Analyzing the Random Coefficient Model Nonparametrically*

Stefan Hoderlein[†], Jussi Klemelä[‡], Enno Mammen[§]

June 17, 2008

Abstract

Linearity in a causal relationship between a dependent variable and a set of regressors is a common assumption throughout economics. In this paper we consider the case when the coefficients in this relationship are random and distributed independently from the regressors. Our aim is to identify and estimate the distribution of the coefficients nonparametrically. We propose a kernel based estimator for the joint probability density of the coefficients. Although this estimator shares certain features with standard nonparametric kernel density estimators, it also differs in some important characteristics which are due to the very different setup we are considering. Most importantly, the kernel is nonstandard, and derives from the theory of Radon transforms. Consequently, we call our estimator the Radon Transform Estimator (RTE). We establish the large sample behavior of this estimator, in particular rate optimality and asymptotic distribution. In addition, we extend the basic model to cover extensions including endogenous regressors and additional controls. Finally, we analyze the properties of the estimator in finite samples by a simulation study, as well as an application to consumer demand using British household data.

Keywords: Random Coefficient Model, Inverse Problems, Kernel Density Estimation, Radon Transform.

*We have received helpful comments from seminar participants in Copenhagen, Göttingen, Frankfurt, Mannheim, Oberwolfach, UCL, Northwestern and Wien (ESEM), as well as from Jim Heckman and Arthur Lewbel.

[†] University of Mannheim, Department of Economics, L7, 3-5, 68131 Mannheim, Germany, email: stefan_hoderlein@yahoo.com. Financial support by Landesstiftung Baden-Württemberg "Eliteförderungsprogramm" is gratefully acknowledged.

[‡] University of Oulu, Department of Mathematical Sciences, P. O. Box 3000, 90014 University of Oulu, Finland, email: klemela@oulu.fi.

[§] University of Mannheim, Department of Economics, L7, 3-5, 68131 Mannheim, Germany, email: emammen@rumms.uni-mannheim.de.

1 Introduction

Heterogeneity of individual agents, in particular consumers or firms, is a prevalent notion throughout economics. In addition, it is often the case that the individuals are, at least approximately, characterized by a linear relationship between a d-vector of explanatory variables, and a dependent variable. Combining these two notions yields in a natural fashion to the random coefficient model (RCM),

$$Y_i = \beta_i^T X_i, \tag{1}$$

where Y_i is an observed continuously distributed random scalar, X_i denotes an observed random d- vector of individual specific regressors and β_i is an unobserved random d-vector of individual coefficients. In this model, the subscript i denotes individual observation, and we may include an intercept, i.e. $X_{i,1} \equiv 1$, so that we may rewrite model (1) as $Y_i = \beta_{i,2} X_{i,2} + ... + \beta_{i,d} X_{i,d} + \varepsilon_i$ with an error term $\varepsilon_i = \beta_{i,1}$. The RCM is arguably the oldest and most important way of expressing the notion of unobserved heterogeneity in econometrics through allowing the marginal effects (summarized in β) to vary across individuals.

Traditionally, the random coefficient model has been investigated under mean independence, i.e. $\mathbf{E}\left[\beta_i | X_i\right] = \beta$, and homoscedasticity, i.e. $Cov\left[\beta_i | X_i\right] = \Sigma_\beta$, see the classic references of Hildreth and Huock (1968) and Swamy (1970), any standard econometrics textbook, e.g., Wooldridge (2002) or the recent survey in Hsiao and Pesaran (2004). While this allows to identify the average marginal effect and the variance, important features of the joint distribution of marginal effects are left unidentified (unless one is willing to assume, e.g., joint normality). These includes the quantiles of the marginals, as well as skewness, kurtosis or symmetry of the distribution or dependence structure of various components of β. Moreover, the question of multimodality, or the related question whether the population consists of a mixture of subpopulations are left unanswered. Finally, there are many instances where it is interesting to evaluate whether the probability density of the parameter is significantly different from zero on a specified set, which corresponds to the notion of whether a restriction on the parameters holds across a heterogeneous population.

We study the random coefficient model (1) under the stronger independence assumption that β is independent from X, or from instruments Z. It is the aim of this paper to show that under this assumption the joint distribution of marginal effects is identified nonparametrically and to propose a sample counterpart estimator which allows to analyze the joint distribution of marginal effects. To give an example how our method works in practice and how it may reveal interesting features of the distribution of marginal effects across the population, consider Figure 1. This figure displays an estimate of the joint distribution of the income and uncompensated own price elasticities of food consumption[1], controlling for household observables. The graph is a contour plot with the solid lines equal to the level lines, akin to lines of similar altitude on a map. It shows a clearly unimodal distribution, which is slightly skewed towards the southwestern corner.

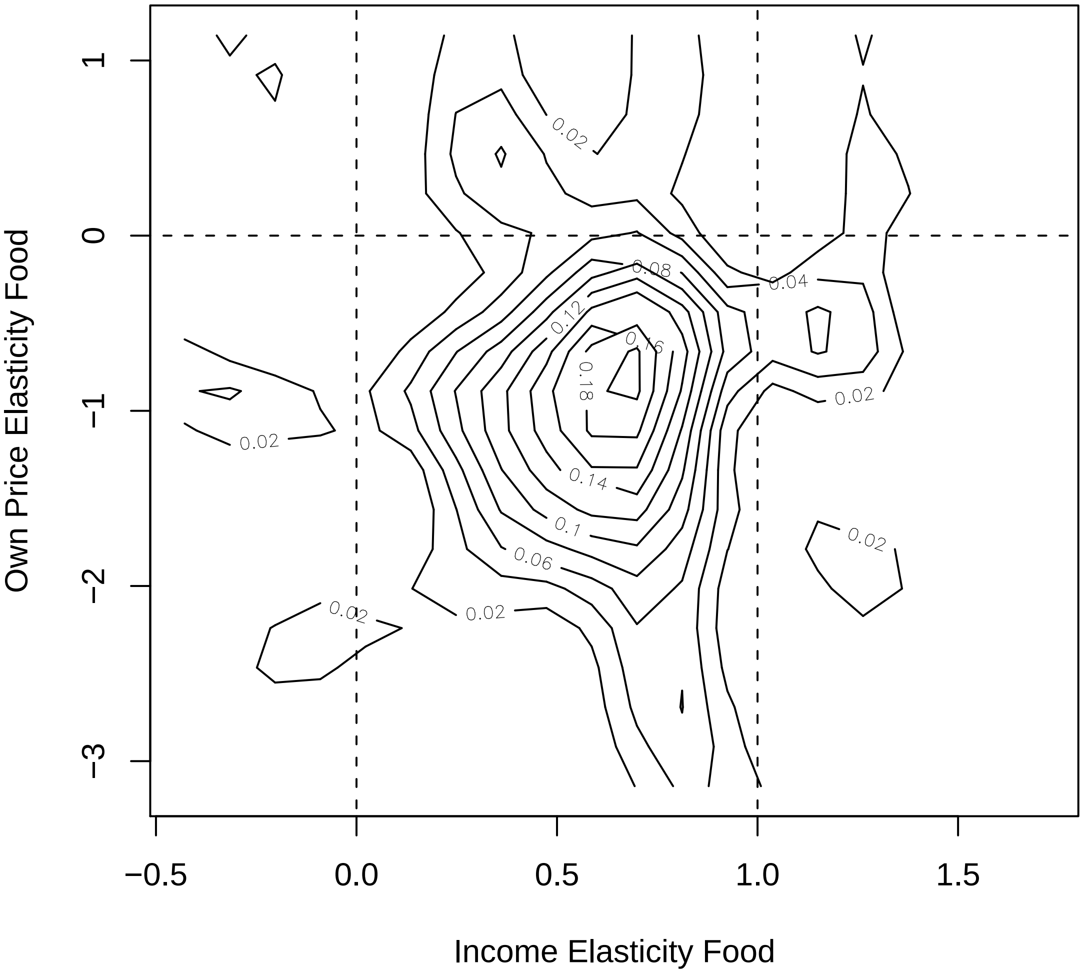

Figure 1: Application of the RT Estimator to Demand for Food – Contour plot of Joint Density of Elasticities

There seems to be little association between the two marginal effects, indicating that individuals with low income elasticities are equally likely to have high and low own price elasticities. More importantly, almost all the income elasticities are between 0 and 1, indicating that food is a normal good, but not a luxury good across the entire population. Quite interestingly,

[1] To be precise, the income elasticities are actually total expenditure elasticities. However, we use the more common terminology.

a small but potentially significant area of the population shows positive uncompensated own price elasticities. While this fact alone could also be interpreted as food being a Giffen good for these individuals, the fact that for all of these individuals food is normal (i.e., the income elasticities are positive) rules out this explanation. Hence, if we are willing to accept both the linearity in model (1) and the independence assumption, we conclude that there is an indication that standard consumer theory may be an invalid description for a fraction of the population[2]. However, for a more careful analysis we have to make sure that the density in this area is significantly positive, which requires an asymptotic distribution theory for our estimator in the first place.

The structure of our RT estimator is simple, and very much resembles a standard kernel density estimator. More precisely, the estimator for the joint density of random coefficients at a fixed position, $f_\beta(b)$ is given by

$$\hat{f}_\beta(b) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(S_i^T b - U_i\right)\left(\hat{f}_S(S_i)\right)^{-1}, \qquad b \in \mathbf{R}^d,$$

where U_i and S_i are suitable transformations of Y_i and X_i, see (2), K_h is an appropriate kernel, and $\hat{f}_S$ denotes an estimator for the density of the transformed regressors. The differences to standard kernel density estimation are the nonstandard kernel, as well as the normalization by the density of transformed regressors.

The details of identification and estimation of $f_\beta(b)$ will occupy much of the first part of this paper. Specifically, for identification we apply the theory of Radon transforms that has been used in computer tomography. Nonparametric estimation in random coefficient models has been considered in Beran and Hall (1992), Beran, Feuerverger and Hall (1996), and Feuerverger and Vardi (2000). The first paper extends the familiar strategy to estimate the first moments and then to make use of the independence assumption, by estimating higher order moments. The other two papers propose to estimate the characteristic function of the response variable and then transform this estimator back. The latter paper also discusses numerical aspects and links random coefficient models to tomography.

Like Beran, Feuerverger and Hall (1996) and Feuerverger and Vardi (2000), our estimator is build upon the Radon transform. However, in contrast to their work, our approach utilizes the Radon transform directly to construct a simple estimator which employs a one-step procedure. Because our approach is direct, it is also much better suited for use as building block in more complicated models appearing in econometrics, e.g., endogenous regressors, or for hypothesis testing (see also Heckman and Vytlacil (1998) for an alternative approach to deal with endogeneity). In particular, it allows us to consider additional covariates in a semiparametric

[2] Related work in consumption includes in particular Foster and Hahn (2000).

fashion, which neither of the approaches mentioned can. Also, we are the first to derive rate optimality. Technically, there are also some parallels between our approach and Korostelev and Tsybakov (1993). However, they estimate a Radon transformed regression function, which is conceptually different. Other approaches in statistics that treat empirical Radon transforms are based on singular value decompositions of the Radon operator (Johnstone and Silverman, 1990) or on wavelet-vaguelette decompositions (Donoho, 1995 and Abramovich and Silverman, 1998). For a discussion of estimators in inverse problems including empirical Radon transforms based on empirical risk minimization, see Klemelä and Mammen (2008). See also Natterer (2001) or Helgason (1999) for an overview on the mathematics of Radon transforms.

The random coefficient model (1) is a mixing model. The distribution of β is the mixing distribution. There exist classical approaches in the statistical literature for the identification of mixing distributions. Mixing models have been used in econometrics to capture heterogeneity. Recent work in econometrics includes Matzkin (2007), Briesch, Chintagunta and Matzkin (2007), Fox and Gandhi (2008) and Bajari, Fox, Kim and Ryan (2007) where also other references can be found. These papers contain results on the identification of mixture distributions in random choice models. An early reference for nonparametric identification in binary choice is Ichimura and Thompson (1998). For this model, Gautier and Kitamura (2008) contains a detailed asymptotic theory for a nonparametric estimator that is based on a singular value decomposition. An alternative route to heterogeneity is to dispense with identifiability of mixing distributions and to check for identification of local average marginal effects, see e.g. Hoderlein (2007) and Hoderlein and Mammen (2007, 2008).

Our estimator of the density of β achieves optimal rates of convergence in Sobolev classes. The estimator depends on the unknown smoothness parameter of the Sobolev class and on the bandwidth of a kernel. We state formulas for the asymptotic variance and the asymptotic bias. For the case of twice differentiable functions we have an asymptotic bias expression that depends on second order partial derivatives in a similar way as in classical kernel smoothing. Thus one can use plug-in estimates of the bandwidth as in classical nonparametric kernel smoothing. We do, however, not give a theoretical discussion of bandwidth choice here. Our estimator makes use of a kernel estimator of the density of the covariates (scaled to the unit sphere). The choice of the kernel and of the bandwidth of the spherical kernel estimator of the design density only affects second order properties of the estimator (as long as the order of the bandwidth lies in a certain range). Further theoretical work is needed to construct methods for the automatic choice of the two smoothing parameters of the estimator.

