.20}$$

$$N^{1/2} T_N^{-\frac{2s+2d-1}{2}} = o(1), \tag{4.21}$$

$$N^{1/2} T_N^{(d-1)/2} \sigma\left(\{f_X < m_N\}\right) = o(1) \tag{4.22}$$

then

$$N^{\frac{1}{2}} s_N^{-1}(b) \left(\hat{f}_\beta(b) - f_\beta(b) \right) \xrightarrow{d} N(0, 1) \tag{4.23}$$

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

The standard error $s_N(b)$ is the standard deviation of

$$Z_N(b) = \frac{2}{|\mathbb{S}^{d-1}|} \sum_{p=0}^{T_N-1} \frac{\chi(2p+1, 2T_N) h(2p+1, d)}{\lambda(2p+1, d) C_{2p+1}^{\nu(d)}(1)} \left(\frac{(2Y-1) C_{2p+1}^{\nu(d)}(X'b)}{\max(f_X(X), m_N)} \right) \tag{4.24}$$

(see equation (4.8)), which can be estimated using an estimate $\hat{f}_X$ of f_X.

The next theorem is concerned with the restrictive case where the density of the covariates is bounded from below and hence the trimming factor m_N is set at zero.

Theorem 4.3 (Asymptotic normality when the density of the covariates is bounded from below). *Suppose f_β^- belongs to $\mathrm{W}^s_\infty(\mathbb{S}^{d-1})$ with $s > 0$, and Assumptions A.1, 3.1 and 4.1 hold. If $\hat{f}_X$, f_X and T_N satisfy*

$$N^{1/2} T_N^{-(d-1)/2} \max_{i=1,\dots,N} \left| f_X(x_i) - \hat{f}_X(x_i) \right| = o_p(1), \tag{4.25}$$

$$N^{-1/2} T_N^{(d-1)/2} = o(1) \tag{4.26}$$

$$N^{1/2} T_N^{-\frac{2s+2d-1}{2}} = o(1), \tag{4.27}$$

then

$$N^{\frac{1}{2}} s_N^{-1}(b)\left(\hat{f}_\beta(b)-f_\beta(b)\right) \xrightarrow{d} N(0,1) \tag{4.28}$$

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

A formula for Z_N for this case is obtained by replacing $\max(f_X(X),m_N)$ with $f_X(X)$ in (4.24). Note that $s_N^2(b)$ grows at the rate of T_N^{2d-1} in this case.

5. Discussion

5.1. **Estimation of Marginals.** In Section 3 we have provided an expression for the estimator of the full joint density of β, from which an estimator for a marginal density can be obtained. Let σ_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 $\overline{\overline{\beta}}$ which is a vector of dimension 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 β increases. In order to obtain a flat density for the marginals of the uniform joint distribution on the sphere it is enough to consider densities with respect to the dominating measure $\overline{\overline{P}}_*\underline{\sigma}_{d-1}$. 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 $\{1,\dots,d-1\}$,

$$\int_{\mathbb{S}^{d-1}} g(b)d\underline{\sigma}_{d-1}(b)=\int_{\mathbb{B}^k}\left[\int_{\mathbb{S}^{d-1-k}} g\left(\rho\left(\overline{\overline{b}}\right)u,\overline{\overline{b}}\right)d\underline{\sigma}_{d-1-k}(u)\right]d\overline{\overline{P}}_*\underline{\sigma}_{d-1}\left(\overline{\overline{b}}\right) \tag{5.1}$$

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

$$f_{\overline{\overline{\beta}}}\left(\overline{\overline{b}}\right)=|\mathbb{S}^{d-1}|\int_{\mathbb{S}^{d-1-k}} f_\beta\left(\rho\left(\overline{\overline{b}}\right)u,\overline{\overline{b}}\right)d\underline{\sigma}_{d-1-k}(u). \tag{5.2}$$

