A Panel Data Estimator for the Distribution and Quantiles of Marginal Effects in Nonlinear Structural Models with an Application to the Demand for Junk Food

Joe Cooprider Stefan Hoderlein Alexander Meister
Boston College* Emory University† University of Rostock‡

July 19, 2022

Abstract

In this paper, we propose a framework to estimate the distribution of marginal effects in a general class of structural models that allow for very general nonlinearities, high dimensional heterogeneity, and unrestricted correlation between the persistent components of this heterogeneity and all covariates. The main idea is to form a derivative dependent variable using two periods of the panel, and use differences in outcome variables of nearby subpopulations to obtain the distribution of marginal effects. We establish constructive nonparametric identification for the population of "stayers" (Chamberlain (1982)), and show generic non-identification for the "movers". We propose natural semiparametric sample counterparts estimators, and establish that they achieve the optimal (minimax) rate. Moreover, we analyze their behavior through a Monte-Carlo study, and showcase the importance of allowing for nonlinearities and correlated heterogeneity through an application to demand for junk food. In this application, we establish profound differences in marginal income effects between poor and wealthy households, which may partially explain health issues faced by the less privileged population.

Keywords: Heterogeneity, Nonparametric, Identification, Random Coefficients.

*Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA, Email: coopridj@bc.edu.

†Department of Economics, Emory University, 1602 Fishburne Dr., Atlanta, GA 30322, USA, Email: stefan_hoderlein@yahoo.com.

‡Institute for Mathematics, University of Rostock, 18051 Rostock, Germany, Email: alexander.meister@uni-rostock.de.

1 Introduction

Motivation. It is commonplace that panel data allows researchers to model the impact of correlated unobserved individual specific heterogeneity, as is illustrated by the fixed effects approach and generalizations to linear random coefficients models (Chamberlain (1982), Wooldridge (2005), Graham and Powell (2012), Arellano and Bonhomme (2012)). A particular challenge, however, arises with the presence of nonlinearities in many microeconometric models, even in models that do not feature a limited dependent variable. This situation arises frequently in economics. While economic models often exhibit qualitative restrictions stemming from constrained optimization of rational agents, e.g., convexity or monotonicity, they feature linearity or additivity only in exceptional cases. In consumer demand which motivates the application of this paper, this has led to the rise and popularity of nonlinear models (e.g., the QUAIDS, see Banks, Blundell and Lewbel (1997)), and nonparametric and nonseparable models in general, because they capture important aspects of the data that are otherwise missed.

But while it is now commonly found that microeconomic relationships should allow for nonlinearities on the individual level, there is even more experimental and observational evidence that individuals differ across the population in ways that are not entirely captured by observable variables. There are basically two ways to deal with this complex unobserved heterogeneity: considering average effects, or recovering the distribution of heterogeneity parameters. The former is easier to obtain than the latter, and frequently less stringent assumptions have to be imposed for its recovery. As a case in point, in a cross-section setup, average treatment effects are identified under general conditions, while to recover heterogeneous functions or parameters one has to, for instance, impose monotonicity of the structural function in a scalar unobservable (see, e.g., Matzkin (2003)), or a linear random coefficients structure (Hoderlein, Klemelä and Mammen (2011)). Moreover, when covariates are endogenous, further restrictions are necessary (see Imbens and Newey (2009), Kasy (2011), or Hoderlein, Holzmann and Meister (2017)).