This paper is organized as follows: In section two, we establish nonparametric identification, and use this result in sections three and four to construct a sample counterpart estimator, and analyze its asymptotic properties. In section five, we discuss extensions towards endogeneity and the inclusion of additional control variables. The small sample and real world performance of our estimator is in the focus in sections six and seven, where we consider simulation and application to consumer demand. Finally, we conclude with an outlook.

2 Nonparametric Identification of the Joint Density of Random Coefficients

Let us first state the model and the setup: throughout this paper, we will always assume to have i.i.d. random vectors (Y_i, X_i, β_i), $i = 1, \ldots, n$, with $Y_i \in \mathbf{R}$ and $X_i, \beta_i \in \mathbf{R}^d$ with the following structural relationship between the variables:

$$Y_i = \beta_i^T X_i, \qquad i = 1, \ldots, n.$$

Our goal is to estimate the density of the vector β_i, which we denote by $f_\beta : \mathbf{R}^d \to \mathbf{R}$. The key identification assumption is that X_i and β_i are independent. Note that we require at this point full independence, which may seem a strong assumption. However, the entire model specification is in principle testable, if one splits the support of X_i into two regions, and then derives the estimator of the density in each region separately. A simple nonparametric density comparison test would then be sufficient to check the specification. But one should be aware that the illposedness of the Radon transform may also cause problems in case of small deviations from independence.

In order to derive our estimator, we use the following transformation $(Y_i, X_i) \mapsto (U_i, S_i)$, $i = 1, \ldots, n$, where

$$S_i = \|X_i\|^{-1} X_i \in \mathbf{S}_{d-1}, \qquad U_i = \|X_i\|^{-1} Y_i \in \mathbf{R}. \tag{2}$$

By $\| \cdot \|$ we denote the Euclidean norm in $\mathbf{R}^d$. Moreover, the unit sphere in $\mathbf{R}^d$ is denoted by $\mathbf{S}_{d-1} = \{z \in \mathbf{R}^d : \|z\| = 1\}$. Then our model becomes:

$$U_i = \beta_i^T S_i, \qquad i = 1, \ldots, n$$

with S_i independent of β_i.

A key concept in the following will be that of a Radon transform, which is defined as the integral of a function over lower dimensional hyper-planes. We parametrize the $d-1$-dimensional hyperplanes in the d-dimensional Euclidean space by a direction vector $s \in \mathbf{S}_{d-1}$ and a distance from the origin $u \in \mathbf{R}$:

$$P_{s,u} = \{z \in \mathbf{R}^d : z^T s = u\}.$$

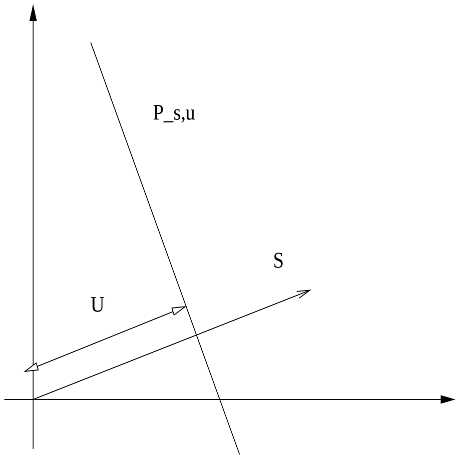

Figure 2: Parametrization of hyperplanes: line $P_{s,u}$ is parametrized with direction vector s and distance u. The lines are parametrized by the length u of the perpendicular from the origin to the line and by the orientation s of this perpendicular.

See Figure 2 for an illustration of the parametrization. The Radon transform of a function $f : \mathbf{R}^d \to \mathbf{R}$ is then formally defined as

$$(Rf)(s,u) = \int_{P_{s,u}} f,$$

where the integration is with respect to the $d-1$-dimensional Lebesgue measure on the hyperplane $P_{s,u}$, for any function $f : \mathbf{R}^d \to \mathbf{R}$ integrable on each hyperplane. Radon observed that a function is completely determined by all its integrals (over lower dimensional hyperplanes). This fact was rediscovered and utilized in computer tomography, see Natterer (2001) or Helgason (1999). The basic observation in our model is that the conditional density of U given S is given by the Radon transform of the density f_β:

$$f_{U|S}(u|s) = Rf_\beta(s,u). \tag{3}$$

Indeed, this conditional density is obtained by integrating f_β for each u, over the plane perpendicular to s. Intuitively, we would now like to invert the operator R to obtain the unknown density of interest f_β from the observable conditional density of the transformed variables, $f_{U|S}$. This, however, is an ill-posed inverse problem. The inverse operator is not smooth, i.e. small changes in the argument may result in big changes in the value. To solve this problem, one has to use a regularized inverse A_h of the Radon transform. A regularized inverse is given by the operator $A_h : \{g : \mathbf{S}_{d-1} \times \mathbf{R} \to \mathbf{R}\} \to \{f : \mathbf{R}^d \to \mathbf{R}\}$, defined by

$$(A_h g)(z) = \int_{\mathbf{S}_{d-1}} \int_{-\infty}^{\infty} K_{r,h}(s^T z - u) g(s,u)\, du d\mu(s), \qquad z \in \mathbf{R}^d, \tag{4}$$

where μ is the Lebesgue measure on the unit sphere $\mathbf{S}_{d-1}$. The definition of $K_{r,h}$ is slightly involved, however, its properties will turn out to make it similar to a smoothing kernel. Formally, it is defined by its Fourier transform[3]

$$\widetilde{K}_{r,h}(t) = \frac{1}{2}\,(2\pi)^{-d+1}|t|^{d-1}L_r(h|t|), \qquad t \in \mathbf{R}, \tag{5}$$

where $h > 0$ is a smoothing parameter, and $L_r : [0,\infty) \to \mathbf{R}$ is a function

$$L_r(t) = \begin{cases} (1-t^r)I_{[0,1]}(t), & 0 < r < \infty \\ I_{[0,1]}(t), & r = \infty, \end{cases} \qquad t \in \mathbf{R} \tag{6}$$

depending on the parameter $0 < r \leq \infty$. From now on, we suppress the dependence of K and L on r and write

$$K_{r,h} = K_h, \qquad L_r = L.$$

The choice of the parameters h and r will be discussed below. It can be checked that the kernel K_h has the following explicit representation:

$$K_h(u) = (2\pi)^{-d}\int_0^\infty \cos(tu)t^{d-1}L(ht)\,dt, \qquad u \in \mathbf{R}.$$

The kernel K_h depends on the bandwidth h and on the dimension d and the order parameter r which both appear in the definition of the function L, see (6). Figure 2 shows this kernel function K_h for smoothing parameter values $h = 0.3, 0.5, 0.6$. Frame a) shows the case when $d = 2$ and $r = 2$, frame b) shows the case when $d = 2$ and $r = \infty$, and frame c) shows the case when $d = 4$ and $r = 2$. By definition of L, the frequency of the oscillations increases with increasing smoothing parameter h.

In the appendix we show that $\|(A_hRf) - f\|_2$ is of order h^s if f has square integrable derivatives of order s, see Lemma 6[4]. The result shows that the operator A_h is a regularized inverse, i.e. that

$$\lim_{h\to 0}\|(A_hRf) - f\|_2 = 0.$$

This suggests to construct the estimator for f_β at a fixed position b as sample counterpart to

$$f_\beta(b) = (A_hf_{U|S})(b) = \int_{\mathbf{S}_{d-1}}\int_{-\infty}^{\infty} K_h(s^Tb - u)f_{U|S}(u|s)\,dud\mu(s), \tag{7}$$

where all quantities are as defined above.

[3] The Fourier transform of an integrable function $g : \mathbf{R}^d \to \mathbf{R}$ is defined by

$$\widetilde{g}(\omega) = \int_{\mathbf{R}^d} \exp\{iz^T\omega\}g(z)\,dz, \qquad \omega \in \mathbf{R}^d,$$

while the inverse Fourier transform of an integrable function $\tilde{g}$ is given by

$$g(z) = (2\pi)^{-d}\int_{\mathbf{R}^d} \exp\{-iz^T\omega\}\widetilde{g}(\omega)\,d\omega, \qquad z \in \mathbf{R}^d.$$

Note that for $f, g \in L_1(\mathbf{R}^d) \cap L_2(\mathbf{R}^d)$, $\int_{\mathbf{R}^d} fg = (2\pi)^{-d}\int_{\mathbf{R}^d} \widetilde{f}\widetilde{g}$ holds.

[4] Here, $\|g\|_2^2 = \int_{\mathbf{R}^d} g^2(x)dx$ denotes the L$_2$-norm.

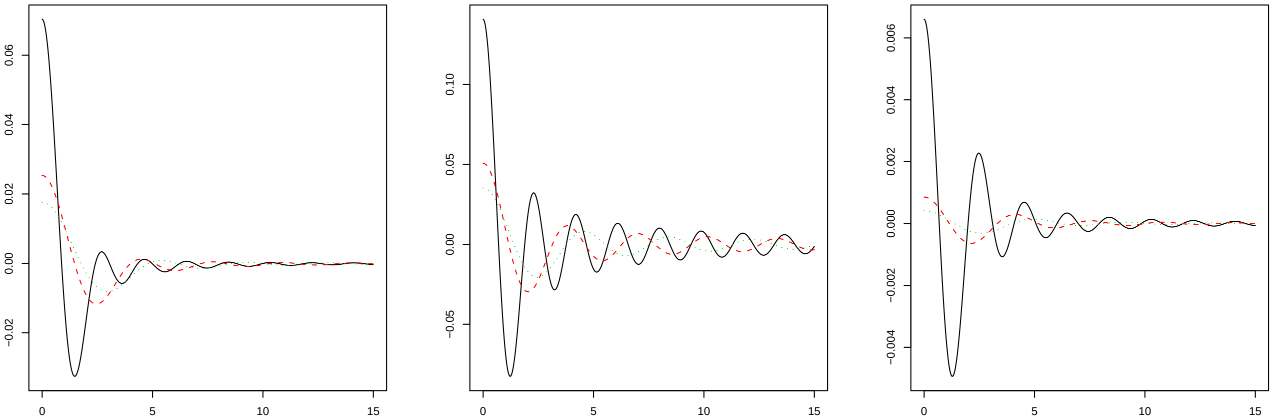

Figure 3: The kernel function K_h when $h = 0.3, 0.5, 0.6$; a) $d = 2$ and $r = 2$; b) $d = 2$ and $r = \infty$; c) $d = 4$ and $r = 2$. The case $h = 0.3$ is shown as a solid line, $h = 0.5$ is shown as a dashed line, and $h = 0.6$ is shown as a dotted line.

3 A Sample Counterpart Estimator

Using the analogy principle, our estimator is defined as sample counterpart to (7), i.e.

$$\hat{f}_\beta(b) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\hat{f}_S(S_i)} K_h\left(S_i^T b - U_i\right), \qquad b \in \mathbf{R}^d. \tag{8}$$

We call our estimator the Radon transform estimator (RTE). In contrast to a standard kernel density estimator which is just a sum of kernels, we require also an estimator $\hat{f}_S$ of the density f_S. At this point it is obvious that trimming will be needed if S_i has density near zero. However, we first consider the non trimming case as the deviations and the intuition are easier. For the estimation of f_S we need an estimator for densities on the unit sphere. In the simulations and in the application we will use a kernel density estimator. A standard kernel smoothing approach for spherical data is given by

$$\hat{f}_S(s) = \frac{c(g)}{n} \sum_{i=1}^{n} G\left(g^{-2}(1 - s^T S_i)\right), \qquad s \in \mathbf{S}_{d-1}, \tag{9}$$

where $G : [0, \infty) \to \mathbf{R}$ is a kernel function, $g > 0$ is the smoothing parameter, and $c(g)$ is the normalization constant:

$$c(g)^{-1} = \int_{\mathbf{S}_{d-1}} G\left(g^{-2}(1 - s^T e)\right) d\mu(s),$$

for any $e \in \mathbf{S}_{d-1}$, see e.g. Klemelä (2000). Here, μ is the Lebesgue measure on $\mathbf{S}_{d-1}$. Note that $\|s - e\|^2 = 2(1 - s^T e)$ for $s, e \in \mathbf{S}_{d-1}$, so that the estimator is a kernel estimator with a spherically symmetric kernel. Reasonable choices for the kernel function are for example $G(t) = e^{-t} I_{[0,\infty)}(t)$, $G(t) = (1 - t) I_{[0,1]}(t)$, and $G(t) = I_{[0,1]}(t)$.