One can use deterministic methods to compute the integral (e.g., Narcowich et al. (2006) for quadrature methods on the sphere) or for example one may use a Monte-Carlo method, by forming

$$\hat{f}^{M}_{\overline{\overline{\beta}}}\left(\overline{\overline{b}}\right) = \frac{|\mathbb{S}^{d-1}|}{M}\sum_{j=1}^{M}\hat{f}_{\beta}\left(\rho\left(\overline{\overline{b}}\right)u_j, \overline{\overline{b}}\right) \tag{5.3}$$

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

5.2. **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.[4] First, the convergence rate of our estimator of the joint density of β slows down as the dimension d of β grows, which is a manifestation of the curse of dimensionality. Treating some coefficients as fixed parameters alleviates this problem. Second, our identification assumption in Section 3 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:

$$Y_i = \mathbb{I}\{\beta_{1i} + \beta_{2i}'X_{2i} + \alpha_1 Z_{1i} + \alpha_2' Z_{2i} \geq 0\} \tag{5.4}$$

where $\beta_1 \in \mathbb{R}$ and $\beta_2 \in \mathbb{R}^{d_X-1}$ are random coefficients, whereas the coefficients $\alpha_1 \in \mathbb{R}$ and $\alpha_2 \in \mathbb{R}^{d_Z-1}$ are nonrandom. The covariate vector $(Z_1, Z_2')'$ is in $\mathbb{R}^{d_Z}$, though the $(d_Z - 1)$-subvector Z_2 might have limited support: for example, it can be a vector of dummies. The covariate vector $(X_2', Z_1)'$ is assumed to be, among other things, continuously distributed. Normalizing the coefficients vector and the vector of covariates to be elements of the unit sphere works well for the development of our procedure, as we have seen in the previous sections. The model (5.4), 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 (5.4) is equivalent to:

$$Y = \mathbb{I}\{(\beta_1^*(Z_2), \alpha_1, \beta_2)(1, Z_1, X_2')' \geq 0\}$$

[4]Hoderlein et al. (2010) suggest a method to deal with non-random coefficients in their treatment of random coefficient linear regression models.

$$= \mathbb{I}\left\{\tau(Z_2)'W \geq 0\right\}.$$

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 (A.30) and (4.1) we obtain

$$f^-_{\tau(Z_2)|Z_2=z_2}(t) = \sum_{p=0}^{\infty} \frac{1}{\lambda(2p+1, d_X+1)} \mathbb{E}\left[\left. \frac{(2Y-1)q_{2p+1,d_X+1}(W,t)}{f_{W|Z_2=z_2}(W)} \right| Z_2 = z_2\right] \tag{5.5}$$

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

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

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

$$f\left(\left. \frac{(\beta_1^*(Z_2), \alpha_1, \beta_2)}{\|\beta_2\|} \right| Z_2 = z_2\right) \tag{5.6}$$

for all $z_2 \in \mathcal{Z}_2$ as well. If we are only interested in the joint distribution of β_2 under a suitable normalization, we can stop here. The presence of the term $\alpha_1 Z_1$ in (5.4) 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 (5.6) gives

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

from which we can, for example, get

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

Define a constant

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

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

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

is identified as well. Let

$$f\left(\frac{(\beta_2', \alpha_1, \alpha_2')'}{\|\beta_2\|}\right) \tag{5.7}$$

denote the joint density of all the coefficient (except for β_1, which corresponds to the conventional disturbance term in the original model (5.4), normalized by the length of β_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 (5.6), and $c\alpha_1$ and $c\alpha_2$ are identified already, therefore the desired joint density (5.7) is identified. Obviously (5.7) also determines the joint density of $(\beta_2', \alpha_1, \alpha_2')'$ under other suitable normalizations as well.

The density (5.5) is estimable: when Z_2 is discrete, one can use the estimator of Section 4 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 (5.6) can be then obtained numerically.