This paper establishes the strength of panel data to allow recovery of the distribution of heterogeneous nonparametric marginal effects, even if covariates are correlated and the time span considered is very short. More precisely, we show that the distribution of marginal effects of a general class of structural models is nonparametrically identified. This allows for arbitrary dependence between the time-invariant unobservable and the covariates of interest, provided as little as two observations are available for the individuals. Formally, we consider a nonparametric and heterogeneous model of the form

$$Y_{k,t} \;=\; \phi(X_{k,t}, A_k) + U_{k,t}\,, \qquad k = 1, \ldots, n;\; t = 1, \ldots, T\,, \tag{1.1}$$

where $Y_{k,t} \in \mathcal{Y} \subseteq \mathbb{R}$, and $X_{k,t} \in \mathcal{X} \subseteq \mathbb{R}^J$ are observable variables, and $A_k \in \mathcal{A} \subseteq \mathbb{R}^\infty$ and $U_{k,t} \in \mathcal{U} \subseteq \mathbb{R}$ are unobserved. Note that in this model, the dimension of A_k is not

restricted, and the structural function ϕ is assumed to be smooth in the sense of being twice continuously differentiable in x_j for all $j = 1, .., J$, with bounded second derivatives, but is otherwise unrestricted. Moreover, we allow for arbitrary dependence (correlation) between any element of A_k and any element of $X_{k,t}$ for any k, t. These facts make our model different from the models of Altonji and Matzkin (2005) and Evdokimov (2010) with which it shares structural similarities, and make it a similar spirited generalization of the random coefficient models of Arellano and Bonhomme (2012) and Graham and Powell (2012).

Main Result. The main result in this paper establishes nonparametric identification of the (marginal) distribution of marginal effects $\partial_{x_j}\phi(x, A)$, for $j = 1, \ldots, J$, and all $x \in \mathcal{X}$, even with many regressors and only two time periods (i.e., $T = 2$), conditionally on $X_{k,1} = X_{k,2} = \cdots = X_{k,T}$. If $T \geq J + 1$, we also show that the joint distribution of all marginal effects, i.e., $\nabla_x\phi(x, A) = (\partial_{x_1}\phi(x, A), \ldots, \partial_{x_J}\phi(x, A))'$ is identified, for all $x \in \mathcal{X}$, see Remark 1. As a corollary, we obtain identification of objects like the average structural marginal effect, as well as the variance of marginal effects. An important limitation of our analysis is that we can only make statements for the population for which $X_{k,1} = X_{k,2} = ... = X_{k,T}$, i.e., we are only identifying the distribution $f_{\nabla_x\phi(x,A)|X_1-X_2=0,X_1=x}$ for the "stayers" (in the sense of Chamberlain (1982)), and for $x \in \mathcal{X}$. To fix ideas, in our demand application this will be the population for which income and prices stay approximately constant. As an important contribution, we establish that this limitation is not an accident of the identification approach taken, but a consequence of a profound non-identification result for nonlinear marginal effects outside of the stayers sub-population. The intuition behind this result is as follows: Suppose the true model is a J-th order polynomial in a scalar $X_{k,t}$ with random coefficients on every term. Then, the number of time periods acts as limiting factor for our ability to learn about this complex models - if J exceeds $T - 1$, there is generic non-identification (i.e., with $T = 2$, at most a linear random coefficients model is identified for $x_2 \neq x_1$, i.e., outside the population of stayers).

Intuition for identification. The essential idea which underlies this strong constructive identification result for the stayers is as follows: Unlike with repeated cross section data, we utilize the fact that we observe individuals repeatedly in a panel to form a derivative dependent variable $\partial Y/\partial X$. Specifically, by considering individuals whose $X_{k,2}$ is close to their $X_{k,1}$ we construct a sample counterpart to the limiting process when taking derivatives. A complication arises because we have to correct for the transitory error $U_{k,t}$. This is done by considering people who have exactly $X_{k,2} = X_{k,1} = x$ for every $x \in \mathcal{X}$ (or, in the sample, almost exactly), because for these individual all changes in $Y_{k,t}$ can be attributed to changes in $U_{k,t}$.