4 Large Sample Behavior of the Radon Transform Estimator

4.1 Rate Optimality

We start the section on asymptotic behavior with a result on rate-optimality of our estimator under Sobolev smoothness conditions. Sobolev smoothness of order $s > 0$ is defined by use of the Sobolev semi-norm $\rho_s(f)$, defined by

$$\rho_s^2(f) = (2\pi)^{-d} \int_{\mathbf{R}^d} \|\omega\|^{2s} \left| \widetilde{f}(\omega) \right|^2 d\omega,$$

for functions $f : \mathbf{R}^d \to \mathbf{R}$ with Fourier transform $\widetilde{f}$. A function f fulfills Sobolev smoothness of order s if $\rho_s^2(f) < \infty$. For integer s this holds if all partial weak derivatives of f of order s are square integrable. Thus Sobolev smoothness naturally generalizes ordinary smoothness of integer orders to the case of fractional orders. Throughout this section we shall make use of the following assumptions:

(A1) The vectors X_i and β_i are independent i.i.d. sequences.

We make the following assumptions on the densities f_β and f_S.

(A2) f_β satisfies the following Sobolev smoothness condition. For some $s > 0$,

$$\rho_s^2(f_\beta) < \infty.$$

Moreover, the density f_β is bounded with bounded support and $\int |\tilde{f}_\beta(\omega)| d\omega < \infty$.

(A3) The density f_S is bounded and the estimator $\widehat{f}_S$ achieves the following rate

$$\sup_{\xi \in \mathbf{S}_{d-1}} \left| \widehat{f}_S(\xi) - f_S(\xi) \right| = o_P\left(n^{-s/(2s+2d-1)} \right). \tag{10}$$

(A4) f_S is bounded away from zero: there exists $0 < C_S < \infty$ with

$$\inf_{\xi \in \mathbf{S}_{d-1}} f_S(\xi) \geq C_S^{-1}. \tag{11}$$

Assumption (A1) is our basic model assumption. Assumption (A2) contains the basic smoothness condition on the density of β. According to (A2) the density f_β has s derivatives. Because of our general notion of smoothness fractional values of s are allowed. The standard example is $s = 2$, which will be discussed in Remark 6. The assumption that the Fourier transform of f_β is integrable is done for technical reasons. It implies that f_β is continuous.

But there exist no mild and tractable conditions that imply this assumption. In Remark 3 we show that this assumption can be avoided at the cost of assuming higher rates of convergence for the design density estimator $\widehat{f_S}$. Assumption (A3) can be easily verified for kernel density estimators $\widehat{f_S}$, as defined in (9). If f_S has bounded partial derivatives of order two and if the bandwidth g is chosen of order $n^{-1/(3+d)}$, then it holds that

$$\sup_{\xi \in \mathbf{S}_{d-1}} \left| \widehat{f_S}(\xi) - f_S(\xi) \right| = O_P\left(\sqrt{\log n} n^{-2/(3+d)} \right).$$

This can be checked by using classical smoothing theory, see Appendix A. Thus for the estimator (9) assumption (A3) holds if $s < (4d-2)/(d-1)$. If f_S is σ-times differentiable the kernel G in the definition of $\widehat{f_S}$ could be replaced by a higher order kernel. Then $\widehat{f_S}$ fulfills (A3) (if the bandwidth g is chosen of order $n^{-1/(2\sigma+d-1)}$) as long as $\sigma/(2\sigma+d-1) > s/(2s+2d-1)$. Here, for (A3) it is not necessary that $\sigma \geq s$, i.e. less smoothness is required for f_S compared with f_β. For more details, see Appendix A. Assumption (A4) will be discussed below.

The following theorem gives the rate of convergence of the estimator $\hat{f}_\beta$ defined in (8).

Theorem 1 *Let assumptions (A1)-(A4) hold, let the kernel L be defined in (6) with $s \leq r \leq +\infty$ and let the smoothing parameter h of $\hat{f}_\beta$ satisfy*

$$h = h_n \asymp n^{-1/(2s+2d-1)}.$$

Then for any bounded subset B of $\mathbf{R}^d$,

$$\int_B \left| \hat{f}_\beta(b) - f_\beta(b) \right|^2 db = O_P\left(n^{-2s/(2s+2d-1)} \right).$$

A proof of Theorem 1 is given in Section 8.

Remark 1 A proof that the rate given in Theorem 1 is the minimax rate can be given similarly as in Theorem 9.5.3 of Korostelev and Tsybakov (1993). It is intuitive to compare the rate with optimal rates for the estimation of a k-order derivative of a density f on $\mathbf{R}^d$. If f has s derivatives the optimal rate is $n^{-(s-k)/(d+2s)}$. Formally, this is equal to the rate in Theorem 1 for $k = s(d-1)/(2s+2d-1)$. Thus, estimation of the density of coefficients in a random coefficients model is asymptotically as hard as the estimation of a derivative of a density of this order. For $s = 2$ we get $(2d-2)/(2d+3)$ which is always smaller as 1.

4.2 The Asymptotic Distribution of the Radon Transform Estimator

In the next theorem we show that our estimator $\hat{f}_\beta(b)$ is asymptotically normal. More precisely,

Theorem 2 *Under the assumptions of Theorem 1,*

$$\sqrt{nh^{2d-1}}\sigma_n^{-1}(b)\left[\hat{f}_\beta(b) - f_\beta(b) - h^s bias_n(b)\right]$$

converges in distribution to a standard normal limit. Here

$$\begin{aligned}\sigma_n^2(b) &= \int_{\mathbf{S}_{d-1}} \frac{1}{f_S(s)}\, h^{2d-1} \int_{-\infty}^{\infty} K_h^2(s^T b - u) Rf_\beta(s,u)\, du d\mu(s),\\ bias_n(b) &= h^{-s}(2\pi)^{-d} \int_{\mathbf{R}^d} \left[L(h\|\omega\|) - 1\right] \widetilde{f}_\beta(\omega) \exp\{-ib^T\omega\}\, d\omega,\end{aligned}$$

and it holds that

$$\begin{aligned}\sigma_n^2(b) &\leq h^{2d-1} C_S C_\beta \mu(\mathbf{S}_{d-1}) \int_{-\infty}^{\infty} K_h^2(u)\, du\\ &= h^{2d-1} C_S C_\beta \mu(\mathbf{S}_{d-1})(2\pi)^{-d}\frac{1}{4}\,(2\pi)^{-2d+2} \int_{-\infty}^{\infty} |t|^{2(d-1)} L^2(h|t|)\, dt\\ &= C_S C_\beta \mu(\mathbf{S}_{d-1})(2\pi)^{-3d+2}\frac{1}{2}\left[\frac{1}{2d-1} - 2\frac{1}{2d-1+r} + \frac{1}{2d-1+2r}\right],\end{aligned}$$

where $C_S^{-1} = \inf_s f_S(s)$ *and where* $C_\beta = \sup_{s,u} Rf_\beta(s,u)$. *Finally,*

$$|bias_n(b)| \leq \rho_s(f_\beta).$$

A proof of Theorem 2 is given in Section 8.

Remark 2 A consistent estimator of $\sigma_n^2(b)$ is given by

$$\widehat{\sigma}_n^2(b) = \frac{h^{2d-1}}{n} \sum_{i=1}^n \frac{1}{\widehat{f}_S(S_i)^2} K_h^2(S_i^T b - U_i).$$

Remark 3 As remarked after the statement of the assumptions there exist no mild and tractable conditions that imply that the Fourier transform of f_β is integrable. This was assumed in (A2). We now argue that this assumption can be avoided at the cost of assuming higher rates of convergence for the design density estimator $\widehat{f}_S$. It can be shown that Theorems 1 and 2 still hold without assuming that $|\widetilde{f}_\beta|$ is integrable if one makes the following additional assumption. In (A3) instead of (10) one has to assume that

$$\sup_{\xi \in \mathbf{S}_{d-1}} \left|\widehat{f}_S(\xi) - f_S(\xi)\right| = o_P\left(n^{-(s+\frac{d}{2})/(2s+2d-1)}\right).$$

For a proof of the modified versions of Theorems 1 and 2 one uses the Cauchy-Schwarz inequality in the last inequality of the proof of Lemma 4. This gives a bound $O(h^{-d/2})$ instead of $O(1)$ in this inequality. The statement of Lemma 4 has to be adjusted accordingly.

If one uses a spherical kernel density estimator $\widehat{f}_S$ and if one makes the assumptions on f_S specified in the appendix then one needs that $\sigma > s + d/2$, i.e. one needs $d/2$ more derivatives for the function f_S as for the density f_β.

Remark 4 The parameter r determines the order of the kernel K_h. This can be seen from the expansion for the bias in Theorem 2. The term $[L(h\|\omega\|) - 1]$ can be absolutely bounded by $\|h\omega\|^s$ for $\|h\omega\| \leq 1$. This directly leads to the bound for the bias at the end of Theorem 2.

Remark 5 The statement of Theorem 2 also holds for choices of the bandwidth h that are not of the order specified in the statement of Theorem 1.

Remark 6 In this remark we treat the case of two times differentiable functions and kernels K_h with $r = 2$, i.e. kernels that correspond to the usual second order kernels in classical kernel smoothing problems. Assume that the Assumptions (A1)-(A4) hold with $s = 2$ and that $\int \|\omega\|^2 |\tilde{f}(\omega)| d\omega < \infty$. Let the kernel L be defined in (6) with $r = 2$ and let the smoothing parameter h of $\hat{f}_\beta$ satisfy

$$h = h_n \asymp n^{-1/(2s+2d-1)}.$$

Then for a fixed point b,

$$\sqrt{nh^{2d-1}} \sigma_n^{-1}(b) \left[\hat{f}_\beta(b) - f_\beta(b) - h^2 bias_n(b) \right]$$

converges in distribution to a standard normal limit where $\sigma_n^2(b)$ is defined as in Theorem 2 and where

$$bias_n(b) = \sum_{j=1}^{d} \frac{\partial^2}{(\partial b_j)^2} f_\beta(b).$$

A proof of this remark is given in Section 8. This result is very similar to classical results on kernel smoothing for two times differentiable densities. The bias is of order h^2, depends only on second derivatives, and is hence estimable given an estimator of the second derivatives. Note also that the bias depends only on the local shape of the function, which shows that our estimation approach indeed localizes.

4.3 Discussion of Condition (A4)

For models where X_i has full dimensional support Assumption (A4) is very mild. There are a number of interesting models that fall into this class, e.g. fixed effects panel data after differencing, i.e. $\triangle Y_i = \beta_i' \triangle X_i$, $i = 1, \ldots, n$, or structural models where there is a lower number of underlying parameters, and the model is linear in underlying parameters, e.g., $Y_i = \beta_i'(\gamma + X_i)$, $i = 1, \ldots, n$.

However, this assumption is restrictive for the case where the design includes an intercept, i.e. $X_{i,1} \equiv 1$. To see why this is the case, consider $d = 2$, so that $X_i = (1, X_{i,2})^T$. Let the density of $X_{i,2}$ be denoted as $f_{2,X}$. Then, assumption (A4) requires that $\liminf_{u \to \pm\infty} u^2 f_{2,X}(u) > 0$, which means in particular that the second moment of $X_{i,2}$ is infinite.

To circumvent this shortcoming, we propose a modification of $\hat{f}_\beta$ that uses trimming to avoid estimation of f_S at regions where this density is too small. One version of a trimming estimator for the density of β is given by

$$\hat{f}_\beta^\tau(b) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\hat{f}_S(S_i)} K_h\left(S_i^T b - U_i\right) I_{\mathbf{S}_{d-1}\setminus C_\tau}(S_i), \qquad b \in \mathbf{R}^d, \tag{12}$$

where C_τ is the following band on the unit sphere

$$C_\tau = \{s \in \mathbf{S}_{d-1} : |s_1| \leq \tau\}, \qquad 0 < \tau < 1. \tag{13}$$

Here τ is a trimming parameter that may depend on n.

We discuss the modified estimator for the case $s = 2$ and with $\widehat{f_S}$ defined as in (9). The discussion can be easily generalized to other values of s. For $s = 2$ we get the following result.

Theorem 3 *Assume that $X_{i,1} \equiv 1$ and make the assumptions (A1)-(A2) with $s = 2$. Let the kernel L be defined in (6) with $2 \leq r \leq +\infty$. Let the smoothing parameter h of $\hat{f}_\beta^\tau$ satisfy $h \to 0$ and $nh^{2d-1} \to \infty$ and let the estimator $\widehat{f_S}$ be defined as in (9) with smoothing parameter g of order*

$$g = g_n \asymp n^{-1/(3+d)}.$$

Assume that f_S has bounded partial derivatives of order 2 and that $d_\tau^{-1}(\log n)^{-1/2}n^{-2/(3+d)} \to 0$ for $d_\tau = \inf_{\xi\in C_\tau} f_s(\xi)$. Then for any bounded subset B of $\mathbf{R}^d$,

$$\int_B \left|\hat{f}_\beta^\tau(b) - f_\beta(b)\right|^2 db \leq C\left(h^4 + \int_B \nu_{h,\tau}(b)^2 db\right) + O_P\left(n^{-1}h^{-2d+1} + d_\tau^{-2}(\log n)n^{-4/(3+d)}\right),$$

where C is a constant,

$$\nu_{h,\tau}(b) = (2\pi)^{-d}\int_{\mathbf{R}^d} I_{D_\tau}(\omega)L(h\|\omega\|)\widetilde{f}_\beta(\omega)\exp\{-ib^T\omega\}\, d\omega$$

and where

$$D_\tau = \left\{z \in \mathbf{R}^d : z = ts,\;\; t \in [0,\infty),\;\; s \in C_\tau\right\}.$$

A proof of Theorem 3 is given in Section 8.