5.3. **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

$$Y = \mathbb{I}\{\beta_1 + \tilde{\beta}' X \geq 0\} \tag{5.8}$$

with

$$X = \Gamma Z + V \tag{5.9}$$

where V is a vector of reduced form residuals and Z is independent of (β, V). Note that Hoderlein et al. (2010) utilize a linear structure of the form (5.9) in estimating a random coefficient linear model. The equations (5.8) and (5.9) yield

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

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

6. 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 $X = (1, X_2, X_3)$ 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 (β_1, β_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 $\lambda = 0.5$. The sample size N is 500 for both models, and the number of Monte Carlo replications is 10,000. The new estimator (4.8) is implemented using the Riesz kernel with $s = 3$ and $l = 3$ (see Proposition A.3). The truncation parameter T_N is set at 3, and the trimming parameter ρ (see Assumption 4.3) is 2. It also requires a nonparametric estimator for f_X, and we use the projection estimator (4.7) based on the same Riesz kernel (i.e. $s = 3$, $l = 3$) and $T_N = 10$. Figures 1 and 2 present the surface plots of the true density (left panel) and the mean of our estimator (right panel), for each of the two specifications. The mean $\mathbb{E}[\hat{f}_\beta]$ is calculated as the empirical average of 10,000 Monte Carlo realizations of $\hat{f}_\beta$. Our estimator (4.8) is defined on $\mathbb{S}^2$ in this case, and we performed an appropriate transformation to plot it as a density on $\mathbb{R}^2$.