More formally, in the case of a scalar $X_{k,t}$ (dropping the cross sectional index k), we take first differences across time to obtain

$$Y_{t+1} - Y_t = \phi(X_{t+1}, A) - \phi(X_t, A) + U_{t+1} - U_t\,, \qquad ,t = 1, \ldots, T-1. \tag{1.2}$$

The first step uses then the observation that conditioning on $X_t = X_{t+1} = x$, for any $x \in \mathcal{X}$, removes the first two terms on the right hand side. Thus, the characteristic function of $Y_{t+1} - Y_t$ conditional on $X_t = X_{t+1} = x$ identifies the characteristic function (ChF) of $U_{t+1} - U_t$.[1]

In a second step, we use the fact that, conditional on $X_t = x$, $X_{t+1} = x + h$, the same first difference equals

$$Y_{t+1} - Y_t \;=\; \partial_x \phi(x, A)h \;+ R_2(x, h, A) + U_{t+1} - U_t \,, \qquad ,t = 1, \ldots, T-1, \tag{1.3}$$

where R_2 is a remainder term. Dividing by h, and establishing the fact that $R_2/h \to 0$ as $h \to 0$, we hence have for small h (and conditional on $X_t = x$, $X_{t+1} = x + h$) approximately

$$\frac{Y_{t+1} - Y_t}{h} \;\approx\; \partial_x \phi(x, A) + \frac{U_{t+1} - U_t}{h}.$$

We then use the fact that we have obtained the ChF of $\frac{U_{t+1}-U_t}{h}$, conditional on $X_t = x, \Delta X = X_{t+1} - X_t = 0$ in the first step. By adding the assumption that $\Delta X \perp \Delta U | X_t$, we are thus able to identify the ChF of $\frac{U_{t+1}-U_t}{h}$, conditional on $X_t = x, \Delta X = h$ as well, as the conditioning on different values of ΔX does not impact the ChF of $\Delta U | X_t$. Under the additional assumption that $A \perp \Delta U | X_t$, we can therefore apply deconvolution to obtain the ChF of $\partial_x \phi(x, A)$, conditional on $X_t = x, \Delta X = 0$.

Importantly, note that we do *not* assume that $U_t \perp X_t$, but rather that all correlation between the ΔU and the ΔX innovations runs through X_t. This allows, e.g., for a specification of the $U_s = \sigma(X_1, ..., X_{t-1})V_s$ for all $s \geq t$ and $V_s \perp X_1, ...X_T, ,$ e.g., in a two period panel we allow for heteroskedasticity of U_1 and U_2 as a function of the level X_1, but not of the innovation ΔX. Similar "timing" assumptions are common in the empirical IO literature. In, e.g., Olley and Pakes (1996), ΔU are shocks to productivity and ΔX are changes to (production) factor input, and the main identifying assumption is that X_t only responds to past shocks in productivity, but not contemporaneous ones because of a time lag in the response of the firm's to shocks in productivity[2]. In fact, in the case of a J-vector X_t, if interest centers on one variable only (e.g., labor input X_{1t}), U_t and ΔU may be arbitrarily correlated with $X_{2t}, ...X_{Jt}$. Moreover, as we sketch in the appendix the model is identified under even weaker conditions that permit for contemporaneous heteroskedasticity in the variable of interest, i.e.,$U_t = \sigma(X_t)V_t$, for instance for the productivity shock to have different effects according to the level of labor input or observable characteristics of the firm like size. However, it would be more difficult to build an estimator on the identification argument, and we therefore do not elaborate on this strategy. Finally, the deconvolution step requires in addition that $A \perp \Delta U | X$, which is again different from full independence between A and U_t.

[1]Note that there is a one-to-one relationship between the ChF and the PDF of continuous random variable.

[2]In this example, A would be a collection time invariant unobservables which may be correlated with factor inputs and observable characteristics X).

The baseline specification allows us to identify the marginal distribution of every marginal effect needing only two time periods. However, its driving force is the time invariance of the correlated unobservable A, and the additive separability of the transitory error U_t, together with the timing assumption. With more time periods, we may relax these assumptions and allow for the structural function ϕ to change over time under restrictions on the way time enters which may be weakened as T becomes large. Several other extensions are briefly discussed in this paper: The approach can be augmented to allow for discrete covariates; however, the effect of interest has to be on a continuous variable. More generally, we may control for additional covariates through a semiparametric specification. Finally, we conjecture that the approach can be extended to a discrete dependent variable if one exogenous regressor with large support is available, similar to Honoré and Lewbel (2002).