Remark 7 Our trimming estimator has the same order of variance as the estimator without trimming but its bias is much more involved. Indeed, the bias expression is too complicated for an intuitive understanding and is not helpful for the practical implementation of the estimator. In our simulations the estimator without trimming worked quite well even for data sampled from a model with intercept.

Remark 8 For the additional bias term, one can use the following crude bound:

$$\begin{aligned}
\|\nu_{h,\tau}\|_2^2 &\asymp \int I_{D_\tau}(\omega)L^2(h\|\omega\|)\left|\tilde{f}_\beta(\omega)\right|^2 d\omega \\
&\leq \int_{\mathbf{S}_{d-1}} d\mu(s) \int_0^\infty dr\, r^{d-1} I_{D_\tau}(rs)L^2(hr) \\
&\leq \int_{\mathbf{S}_{d-1}} d\mu(s) \int_0^{h^{-1}} dr\, r^{d-1} I_{C_\tau}(s) \\
&\asymp \tau \int_0^{h^{-1}} r^{d-1}\, dr \\
&\asymp \tau h^{-d}.
\end{aligned}$$

5 Extensions

5.1 Endogeneity

Frequently in econometrics there are reasons to believe that the independence assumption between regressors and unobservables is violated. In consumer demand for instance, we may think of the distribution of coefficients as being generated by heterogeneity in preferences across the population. However, the assumption of independence of preferences and regressors like household characteristics or total expenditures may be rightfully questioned. Hence some way of dealing with endogeneity may also be desirable in our setup.

The standard concept for handling this type of endogeneity are instruments. In the textbook linear model these are variables that are uncorrelated with the unobservables but correlated with the endogenous regressors. In our setup, we devise a similar solution. More precisely, one possible specification that retains the linear structure and blends in nicely with the textbook models is the following:

$$\begin{aligned}
Y_i &= X_i^T \beta_i \\
X_i &= \Gamma Z_i + V_i,
\end{aligned}$$

where the notation is as above, but Z_i denotes a random L-vector of instruments, Γ is a nonrandom $d \times L$ matrix of coefficients, and V_i denotes a random d-vector of residuals. Under standard conditions, there exists a root n consistent estimator of Γ, denoted by $\widehat{\Gamma}$.

For identification of $f_\beta(b)$ we require that Z_i be (jointly) independent of (β_i, V_i), as is easily seen by simply plugging in the second equation into the first and rearranging terms. This is a straightforward, but interesting finding: In Hoderlein (2007), Hoderlein and Mammen (2007, 2008) we consider the case where Y_i is a nonseparable function of regressors and Borel space valued unobservables, e.g., preferences, without assuming monotonicity in unobservables. In this scenario, if regressors are endogenous, we argue that joint independence of unobservables in both equations, i.e. (β_i, V_i) above, from the instruments is sufficient to identify the average structural derivatives, but not the individuals marginal effect. Here, however, we are able to identify every individuals' marginal effect due to the assumption of linearity across the population.

With respect to estimation, we would apply the theory of the previous section to

$$Y_i = \delta_{0i} + Z_i^T \delta_{1i}.$$

This can be easily done as long as $\delta_{1i} = \Gamma^T \beta$ has not a degenerate distribution and it needs some additional theoretical work otherwise. The density f_δ gives the joint density f_β, by using $\beta_i = \left[\Gamma^T\right]^- \delta_{1i}$ where $[A]^-$ denotes the Moore Penrose inverse of a matrix A. More precisely, let $\hat{H} = \left(\left[\widehat{\Gamma}^T\right]^-, B\right)^T$ with $B = (0_{L-d\times d}, I_{L-d})$ and $\widehat{\Gamma}$ is a root n consistent estimator for Γ. Then,

$$\hat{f}_\beta(b) = \int \det \left|\hat{H}^{-1}\right|^{-1} \hat{f}_\delta \left(\hat{H}^{-1} (b, \delta)^T\right) d\delta.$$

It is straightforward to show that $\hat{f}_\beta(b) = \bar{f}_\beta(b) + O_p\left(n^{-1/2}\right)$, where $\bar{f}_\beta(b) = \det\left|H^{-1}\right|^{-1}$ $\hat{f}_\delta\left(H^{-1} (b, \delta)^T\right)$.

5.2 Controlling for the Influence of other Variables and for Nonlinearities

Another frequent event in econometrics is that some variables are of greater relevance for the researcher than others. In particular, it is often the case that a set of variables play only the role of controls. In regression analysis, this fact has lead to semiparametric models like the popular partially linear model, i.e. $Y_i = m(X_{1i}) + X_{2i}^T\delta + V_i$, where $X_i = \left(X_{1i}^T, X_{2i}^T\right)^T$, V_i denotes a mean independent error, m is a smooth function, and δ a vector of fixed coefficients.

When estimating the joint density of the random coefficients, we face again the same problem of curse of dimensionality that is inherent to the nonparametric literature. Hence we may translate the same semiparametric solution to our model, i.e., we assume that

$$Y_i = X_{1i}^T\beta_i + X_{2i}^T\delta, \tag{14}$$

where δ does not vary across individuals. Note that under our assumptions,

$$\mathbb{E}\left[Y_i|X_i\right] = X_{1i}^T\mu_\beta + X_{2i}^T\delta,$$

where $\mu_\beta = \mathbb{E}\left[\beta_i\right]$, and hence

$$Y_i - \mathcal{L}\left(Y_i|X_i\right) = X_{1i}^T\left(\beta_i - \mu_\beta\right),$$

where $\mathcal{L}\left(Y_i|X_i\right)$ denotes the linear projection of Y_i on X_i. This suggests to use the residuals of an OLS regression of Y_i on all X_i as new dependent variable, say $\tilde{Y}_i$, and use the data $\left(\tilde{Y}_i, X_{1i}\right)$ to obtain an estimator of $f_{\beta-\mu_\beta}$ as above. Plugging in an estimator for μ_β yields an estimator for f_β. By similar arguments as in the previous subsection, it follows that the asymptotic distribution of this estimator for f_β does not differ from that of $\hat{f}_\beta$ as detailed above, since both $\tilde{Y}_i$ and the OLS estimator for μ_β are root n consistent.

There are two alternative ways to deal with the control variables. The first is related to the famous Frisch-Waugh partitioned regression principle and still uses the model as defined in equation (14): Let $\tilde{Y}_i = Y_i - \mathcal{L}\left(Y_i|X_{2i}\right)$, and $\tilde{X}_{1i} = X_{1i} - \mathcal{L}\left(X_{1i}|X_{2i}\right)$. Then, apply our estimator using the data $\left(\tilde{Y}_i, \tilde{X}_{1i}\right)$ to obtain an estimator for f_β. This approach is related to the estimation idea in Christopeit and Hoderlein (2006), and works only if the mean regressions of Y_i and X_i are truly linear, and $\tilde{X}_{1i}$ is fully independent from X_{2i}, and not just mean independent. Moreover, with this procedure the distribution of the original intercept α is not identified.

The second alternative way to treat additional control variables does not assume that model (14) holds, but assumes that β_i depends in a nonparametric fashion on covariates, i.e. $\beta_i = \beta\left(X_{2i}, A_i\right)$, where β is smooth in x_2 and A_i denotes unobservables. In this case, an estimator for the conditional density of β_i given X_{2i} can be obtained by carrying $X_{2i} = x_2$ through all arguments. Consequently, an estimator has the form

$$\hat{f}_{\beta|X_2}(b,x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\hat{f}_{SX_2}(S_i, x_2)} K_h\left(S_i^T b - U_i\right) \mathcal{K}_\eta(X_{2i} - x_2), \qquad b \in \mathbf{R}^d.$$

where $\mathcal{K}_\eta$ is a standard multivariate kernel with bandwidth vector η. The large sample behavior of such an estimator is straightforward using the tools introduced.

The following modification of our model takes care of nonlinearities:

$$Y_i = \sum_{j=1}^{d} \beta_{i,j} m_j(X_{i,j}). \tag{15}$$

Here m_j are functions that are purely nonparametric or that are parametrically specified. For j with $\mathbb{E}[\beta_i] \neq 0$ the functions m_j can be estimated by using

$$\mathbb{E}[Y_i|X_i] = \sum_{j=1}^{d} \mathbb{E}[\beta_i] m_j(X_{i,j}). \tag{16}$$

For the estimation of other components one can use estimates of $\mathbb{E}[Y_i^2|X_i]$. Equation (16) constitutes a nonparametric additive model. This model can be fitted e.g. by the smooth backfitting approach of Mammen, Linton and Nielsen (1999). Smooth backfitting has been used in Borak, Härdle, Mammen and Park (2007), Mammen, Støve and Tjøstheim (2008) and Connor, Hagmann and Linton (2008) for models related to (15) to identify temporary and individual effects, respectively, not captured by explanatory variables.

6 Simulation

It is one of the particularly interesting features of our estimator that it allows to identify different subpopulations within the overall population. Whereas our application will turn out to reveal an unimodal population, in our simulation study we will consider a bimodal case and we will focus on the estimators ability to recover and display a population generated by a mixture of normals in a small sample. The details of our simulation study are as follows. For $j = 1, 2$, we choose ξ_j to be bivariate normal, i.e., $\xi_j \backsim \mathcal{N}(\mu_j, \Sigma)$, where $\mu_1 = (-3, -3)', \mu_2 = (3, 3)'$, and $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. The overall population is composed to equal parts out of the two subpopulations, i.e., $\beta = \mathbb{I}\{\theta \geq 0\}\xi_1 + \mathbb{I}\{\theta < 0\}\xi_2$, and $\mathbb{P}\{\theta \geq 0\} = 0.5$. The marginal distribution of the bivariate β is as shown in figure 3.

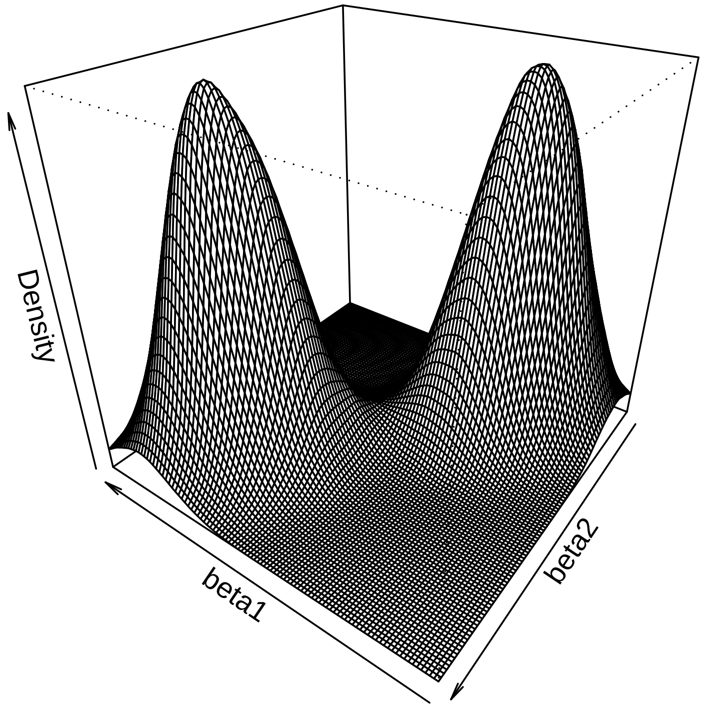

Figure 4: True Joint Density of β

We assume that $\alpha \backsim \mathcal{N}(0, 2)$, and $\alpha \perp \beta$. Moreover, the regressors X are $\mathcal{N}(0, 2.5\Sigma)$, and

$(\alpha, \beta) \perp X$. The model is given by

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2.$$

For $n = 500, 1000, 2500$ observations we calculate the average L$_2$-error (ALE) of our estimator:

$$ALE = \left\{ \int [\widehat{f}_\beta(b) - f_\beta(b)]^2 f_\beta(b) \; db \right\}^{1/2}.$$

This integral is calculated by Monte Carlo integration. The calculation is repeated 500 times. The average of the values for the ALE gives an approximation for the mean average L$_2$-error ($MALE$):

$$MALE = E[ALE] = E\left[\left\{ \int [\widehat{f}_\beta(b) - f_\beta(b)]^2 f_\beta(b) \; db \right\}^{1/2} \right].$$

We select the optimal bandwidths beforehand by doing a grid search with respect to finding the vector of bandwidths that minimizes the $MALE$ in 100 replications. Figure 4 shows contour plots of the estimator $\widehat{f}_\beta$ for two quantiles of the distribution of ALE. More specifically, in figure 4 we show the true DGP , i.e. f_β, as solid lines, while to give a feeling for a "good" and a "poor" realization of the estimator, we display the realization whose ALE are at the 0.8 and 0.2 quantiles of the distribution of ALE, respectively.