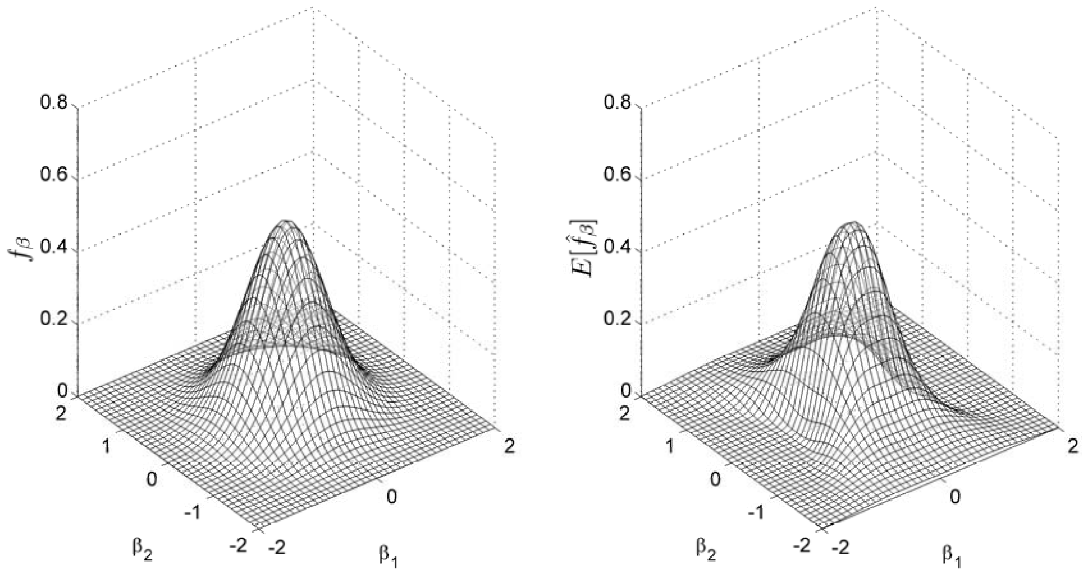

FIGURE 1. Simulation result: Model 1

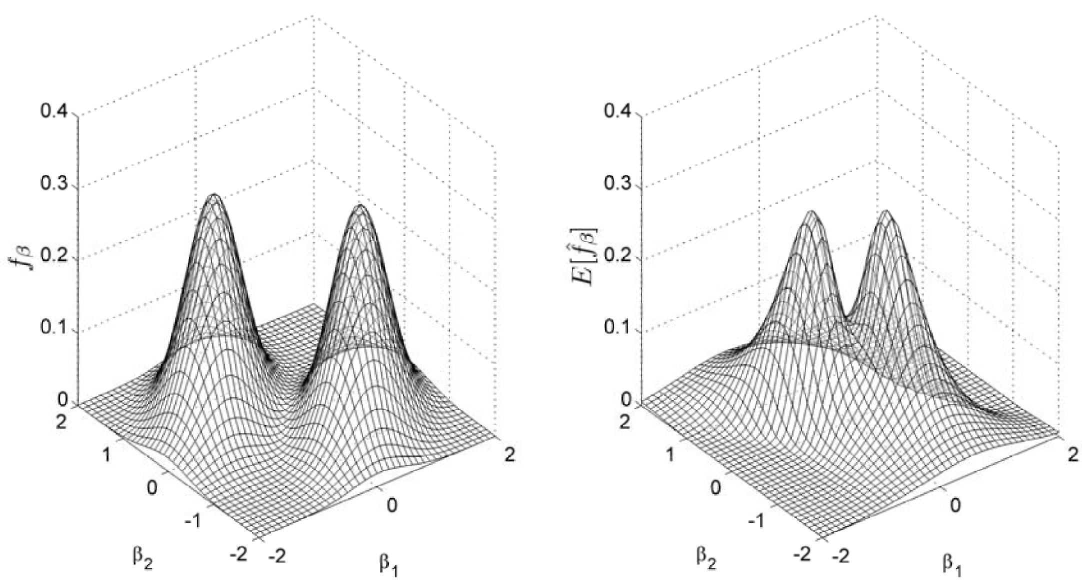

FIGURE 2. Simulation result: Model 2

In the case of model 1, with the reasonable sample size, the location of the peak of the density, as well as its shape, are successfully recovered by our procedure. For model 2, again, our procedure works well: the estimated surface plot nicely captures the shape 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.

7. 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 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 and practical interest to obtain an adaptive procedure for choosing the smoothing parameters of our estimator, though it is a task we defer to subsequent investigations.[5] With appropriate under-smoothing, the estimator is shown to be asymptotically normal, providing a theoretical basis for nonparametric statistical inference for the random coefficients distribution.

[5] Gautier and Le Pennec (2011) consider a needlet-based procedure and discuss its rate optimality in a minimax sense and adaptation.

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] Baldi, P., Kerkyacharian, G., Marinucci, D. and Picard, D. (2009): "Adaptive Density Estimation for Directional Data Using Needlets". *Annals of Statistics*, 37, 3362-3395.

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

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

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

[7] 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.

[8] Briech, R., Chintagunta, P. and Matzkin, R. (2007): "Nonparametric Discrete Choice Models with Unobserved Heterogeneity". *Journal of Business and Economic Statistics*, 28, 291307.

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

[10] 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.

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

[12] 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.

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

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

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

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

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

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

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

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

[21] Gaïffas, S. (2005): "Convergence Rates for Pointwise Curve Estimation with a Degenerate Design". *Mathematical Methods of Statistics*, 14, 1-27.

[22] Gaïffas, S. (2009): "Uniform Estimation of a Signal Based on Inhomogeneous Data". *Statistica Sinica*, 19, 427-447.

[23] Gautier, E., and Le Pennec (2011): "Adaptive estimation in nonparametric random coefficients binary choice model by needlet thresholding". Preprint arXiv:1106.3503.

[24] Guerre, E. (1999): "Efficient Random Rates for Nonparametric Regression Under Arbitrary Designs". Working Paper.

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

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

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

[28] Hall, P., Marron, J. S., Neumann, M. H., and Tetterington, D. M. (1997): "Curve estimation when the design density is low". *Annals of Statististics*, 25, 756770.

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

[30] Harding, M. C. and Hausman, J. (2007): "Using a Laplace Approximation to Estimate the Random Coefficients Logit Model by Non-linear Least Squares". *International Economic Review*, 48, 1311-1328.

[31] 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.

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

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

[34] 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.

[35] Hoderlein, S., Klemelä, J., and Mammen, E. (2010): "Analyzing the Random Coefficient Model Nonparametrically". *Econometric Theory*, 26, 804-837.

[36] 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.

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

[38] 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.

[39] 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.

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

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

[42] Narcowich, F., Petrushev, P., and Warda, J. (2006): "Decomposition of Besov and TriebelLizorkin spaces on the sphere". *Journal of Functional Analysis*, 238, 530-564.

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

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

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

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

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

E-mail address: eric.gautier@ensae-paristech.fr

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

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