Translation to finite sample estimation. In the (finite) sample, we thus use the difference between people who are on or very near the diagonal (as approximation of $\Delta X = 0$) from those who are near, but not quite as near, to the diagonal (as approximation of $\Delta X = h$). Similar to regression discontinuity models we discard the rest of the sample as it does not contain (point) identifying information. Using this subsample, the difference in the distribution of $Y_{k,t}$ is then due to the (heterogeneous) causal marginal effect of $X_{k,t}$. This effect depends, obviously, in general on the position $X_t = X_{t+1} = x$ we consider; by letting the position x vary, we obtain an arbitrary nonlinear relationship. Fig. 1 illustrates the data used in the sample. Finally, that this works only near the diagonal (i.e., only for the stayers) is due to the fact that higher order terms in the derivative approximation only disappear in this neighborhood, which we establish formally in a non-identification section.

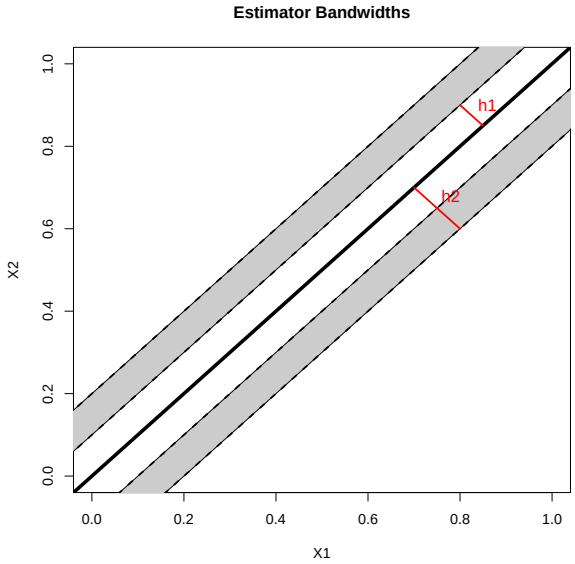

Figure 1: The shaded region is the region we will use, which is more than h_1 away from where $X_1 = X_2$ but less than h_2 away from where $X_1 = X_2$.

When it comes to estimation, we follow a semiparametric route. That is, we assume that the distribution of marginal effects follows a known parametric distribution governed by a finite parameter $\theta(x)$ which depends on the position $X_1 = X_2 = x$ at which we evaluate the conditional distribution. As such, our approach can be described as conditionally parametric. The advantage of such a procedure is as follows: Since our identification argument and the associated sample counterparts estimator is based on the (conditional) ChF, we avoid having to invert these estimators to obtain the (conditional) PDF. In the sample, this inversion step comes at the cost of having to pick an additional regularization parameter. Moreover, since one of the main objectives of our approach is to get an estimator for the quantiles of marginal effects as well, we avoid having to add another cumbersome inversion. Instead, the conditional parametric approach obtains all of these quantities: the conditional characteristic function, density, as well as the quantiles in one convenient step. Moreover, the ChF need not be observed for every value of the argument (s, say).

The core principle employed in our estimator is a minimum contrast (distance) step. We first form the sample counterpart to the identified nonparametric characteristic function for every value of $X_1 = X_2 = x$, and then pick the the parameter $\theta(x)$ that minimizes the contrast (distance) between the approximating parametric specification and this object. For this estimator, we establish the (optimal) minimax rate, and establish that our estimator achieves this rate. The rate is governed by the dimensionality of X and the fact that we work with the set $X_1 = X_2 = x$. If there is no $U_{k,t}$ and X is scalar, the rate is equivalent to a two dimensional nonparametric regression. Having, in addition, a $U_{k,t}$ that follows an ordinary smooth distribution slows the convergence rate down by the expected factor, α, due to the added deconvolution step in removing the influence of $U_{k,t}$.