To provide a comparison, we also analyze the behavior of an infeasible ("oracle") estimator. The oracle estimator makes direct use of the unobserved random coefficients β_i and is given as the kernel density estimator of f_β, i.e.

$$\bar{f}_\beta(b) = n^{-1} \sum_{i=1}^{n} \mathcal{K}_h(\beta_i - b),$$

where $\mathcal{K}_h$ denotes a two dimensional product kernel, i.e. $\mathcal{K}_h(\beta_s - b) = h^{-2} K(h^{-1}(\beta_{1,s} - b_1))$ $K(h^{-1}(\beta_{2,s} - b_2))$, and K denotes the Epanechnikov kernel. For both the RTE and the oracle estimator, no higher order bias reduction (e.g., by higher order kernels or higher order local polynomials) was employed so that the result are comparable from this perspective. For $n =$ 1000 figure 5 shows the density of ALE of both, oracle and RT estimators.

As expected, the density of the ALE of the oracle estimator has most of its mass to the left of the one of the RTE, and consequently the infeasible oracle estimator outperforms the RT. However, the extend of the outperformance is tolerable: Both the median and the mean, as well as many quantiles of the ALE distribution are approximately twice as large, while the spread in the ALE is roughly comparable. To summarize how our results change with changing

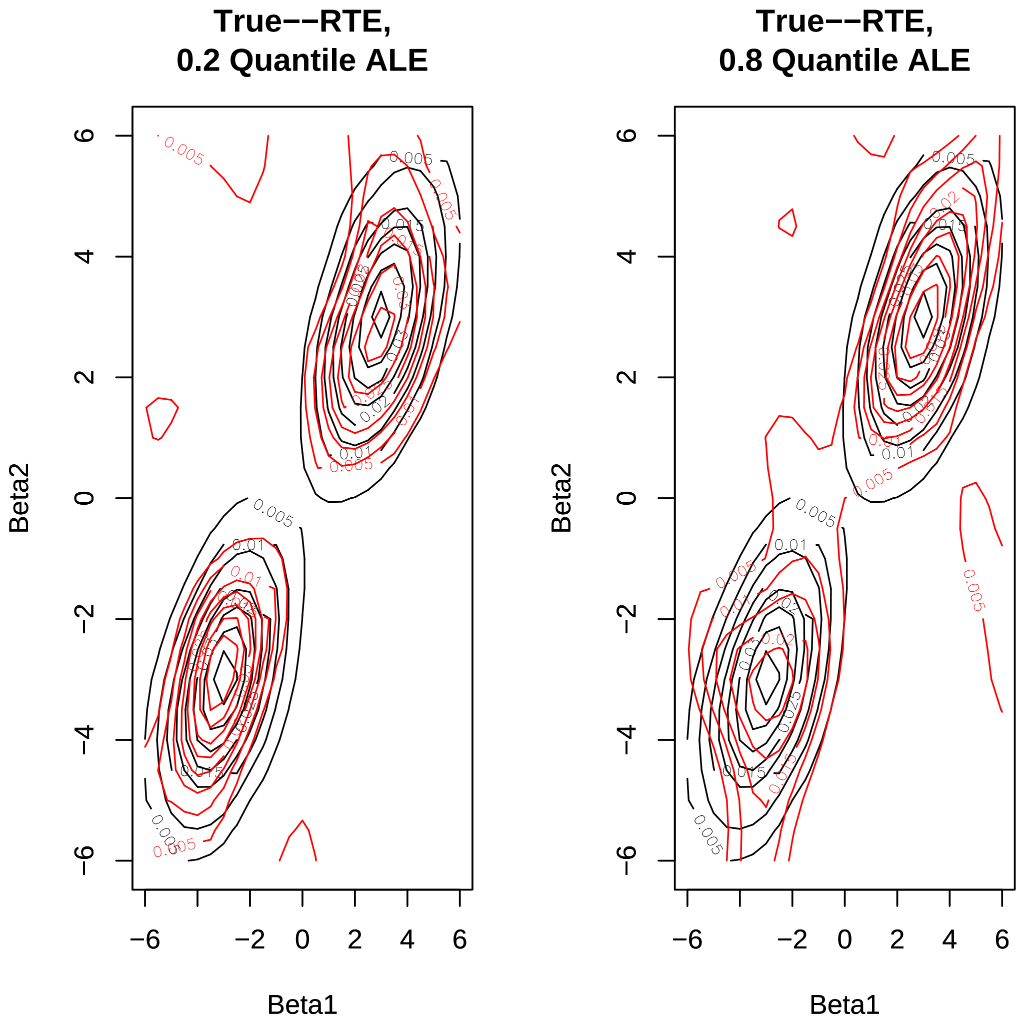

Figure 5: The RT Estimator with $n = 1000$ Observations. Realizations at 0.2 and 0.8 Quantile (light grey) vs. True Model (solid)

sample size, consider the following table which contains the MALE at $n = 500, 1000$ and 2500.

Data Size	$n = 500$	$n = 1000$	$n = 2500$
(I) *Mean ALE* $Radon\ Transform\ Estimator$	1.3431	0.8825	0.5769
(II) *Mean ALE* $Oracle\ Estimator$	0.6314	0.4420	0.3337
Ratio $(I)/(II)$	2.1272	1.9966	1.7288

Table 6.1: Comparison of MALE of
RTE and Oracle for different data size n.

Obviously, both estimators improve as n becomes larger, as can be seen from the first two rows. But note that the relative inefficiency of the RTE compared to the Oracle increases implying that as the data size increases, the behavior of our RTE estimator resembles more and more

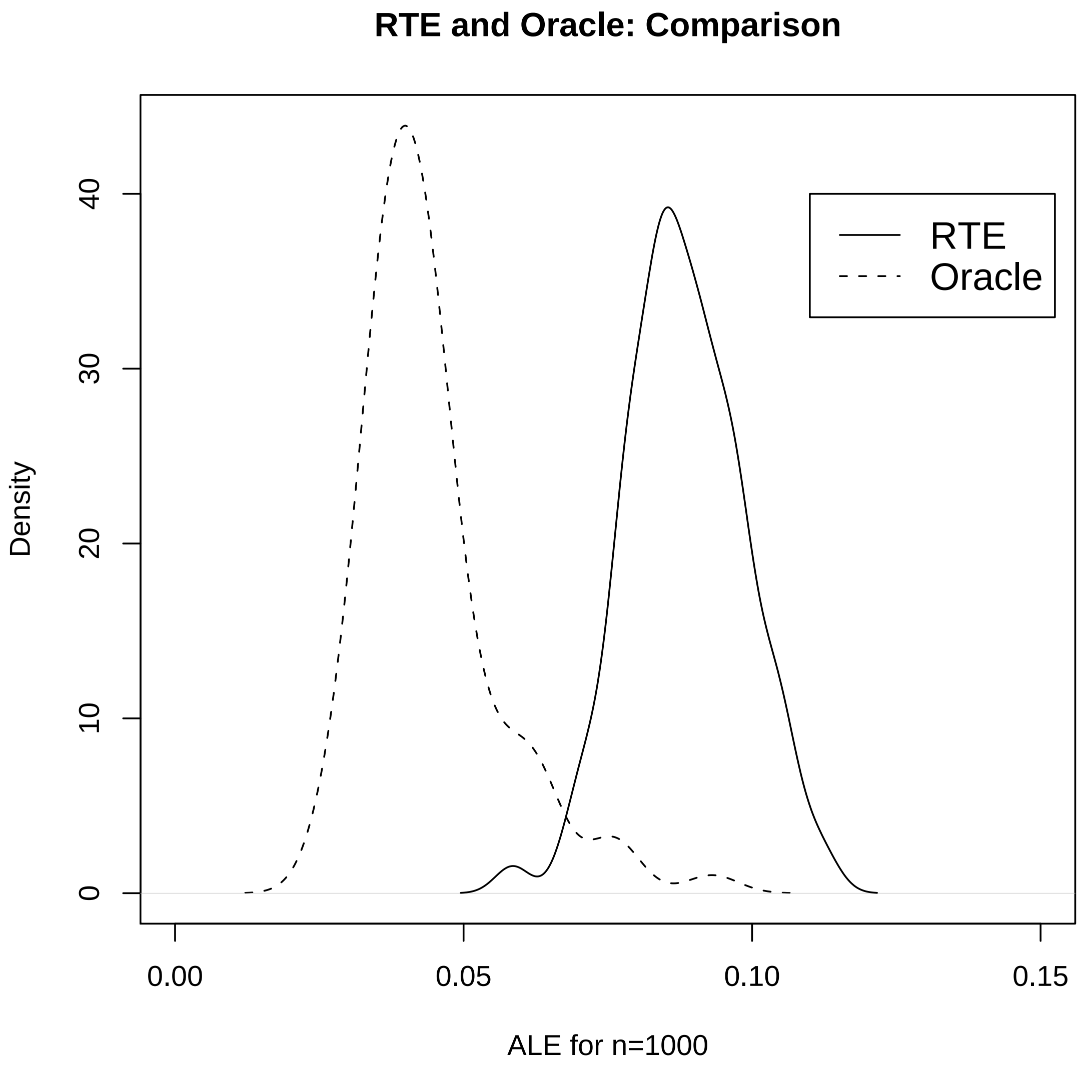

Figure 6: Densities of ALE for RT estimator and oracle estimator

that of an infeasible benchmark estimator.

Is this relative inefficiency compared with an infeasible optimum really severe in an application? To assess this question and find possible explanations, consider the following figure (fig. 6) which shows a comparison of RTE and oracle estimator at the respective median ALE (in a sense, a “typical” realization of the DGP).

Two things are apparent: First and unsurprisingly, the infeasible oracle estimator gives a more accurate approximation to the true DGP. Second, with respect to the main features the differences in quality of the fit are rather small. The main features, in particular the location and height of the two peaks, are almost equally well captured. The only obvious difference are the wiggles of the RT estimator in the tails of the distribution, which actually account for a good part of the difference in the ALEs. It is important to note that these wiggles have

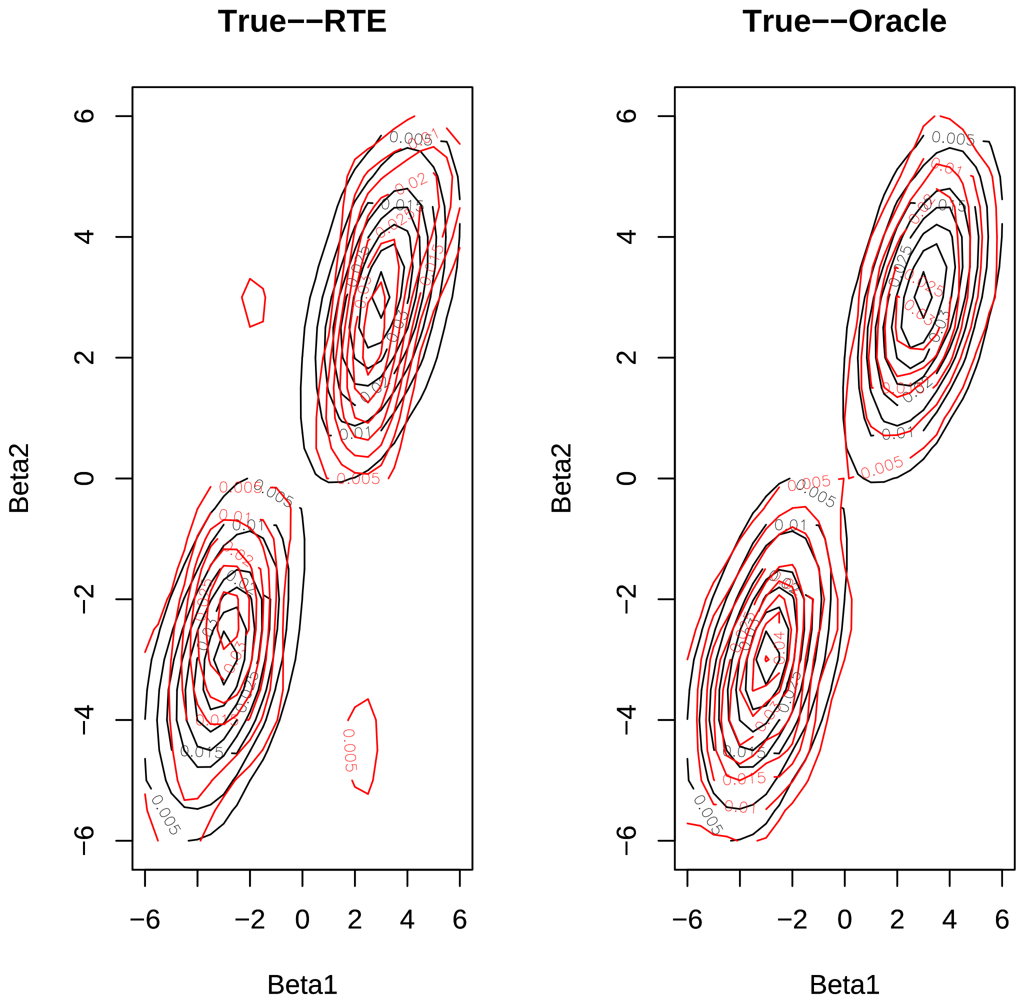

Figure 7: Median ALE Realizations of Estimator of Density of Random Coefficient of RT (light grey) vs. True Model (solid) and Oracle (light grey) vs. True Model (solid)

been less pronounced in models without intercept (not reported here). Consequently, they may be understood as the small sample effects of limits in the identifiability of the distribution of marginal effects.

But note also from both the graphical and the numerical evidence that the behavior of the RTE is acceptable even for rather moderate data sizes, i.e. $n = 500$ or 1000. We conclude that in this experiment the RTE works well with a data size commonly encountered in practice, provided the researcher limits the analysis to models that do not have heterogeneity of marginal effects in too many dimensions.

7 Application to Consumer Demand for Food

In this section we focus on applying our kernel estimator to a real world application, as already outlined in the introduction. Our motivation comes from consumer demand, and we use British household data. The section will consist of three subsections: In the first, we give a short motivation of our approach, and how it compares to similar work in the demand literature. Then we will provide a data description and will discuss some related issues. Finally, we give an overview about the results.

7.1 Unobserved Heterogeneity in Consumer Demand

For a long time unobserved heterogeneity has been a major issue in the demand literature. This importance is driven by the data: For any given level of income and prices the observed demand (expressed in budget shares) varies enormously across individuals. Correspondingly, the R^2 of cross regressions has been extremely low. As a consequence, the focus in the demand literature has been on ways of modeling this unobserved heterogeneity.

In an important paper, Lewbel (2001) provides a framework for modeling heterogeneity in consumer demand, see also related work by Hoderlein (2007). Lewbel (2001) discusses also a case of importance for this paper, namely that of a heterogeneous linear (in his case, almost ideal) population, but he does not propose any estimator for the distribution of random coefficients. Our approach allows now to estimate the distribution of coefficients of a linear model in a heterogeneous population. More specifically, we consider the model

$$W_i = \alpha_i + \gamma_i^T P_i + \lambda_i \left[Y_i - H\left(P_i\right)\right] + \delta^T Z_i \tag{17}$$

where W_i is the d-vector of budget shares, P_i is a d-vector of log prices and Y_i denotes log nominal income. $H\left(P_i\right)$ is a log price index. For simplicity, we consider one good only, and we take the standard shortcut and choose the log GDP deflator as $H\left(P_i\right)$, so that $Y_i - H\left(P_i\right)$ represents log real income. α_i, λ_i and γ_i in turn are now random parameters that vary across the population. To mitigate the curse of dimensionality, the coefficient on observable demographic characteristics Z_i have been made invariant across individuals. For these theoretical demand functions homogeneity requires $\gamma_i 1_d = 0$.

Following Hoderlein and Lewbel (2006), we replace all prices other than the own price by a single price index, and use their result that homogeneity still holds in the dimension reduced regression[5]. Imposing the specification of Hoderlein and Lewbel, homogeneity of degree zero, and applying partitioned regression, we obtain

$$W_i - \mathcal{L}\left(W_i|Z_i\right) = \tilde{\alpha}_i + \beta_i^T \left[X_i - \mathcal{L}\left(X_i|Z_i\right)\right]$$

[5] Under certain not very restrictive conditions on the stochastic process of prices which we assume to be true.

as estimating equation, where $X_i = (X_{1i}, X_{2i})^T = (Y_i - H(P_i), P_{F,i} - P_{R,i})^T$, $P_{F,i}$ denotes food price, $P_{R,i}$ denotes the price index for the remainder, and $\tilde{\alpha}_i$ contains remaining factors as described in section 5.2. From this estimating equation we may determine in particular the joint distribution of marginal income and uncompensated own price semi-elasticities.

The focus of our analysis is on the income elasticity, and the compensated own price elasticity. These may be obtained by using the fact that $\varepsilon_{Inc,i} = \beta_{Inc,i}/W_i + 1$, and $\varepsilon_{PF,i} = \beta_{Food,i}/W_i - 1$. Using the identifying independence assumption between coefficients and regressors, one can obtain the joint density of elasticities by applying a transformation formula to the estimators of $f_{\tilde{\alpha}\beta}$ and f_W.

7.2 The Data

Every year, the FES reports the income, expenditures, demographic composition and other characteristics of about 7,000 households. The sample surveyed represents about 0.05% of all households in the United Kingdom. The information is collected partly by interview and partly by records. Records are kept by each household member, and include an itemized list of expenditures during 14 consecutive days. The periods of data collection are evenly spread out over the year. The information is then compiled and provides a repeated series of yearly cross-sections.

The category of goods we consider is food related, and consists of the subcategories food bought and catering, which are self explanatory. Together our food category accounts for 28% of expenditures on average. We removed outliers by excluding the upper and lower 2.5% of the population. Income in demand is total expenditure under the assumption of additive separability of the preferences. It is roughly defined as all (nominal) expenditures on nondurable goods excluding some that are known to contain measurement error.

7.3 Results

When estimating our consumer demand model as defined by (17) with random coefficient by applying the RTE to the FES food data, we find the results as detailed in fig. 1 in the introduction which shows an estimate of the joint distribution of the log real income and uncompensated own price elasticities of food consumption. We control for the preference heterogeneity associated with household observables in the following way: First, we stratify the population to obtain a relatively homogeneous subpopulation, which is equivalent to controlling for the influence of discrete controls nonparametrically. Like much of the demand literature we focus on one subpopulation (namely two person households, both adults, at least one working and the head of household a white collar worker), to minimize measurement error. For brevity of exposition, we do not report results controlling for endogeneity. However, we do control for the influence of other characteristics by partitioning them out as described above.

The resulting elasticities were computed as detailed in section 5.2 and 7.2. As already mentioned above, the contour plot displayed in fig. 1 suggests a clearly unimodal distribution, which is skewed towards the southwestern corner. It is interesting to note that the own price elasticities are more spread out than the income elasticities. Frequently, in applied demand analysis price elasticities are only imprecisely estimated, and the results vary a lot according to data and subpopulation considered. Though this was often attributed to insufficient price variation, this graph suggests that it might be due to the large heterogeneity in price effects.

As already mentioned, it is impossible to assess whether there is a significant part of the population showing positive own price elasticities without a formal test, but it appears to be at least possible. With respect to other outlying low density areas as those at negative income and own price elasticities, in light of the simulation evidence gathered one should caution against over-interpreting these areas. Indeed, these areas correspond most likely again to wiggles, and are not a genuine feature of the data.

What we would like to analyze in this section is the effect of specification on our results. Even though an approximately linear structure of budget shares in income and prices is frequently postulated, there may be obvious doubts whether the same truly holds in reality for all individuals. Hence, we consider an alternative regression, namely one were we apply our model to the relationship between log expenditure for food and the same covariables as in model (17), and look whether our qualitative findings remain robust to the change in specification.

As a result of the application of our RTE we obtain again real income and uncompensated own price elasticities of food consumption, see figure 7.

Several things are noteworthy. First, the results by and large agree with the findings of the budget share specification. The income elasticities indicate again that food is a normal, but not a luxury good for almost the entire population. Observe that the bulk of the income elasticities are somewhat lower in this specification, which is more in line with parametric findings. Second, the own price elasticities are mostly non-positive. However, again a potentially significant fraction of the population displays non-rational behavior. Third, the price elasticities are perhaps even more spread out than in the budget share specification, emphasizing our previous point about very heterogeneous price effects. Fourth, there is again little correlation between the two marginal effects, but note that there is some evidence of conditional heteroscedasticity, as the distribution of income effects seems more spread out for individuals with low own price elasticities.

In summary, both specifications produce qualitatively similar results, although some of the details vary. Another interesting finding is obtained by comparison with the evidence

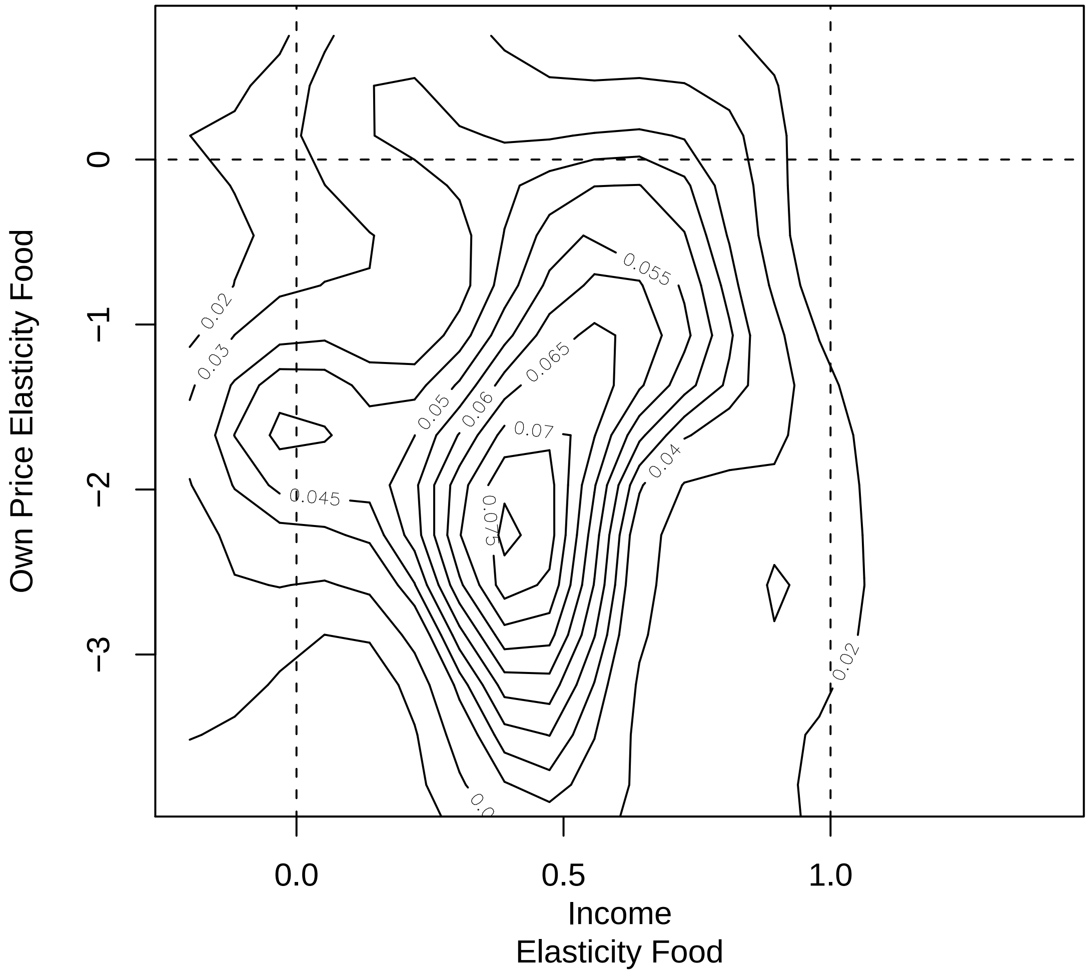

Figure 8: Application of the RT Estimator to Demand for Food – Contour plot of Joint Density of Elasticities, Dependent Variable log Food Expenditure

in the simulations. In the application, there is no evidence whatsoever of subpopulations. Indeed, the population displays a lot of heterogeneity, but the coefficients vary smoothly across the population with a clear maximum, but without much clustering. Though the existence of "types" is frequently postulated in economic theory, the evidence gathered here suggests otherwise for the case of food demand.

8 Summary and Outlook

The random coefficient model allows for a great deal of heterogeneity in marginal effects of individual agents. In this paper we consider the linear random coefficient model that allows for a nonparametric treatment of this unobserved heterogeneity. We establish a nonparametric identification relation for the underlying mixing distribution, and propose a structurally simple sample counterpart estimator, which we call the Radon-Transform (RT) estimator. The large sample behavior of the RT estimator is also obtained, and can be handled by arguments that are standard in nonparametric smoothing. Through a simulation study, as well as through an application to consumer demand, we establish that the RT estimator works well in data sets commonly encountered in practice. Analyzing other areas of applied economics by the RT method may also reveal interesting facts about rationality, about the existence of "types" or about a variety of other potential questions which were not analyzed in this paper. Ultimately, finding additional areas of application will determine whether this new view on the random coefficient model will be successful.

Appendix I: Proofs

Preliminary Lemmas

We first compare $\widehat{f_\beta}$ with a theoretical estimator $\bar{f}_\beta$ which is defined as $\hat{f}_\beta$, but with $\hat{f}_S(S_i)$ replaced by $f_S(S_i)$:

$$\bar{f}_\beta(b) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{f_S(S_i)} K_h\left(S_i^T b - U_i\right), \qquad b \in \mathbf{R}^d. \tag{18}$$

We now prove that these two estimators are asymptotically equivalent.