Importantly, this paper contains an application to consumer demand for junk food. Because of the relationship to obesity and other adverse health effects, this is a question of obvious importance for the society (see also the short literature review in the applied section). A key concern is that "poor" households - which we define to be households with low total expenditure for goods that Nielsen scanner data tracks - spend marginally more on junk food than wealthy, high income households. This means that a model that forces all households to have the same "income" and price elasticities, i.e., a linear random coefficients model, is not able to capture this important feature. Similarly, we want to control for unobserved factors that are correlated with poverty, e.g., education levels, in particular regarding nutrition, and hence it is imperative to allow for the unobservables to be correlated. Therefore, we feel that our approach, which allows for nonlinearities, high dimensional heterogeneity, and complicated correlation patterns, is particularly well suited for this application.

When applying our approach to the Nielsen Homescan data, we indeed find evidence of the aforementioned nonlinearities. Indeed, for every dollar spent on Nielsen products, poor households seem to consume twice as much junk food on average compared to wealthy households,

even implicitly controlling for persistent correlated effects like education. Moreover, there also seems to be more heterogeneity within poor households (compared to wealthy ones), perhaps a function of the larger degree of addiction to an unhealthy lifestyle of at least parts of this subpopulation. It is interesting to muse about the reason for the significant correlation between expenditure levels and marginal effects, even after controlling for fixed factors. We also find very reasonable price elasticities that increase in the own price. Since we use a bundle of goods and Stone-Lewbel prices, we feel that this reflects heterogeneity in the composition of junk food. The more high level it is, the higher the price and the more elastic demand. More details can be found in the section on the application below.

Related Literature: Analyzing nonlinear panel data models has a long tradition, dating back to the conditional ML approach by Rasch (1960, 1961); see also Andersen (1970) and Chamberlain (1982, 1984) for models with non-additive individual heterogeneity. Nonlinear parametric panel data models have frequently been analyzed. For an overview of work related to discrete choice models, see Arellano (2003). Most closely related to our work is that of Graham and Powell (2012), and Arellano and Bonhomme (2013), who consider estimation of moments and the distribution of random coefficients in a linear correlated coefficient panel data model. Compared to this line of work, we allow for the structural model to be arbitrarily nonlinear. Chamberlain (2010) discusses the identification of the dynamic panel data binary choice model, and why the logistic distribution assumption is required for identification of β_o, unless one is willing to assume unbounded support for one of the regressors, as is the case in Manski (1987). Recent work on multinomial choice in panels that is related to ours includes Chernozhukov, Fernandez Val and Newey (2019) which features a non-identification counterexample outside of the stayers subpopulation as well. For other nonlinear fixed effects models, see also Hausman, Hall, and Griliches (1984) for panel count data and Honoré (1992) for panel censored regression. Like all of this work, our approach assumes a fixed number of time periods. Indeed, it is one of the appealing features of our approach that we only require $T = 2$.

All of the work just described is concerned with a specific semiparametric model, e.g., the dynamic binary choice model. Approaches that are closer in spirit to our work are those of Chernozhukov, Fernandez-Val, Hahn, and Newey (2014), who consider discrete variation, whereas we consider derivatives, and Graham and Powell (2012), who focus on a linear heterogeneous population (i.e., the structure is linear in the coefficients, with coefficients that vary across the population) and not on a fully nonseparable structure. Other than the differences mentioned above, Graham and Powell (2012) also require (at least) as many time periods as regressors plus one, while we require only two time periods, even with a large number of regressors. Less closely related is the work on the correlated random coefficients models in panel data, see in particular Wooldridge (2005) and Murtazashvili and Wooldridge (2008). This line of work studies the linear random coefficients model as well, but imposes restriction on the correlation between

time invariant individual specific effects and covariates of interest. In contrast, our approach allows for unobserved heterogeneity to enter nonlinearly and does not limit its correlation with the covariates of interest.