Lemma 4 *Under the assumptions of Theorem 1 it holds that*

$$\begin{aligned} \bar{f}_\beta(b) - \hat{f}_\beta(b) &= o_P\left(n^{-s/(2s+2d-1)}\right) \qquad \textit{for } b \in B, & (19)\\ \int_B \left|\bar{f}_\beta(b) - \hat{f}_\beta(b)\right| db &= o_P\left(n^{-s/(2s+2d-1)}\right) & (20)\end{aligned}$$

Proof. We only prove (19). Claim (18) follows by similar arguments. We have that

$$\bar{f}_\beta(b) - \hat{f}_\beta(b) = \frac{1}{n}\sum_{i=1}^{n} \Delta(S_i)\, K_h\left(S_i^T b - U_i\right).$$

with

$$\Delta(S_i) = \frac{1}{f_S(S_i)} - \frac{1}{\hat{f}_S(S_i)}.$$

We will show that

$$\mathbb{E}\left|\mathbb{E}\left[K_h\left(S_i^T b - U_i\right)|S_i\right]\right| = O(1), \tag{21}$$

$$\sup_{s\in\mathbf{S}_{d-1}} \mathbb{E}\left[K_h\left(S_i^T b - U_i\right)^2 |S_i = s\right] = O(h^{-2d+1}). \tag{22}$$

The statement of the lemma immediately follows from these two claims. This can be seen by using the decomposition

$$\begin{aligned}
\bar{f}_\beta(b) - \hat{f}_\beta(b) &= \frac{1}{n}\sum_{i=1}^{n} \Delta(S_i)\left\{K_h\left(S_i^T b - U_i\right) - \mathbb{E}\left[K_h\left(S_i^T b - U_i\right)|S_i\right]\right\} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \Delta(S_i)\mathbb{E}\left[K_h\left(S_i^T b - U_i\right)|S_i\right].
\end{aligned}$$

The second term can be easily bounded by using (21) and

$$\max_{1\le i\le n} |\Delta(S_i)| = o_P\left(n^{-s/(2s+d-1)}\right). \tag{23}$$

Equation (23) follows from assumptions (A3) and (A4); the convergence of $\max_{1\le i\le n}(\hat{f}_S(S_i) - f_S(S_i))$ with the rate $o_P(n^{-s/(2s+d-1)})$ follows from assumption (A3) and the fact $\max_{i=1,\dots,n} \frac{1}{\hat{f}_S(S_i)} = O_P(1)$ follows from assumption (A4) and Stute (1984). The first term can be bounded by using (22) and (23).

We now show claim (21). For fixed $s \in \mathbf{S}_{d-1}$, denote with $\widetilde{Rf}_\beta(s,t)$ the Fourier transform of $u \mapsto Rf_\beta(s,u)$ and with $\widetilde{f}_{U|S}(t|s)$ the Fourier transform of $u \mapsto f_{U|S}(u|s)$. We have that

$$\widetilde{Rf}_\beta(s,t) = \widetilde{f}_\beta(ts), \qquad t \in \mathbf{R}, \tag{24}$$

because of

$$\widetilde{Rf}_\beta(s,t) = E[\exp(itU)|S=s] = E[\exp(its^T\beta)|S=s] = E[\exp(its^T\beta)] = \widetilde{f}_\beta(ts).$$

Equation (24) is also called the projection theorem, see Natterer (2001). For fixed $s \in \mathbf{S}_{d-1}$ and fixed $z \in \mathbf{R}^d$ the Fourier transform of $u \mapsto K_h(s^T z - u)$ is

$$t \mapsto \widetilde{K}_h(t)\exp\{-its^T z\}.$$

Thus, for fixed $s \in \mathbf{S}_{d-1}$ and fixed $z \in \mathbf{R}^d$,

$$\begin{aligned}
&\mathbb{E}\left[K_h\left(S_i^T z - U_i\right) \mid S_i = s\right] \qquad (25)\\
&= \int_{-\infty}^{\infty} K_h(s^T z - u) Rf_\beta(s,u)\, du \\
&= (2\pi)^{-1}\int_{-\infty}^{\infty} \widetilde{K}_h(t)\exp\{-its^T z\}\widetilde{Rf}_\beta(s,t)\, dt \\
&= (2\pi)^{-1}\frac{1}{2}(2\pi)^{-d+1}\int_{-\infty}^{\infty} |t|^{d-1} L(h|t|)\exp\{-its^T z\}\widetilde{f}_\beta(ts)\, dt.
\end{aligned}$$

This implies

$$\mathbb{E}\left|\mathbb{E}\left[K_h\left(S_i^T b - U_i\right)|S_i\right]\right|$$

$$\begin{aligned} &= 2(2\pi)^{-1}\frac{1}{2}(2\pi)^{-d+1}\int_{\mathbf{S}_{d-1}} f_S(s)d\mu(s)\left|\int_0^\infty t^{d-1}L(ht)\exp\{-its^Tb\}\widetilde{f}_\beta(ts)\,dt\right| \\ &\le \int_{\mathbf{R}^d} L(h\|\omega\|)\left|\widetilde{f}_\beta(\omega)\right| d\omega\; O(1) \\ &\le (2\pi)^{-d}\int_{\mathbf{R}^d}\left|\widetilde{f}_\beta(\omega)\right| d\omega\; O(1) = O(1). \end{aligned}$$

This shows claim (21).

For the proof of claim (22) note that uniformly for $s \in \mathbf{S}_{d-1}, b \in B$

$$\begin{aligned} \mathbb{E}\left[K_h\left(S_i^Tb - U_i\right)^2 | S_i = s\right] &= \int_{-\infty}^{\infty} K_h^2(u)\,du\; O(1) \\ &= \int_{-\infty}^{\infty} |t|^{2d-2}L^2(h|t|)\,dt\; O(1) = O(h^{-2d+1}). \end{aligned} \tag{26}$$

Here we have used that $f_{U|S}$ is bounded because we assumed in (A2) that the density f_β of β is bounded with bounded support, and $f_{U|S}$ is written in (3) as a Radon transform of f_β. This shows claim (22). □

Proof of Theorem 1

Because of Lemma 4 it suffices to prove the theorem for the estimator $\bar{f}_\beta$ defined in (18). Theorem 1 follows from Lemma 5 and Lemma 7 given below. Lemma 5 states a bound for the bias of $\bar{f}_\beta$.

Lemma 5 *Under the assumptions of Theorem 1 it holds that,*

$$\|\mathbb{E}\bar{f}_\beta - f_\beta\|_2^2 \le \rho_s(f_\beta)^2 h^{2s}.$$

Proof. Using (3) we get for $b \in \mathbf{R}^d$,

$$\begin{aligned} \mathbb{E}\bar{f}_\beta(b) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\frac{1}{f_S(S_i)}\,\mathbb{E}\left[K_h\left(S_i^Tb - U_i\right) \mid S_i\right]\right) \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\frac{1}{f_S(S_i)}\int_{-\infty}^{\infty} K_h(S_i^Tb - u)Rf_\beta(S_i,u)\,du\right) \\ &= A_hRf_\beta(b), \end{aligned}$$

with A_h defined in (4). The lemma follows by application of the following lemma. □

Lemma 6 *For a function $f \in L_2(\mathbf{R}^d)$ and for the operator A_h defined with $s \le r \le \infty$ in (6) it holds that*

$$\|A_hRf - f\|_2^2 \le h^{2s}\rho_s(f)^2.$$

Proof. Proceeding as in (25) we get that for fixed $s \in \mathbf{S}_{d-1}$ and fixed $b \in \mathbf{R}^d$,

$$\begin{aligned}
&\int_{-\infty}^{\infty} K_h(s^T z - u) Rf(s,u)\, du \\
&\quad = (2\pi)^{-1} \int_{-\infty}^{\infty} \widetilde{K}_h(t) \exp\{-its^T z\} \widetilde{Rf}(s,t)\, dt \\
&\quad = \frac{1}{2} (2\pi)^{-d} \int_{-\infty}^{\infty} |t|^{d-1} L(h|t|) \exp\{-its^T z\} \widetilde{f}(ts)\, dt.
\end{aligned}$$

Thus,

$$\begin{aligned}
A_h Rf(b) &= (2\pi)^{-d} \int_{\mathbf{S}_{d-1}} d\mu(s) \int_0^{\infty} t^{d-1} L(ht) \exp\{-its^T b\} \widetilde{f}(ts)\, dt \\
&= (2\pi)^{-d} \int_{\mathbf{R}^d} L(h\|\omega\|) \exp\{-ib^T \omega\} \widetilde{f}(\omega)\, d\omega.
\end{aligned}$$

This implies:

$$A_h Rf(b) - f(b) = (2\pi)^{-d} \int_{\mathbf{R}^d} \left[L(h\|\omega\|) - 1\right] \widetilde{f}(\omega) \exp\{-ib^T \omega\}\, d\omega. \tag{27}$$

We have that

$$|L(\|h\omega\|) - 1| \le \begin{cases} \|h\omega\|^s, & \text{when } \|h\omega\| \le 1, \\ 1, & \text{when } \|h\omega\| \ge 1. \end{cases}$$

This gives $|L(\|h\omega\|) - 1| \le \|h\omega\|^s$ and thus,

$$\begin{aligned}
\|A_h Rf(b) - f\|_2^2 &= (2\pi)^{-d} \int_{\mathbf{R}^d} \left[L(h\|\omega\|) - 1\right]^2 \left|\widetilde{f}(\omega)\right|^2 d\omega. \\
&\le (2\pi)^{-d} \int_{\mathbf{R}^d} \|h\omega\|^{2s} \left|\widetilde{f}_\beta(\omega)\right|^2 d\omega \\
&\le h^{2s} \rho_s(f)^2.
\end{aligned}$$

We have proved the lemma. □

We now prove a bound for the variance of $\bar{f}_\beta$.

Lemma 7 *Under the assumptions of Theorem 1 it holds that*

$$\int_{\mathbf{R}^d} \mathrm{Var}\left(\bar{f}_\beta\right) \le C n^{-1} h^{-2d+1},$$

for a positive constant C.

Proof. We have that for $b \in \mathbf{R}^d$ (see also (26)),

$$\begin{aligned}
\mathrm{Var}\left(\bar{f}_\beta(b)\right) &\le n^{-2} \sum_{i=1}^{n} \mathbb{E}\left(\frac{1}{f_S(S_i)^2}\, \mathbb{E}\left[K_h\left(S_i^T b - U_i\right)^2 \mid S_i\right]\right) \\
&= n^{-1} \int_{\mathbf{S}_{d-1}} d\mu(s)\, \frac{1}{f_S(s)} \int_{-\infty}^{\infty} K_h^2(s^T b - u) Rf_\beta(s,u)\, du
\end{aligned}$$

$$
\begin{aligned}
&\leq n^{-1}C_S \mathrm{mes}(A) C_\beta \mu(\mathbf{S}_{d-1}) \int_{-\infty}^{\infty} K_h^2(u)\, du \\
&= n^{-1}C_S \mathrm{mes}(A) C_\beta \mu(\mathbf{S}_{d-1}) \\
&\quad (2\pi)^{-d}\frac{1}{4}(2\pi)^{-2d+2}\int_{-\infty}^{\infty} |t|^{2(d-1)} L^2(h|t|)\, dt,
\end{aligned}
$$

where $\mathrm{mes}(A)$ is the Lebesgue measure of A: $\mathrm{mes}(A) = \int_A dx$, and as above $C_S^{-1} = \inf_s f_S(s)$ and $C_\beta = \sup_{s,u} Rf_\beta(s,u)$. We have proved the lemma. □

Proof of Theorem 2 and Remark 6

Asymptotic normality of $\hat{f}_\beta$ immediately follows from the fact that

$$
n^{s/(2s+2d-1)} n^{-1} \sup_u |K_h(u)| = O\left(n^{(-s-d+1)/(2s+2d-1)}\right) = o(1).
$$

The variance $\sigma_n(b)^2$ can be calculated as in Lemma 7. The bound on $\sigma_n(b)^2$ follows by trivial calculations. The bias term $bias_n(b)$ can be calculated by using (27). Its expansion in Remark 6 follows from

$$
\begin{aligned}
&(2\pi)^{-d}\int_{\mathbf{R}^d} [L(h\|\omega\|) - 1]\, \widetilde{f}_\beta(\omega) \exp\{-ib^T\omega\}\, d\omega \\
&= (2\pi)^{-d}\int_{\mathbf{R}^d} (h\|\omega\|)^2 \widetilde{f}_\beta(\omega) \exp\{-ib^T\omega\}\, d\omega \\
&\quad -(2\pi)^{-d}\int_{(h\|\omega\|)^2\geq 1} \left[1 + (h\|\omega\|)^2\right] \widetilde{f}_\beta(\omega) \exp\{-ib^T\omega\}\, d\omega.
\end{aligned}
$$

The first summand is equal to

$$
h^2 \sum_{j=1}^d \frac{\partial^2}{(\partial b_j)^2} f_\beta(b).
$$

The second summand can be absolutely bounded by

$$
\begin{aligned}
&(2\pi)^{-d}\int_{(h\|\omega\|)^2\geq 1} \left[1 + (h\|\omega\|)^2\right] \left|\widetilde{f}_\beta(\omega)\right| d\omega \\
&\leq h^2 (2\pi)^{-d} \int_{\|\omega\|^2 \geq h^{-2}} 2\|\omega\|^2 \left|\widetilde{f}_\beta(\omega)\right| d\omega.
\end{aligned}
$$

This of order $o(h^2)$ because we had assumed that $\|\omega\|^2 \left|\widetilde{f}_\beta(\omega)\right|$ is integrable. □

Proof of Theorem 3

We shortly discuss the bias of the modified estimator $\bar{f}_\beta^\tau$, where

$$
\bar{f}_\beta^\tau(b) = \frac{1}{n}\sum_{i=1}^n \frac{1}{f_S(S_i)} K_h\left(S_i^T b - U_i\right) I_{\mathbf{S}_{d-1}\setminus C_\tau}(S_i), \qquad b \in \mathbf{R}^d.
$$

The variance can be analyzed as in Lemma 7 and for the remaining proof of Theorem 3 one can proceed as in the proof of Theorem 1.

Lemma 8 *Under the assumptions of Theorem 3 it holds that*

$$\left\|\mathbb{E}\bar{f}_\beta^\tau - f_\beta\right\|_2^2 \le 2\rho_2(f_\beta)^2 h^4 + 2\|\nu_{h,\tau}\|_2^2,$$

where C_τ is the band defined in (13) and where D_τ has been defined in the statement of Theorem 3.

Proof. We have that for $z \in \mathbf{R}^d$,

$$\begin{aligned}
&\mathbb{E}\bar{f}_\beta^\tau(z)\\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\frac{1}{f_S(S_i)}\,\mathbb{E}\left[K_h\left(S_i^T z - U_i\right) \mid S_i\right] I_{\mathbf{S}_{d-1}\setminus C_\tau}(S_i)\right)\\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\frac{1}{f_S(S_i)}\int_{-\infty}^{\infty} K_h(S_i^T z - u) Rf_\beta(S_i, u)\, du\, I_{\mathbf{S}_{d-1}\setminus C_\tau}(S_i)\right)\\
&= B_{h,\tau} Rf_\beta(z),
\end{aligned}$$

where

$$(B_{h,\tau} g)(z) = \int_{\mathbf{S}_{d-1}\setminus C_\tau} d\mu(s) \int_{-\infty}^{\infty} K_h(s^T z - u) g(s,u)\, du, \qquad z \in \mathbf{R}^d$$

for functions $g : \mathbf{S}_{d-1} \times \mathbf{R} \to \mathbf{R}$. Thus,

$$\begin{aligned}
B_{h,\tau} Rf_\beta(z) &= 2(2\pi)^{-1}\frac{1}{2}(2\pi)^{-d+1}\int_{\mathbf{S}_{d-1}\setminus C_\tau} d\mu(s)\int_0^\infty t^{d-1} L(ht)\exp\{-its^T\omega\}\widetilde{f}_\beta(ts)\, dt\\
&= (2\pi)^{-d}\int_{\mathbf{R}^d\setminus D_\tau} L(h\|\omega\|)\exp\{-iz^T\omega\}\widetilde{f}_\beta(\omega)\, d\omega\\
&= \nu_{h,\tau}(z) + A_h Rf_\beta(z),
\end{aligned}$$

where A_h is defined as in (4). Proceeding as in the proof of Lemma 5 we get the statement of the lemma. □

References

Abramovich, F. and Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85** 115-129.

Bajari, P., Fox, J.T., Kim, K. and Ryan, S. (2007). Identification and estimation of random utility models, *Working Paper,* University of Chicago.

Beran, R. and Feuerverger, A. and Hall, P. (1996). On Nonparametric Estimation of Intercept and Slope Distributions in Random Coefficient Regression. *Ann. Statist.* **24** 2569-2692.

Beran, R. and Hall, P. (1992). Estimating Coefficient Distributions in Random Coefficient Regressions. *Ann. Statist.* **20**, 1970-1984.

Borak, S., Härdle, W., Mammen, E. and Park, B.U. (2007). Time Series Modelling with Semiparametric Factor Dynamics, *Working Paper.*

Briesch, R.A., Chintgunta, P.K. and Matzkin, R.L. (2007). Nonparametric discrete choice models with unobserved heterogeneity, *Working Paper.*

Christopeit, N. and Hoderlein, S. (2006). Local Partitioned Regression, *Econometrica* **74**, 787-817.

Connor, G., Hagmann, M. and Linton, O.B. (2008). Efficient Estimation of a Semiparametric Characteristic-Based Factor Model of Security Returns, *Working Paper.*

Deaton, A. and J. Muellbauer, (1980). An Almost Ideal Demand System, *American Economic Review,* 70, 312-26.

Devroye, L. and L. Györfi, (1985). *Density Estimation: The L^1 View*, Wiley, New York.

Donoho, D. L. (1995). Nonlinear solutions of linear inverse problems by wavelet-vaguelette decomposition. *J. Applied and Comput. Harmonic Anal.* **2** 101-126.

Feuerverger, A. and Vardi, Y. (2000). Positron Emission Tomography and Random Coefficients Regression. *Ann. Inst. Statist. Math.* **52** 123-138.

Foster, A. and J. Hahn (2000). A Consistent Semiparametric Estimation of the Consumer Surplus Distribution, *Economics Letters*, 69, 245-251.

Fox, J.T. and Gandhi, A. (2008). Identifying heterogeneity in economic choice and selection models using mixtures, *Working Paper,* University of Chicago.

Gautier, E. and Kitamura, Y. (2008). Nonparametric estimation in random coefficients binary choice models. *preprints.*

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. I. H. Poincaré*, **38**(6), 907-921.

Hall, P., Watson, G. S., and Cabrera, J. (1987). Kernel density estimation with spherical data. *Biometrika*, **74**, 751-62.

Heckman, J. and Vytlacil, E. (1998). Instrumental Variables Methods for the Correlated Random Coefficients Model: Estimating the Average Return to Schooling when the Return is Correlated with Schooling, *Journal of Human Resources*, 33, 974-987.

Helgason, S. (1999). The Radon Transform. Progress in Mathematics **5**, Birkhäuser, Boston.

Hildreth, C. and Huock, J.P. (1968). Some Estimators for a Linear Model with Random Coefficients, *Journal of the American Statistical Association*, 63, 584 - 92.

Hoderlein, S. (2007). How many consumers are rational, *Working Paper,* University of Mannheim.

Hoderlein, S. and Lewbel, A. (2006). Price Dimension Reduction in Demand Systems With Many Goods, *Working Paper,* University of Mannheim.

Hoderlein, S. and Mammen, E. (2007). Identification of Marginal Effects in Nonseparable Models without Monotonicity, *Econometrica* **75**, 1513-1518.

Hoderlein, S. and Mammen, E. (2008). Identification and Estimation of Local Average Derivatives in Nonseparable Models without Monotonicity, *Working Paper,* University of Mannheim.

Hsiao, C., and Pesaran, H. (2004). Random Coefficient Panel Data Models, *IZA Discussion Papers,* No. 1236.

Ibragimov, I. A. and R. Z. Hasminskii, (1981). *Statistical Estimation: Asymptotic Theory*, Springer. Originally published in Russian in 1979.

Ichimura, H. and Thompson, T.S. (1998), Maximum Likelihood estimation of a binary choice model with random coefficients of unknown distribution, *J. of Econometrics* **86**, 269-295.

Johnstone, I. and B. Silverman (1990), Speed of Estimation in Positron Emission Tomography, *Ann. Statist.* **18**, 251-280

Klemelä, J. (2000). Estimation of densities and derivatives of densities with directional data. *J. Multivariate Anal.* **73** 18-40.

Klemelä, J. and Mammen, E. (2008). Empirical risk minimization in inverse problems. *preprint*

Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image Reconstruction, Lecture Notes in Statistics, 82* Springer, New York.

Lewbel, A., (1999). Consumer Demand Systems and Household Expenditure, in PESARAN, H. and M. WICKENS (Eds.), Handbook of Applied Econometrics, Blackwell Handbooks in Economics.

— , (2001); Demand Systems With and Without Errors, *American Economic Review*, 611-18.

Mammen, E., Linton, O.B. and Nielsen, J.P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, *Annals of Statistics* **27**, 1443-1490.

Mammen, E., Støve, B., and Tjøstheim, D. (2008). Nonparametric additive models for panels of time series, *Econometric Theory*, forthcoming.

Matzkin, R.L. (2007). Heterogeneous choice, *Working Paper.*

Natterer, F. (2001). The Mathematics of Computerized Tomography, SIAM Classics in Applied Mathematics, Vol. 32.

Stute, W. (1984). The oscillation behaviour of empirical processes: the multivariate case. *Ann. Probab.* **12**(2) 361-379.

Swamy, P.A.V.B. (1970). Efficient Inference in a Random Coefficient Model, *Econometrica*, 38(2), pages 311-23.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge.

A Convergence rates for the spherical kernel estimator

We derive the rates of convergence in the sup-norm of the spherical kernel density estimator $\hat{f}_S$, defined in (9). In the Euclidean case corresponding results have been proved by Ibragimov and

Hasminskii (1981) and Devroye and Györfi (1985). We make the bias-variance decomposition:

$$\sup_{\xi\in\mathbf{S}_{d-1}}\left|\hat{f}_S(\xi)-f_S(\xi)\right|\le\sup_{\xi\in\mathbf{S}_{d-1}}\left|E\hat{f}_S(\xi)-f_S(\xi)\right|+\sup_{\xi\in\mathbf{S}_{d-1}}\left|\hat{f}_S(\xi)-E\hat{f}_S(\xi)\right|,$$

where $f_S:\mathbf{S}_{d-1}\to\mathbf{R}$ is the unknown true density of an i.i.d. sample $S_1,\ldots,S_n\in\mathbf{S}_{d-1}$ on the sphere $\mathbf{S}_{d-1}$.

We can write

$$\sup_{\xi\in\mathbf{S}_{d-1}}\left|\hat{f}_S(\xi)-E\hat{f}_S(\xi)\right|=\sup_{h\in\mathcal{H}}|\nu_n(h)|,$$

where $\nu_n(h)$ is the centered empirical operator

$$\nu_n(h)=\frac{1}{n}\sum_{i=1}^{n}(h(S_i)-Eh(S_i))$$

and

$$\mathcal{H}=\left\{h(u)=c(g)G\left(g^{-2}(1-s^Tu)\right):s\in\mathbf{S}_{d-1}\right\}.$$

Using the same arguments as in the proof of Theorem 3.1 in Stute (1984) we get that

$$\sup_{h\in\mathcal{H}}|\nu_n(h)|=O_p\left(\sqrt{\frac{\log g^{-1}}{ng^{d-1}}}\right). \tag{28}$$

See also Giné and Guillou (2002) for more recent results. Stute (1984) assumes that the density is bounded away from zero on its compact support but Giné and Guillou (2002) assume more generally that $\mathcal{H}$ is a VC-class. A sufficient condition for the kernel is continuity and finite variation.

Turning to the bias term, we can prove that

$$\sup_{\xi\in\mathbf{S}_{d-1}}\left|E\hat{f}_S(\xi)-f_S(\xi)\right|=O(g^\sigma) \tag{29}$$

where $\sigma\ge 2$ is even. The proof is similar to the proof of Theorem 3.7 in Klemelä (2000). See also Hall, Watson, and Cabrera (1987), who consider the case $\sigma=2$. Theorem 3.7 in Klemelä (2000) considers L_p risk for $1\le p<\infty$, but the proof is based on a pointwise expansion of a convolution, and thus it can be applied also for the L_∞-norm. The smoothness assumptions in Theorem 3.7 of Klemelä (2000) are stated in terms of the iterated Laplacian of f_S. In addition, one needs certain assumptions on the kernel G. Specifically, we need to assume that G is a kernel of order σ. Denote $\alpha_i(G)=\int_0^\infty t^{(i+d-2)/2}G(t)\,dt$, where $i\ge 0$. We need to choose $G:[0,\infty)\to\mathbf{R}$ in such a way that $\alpha_i(|G|)<\infty$ for $i=0,\sigma$, $\alpha_0(G)\ne 0$, and $\int_0^{g^{-2}}t^{(2i+d-2)/2}G(t)\,dt=o(g^{\sigma-2i})$ for $i=1,\ldots,\sigma/2-1$. Note that unlike in the Euclidean case we do not need to assume that the "odd moments" of the kernel vanishes, because due to the properties of derivation on the sphere, the derivatives of odd order vanish. Choosing

$$g=n^{-1/(2\sigma+d-1)}$$

and combining (28) and (29) we get

$$\sup_{\xi \in \mathbf{S}_{d-1}} \left| \hat{f}_S(\xi) - f_S(\xi) \right| = O_P \left(\sqrt{\log n} n^{-\sigma/(2\sigma+d-1)} \right).$$

In our setting the spherical density f_S is the density of $S = X/\|X\|$, where $X \in \mathbf{R}^d$ is a Euclidean random vector. The density f_S is obtained by radial integration from the density f_X. Thus, the smoothness of f_X implies the smoothness of f_S. For example, if X is Gaussian, then f_S has the required smoothness for any even σ.