Finally, related is also the literature on nonseparable models using panel data, in particular Altonji and Matzkin (2005), Evdokimov (2010), Hoderlein and White (2012), Chernozhuov, Fernandez-Val, Hoderlein, Holzmann and Newey (2015) and Fernandez Val, Freeman and Weidner (2021). Unlike our paper, Altonji and Matzkin (2005) impose constraints on the correlation between A_k and the $X_{k,t}$ process, but are more general in the structural function ϕ in that they allow interaction between the transitory error $U_{k,t}$ and the other variables, and focus on averages. Evdokimov (2010) also imposes additivity of the error $U_{k,t}$, but assumes that A_k is a scalar and ϕ is monotonic in this scalar, while we allow for many non-monotonic unobservable factors. Hoderlein and White (2012) and Chernozhuov, Fernandez-Val, Hoderlein, Holzmann and Newey (2015) again admit a more general structural function ϕ (as in Altonji and Matzkin (2005)), but are only able to identify averages of the marginal effects, even though Chernozhuov, Fernandez-Val, Hoderlein, Holzmann and Newey (2015) use distributional information. Instead, in this paper we use a deconvolution step to purge the model from the influence of $U_{k,t}$. This also allows to impose different, and arguably weaker, assumptions on the $U_{k,t}$.process. In particular, we do not require the stationarity assumption in their papers (see also Manski (1987)). Finally, Fernandez Val, Freeman and Weidner (2021) consider a nonseparable model that is more general than ours, but focus again on average and quantile effects and discuss estimation of their model through a low rank factor structure approximation only in the case of binary (treatment effect) regressors, and do not exploit continuity of random variables.

Outline of the Paper: Section 2 introduces the model and the precise assumptions we require. In Section 3, we present the general non-identification result for arbitrary values $x_2 \neq x_1$, which motivates our focus on the set of stayers. Section 4 then presents the main constructive nonparametric identification result and discusses extensions. Section 5 establishes the asymptotic lower bound for any estimator under this scenario. In Section 6, we introduce our conditional parametric estimator and the modeling assumptions, establish an upper bound under these conditions, and show that our estimator achieves the minimax rate. Section 7 analyzes the finite-sample performance of our estimators using several example of nonlinear heterogeneous DGPs. Section 8 discusses the application to consumer demand for junk food. The final section contains a summary and concluding remarks.

2 The Model: Basic Structure and Main Assumptions

We consider the panel data model

$$Y_{k,t} \;=\; \phi(X_{k,t}, A_k) + U_{k,t}\,, \qquad k = 1, \ldots, n;\; , t = 1, \ldots, T\,, \tag{2.1}$$

where all $X_{k,t}$ and $Y_{k,t}$ are observed. Therein, the random vectors $(X_{k,t}, A_k, U_{k,t})_{t=1,\dots,T}$ are i.i.d. (i.e. independent copies) for all $k = 1, \dots, n$. Therefore, when addressing identification issues, we omit the index k in the notation of all random variables. We assume

(A1) The random vector $X := (X_1, \dots, X_T)$ has a T-dimensional Lebesgue density.

Our goal is to identify the conditional distribution $\mathcal{L}(Z_j \mid X)$ of the random variable

$$Z_j \;:=\; \frac{\partial \phi}{\partial x}(x, A) \mid_{x=X_j},$$

given X. From a famous result in probability theory (e.g. p. 439, Theorem 33.3, Billingsley (1995)), we learn that there exists a function ζ_j from the domain $\mathbb{R}^T$ to the set of all probability measures on the Borel σ-field $\mathfrak{B}(\mathbb{R})$ of $\mathbb{R}$ such that

$$\big\{\zeta_j(X)\big\}(B) \;=\; P[Z_j \in B \mid X], \text{ a.s.},$$

for all elements B of the Borel σ-field $\mathfrak{B}(\mathbb{R})$. This equation, however, does not determine the value of the mapping ζ_j at any fixed $x \in \mathbb{R}^T$. In particular, the value of ζ_j at one singular $x \in \mathbb{R}^T$ can be changed without switching to an observationally non-equivalent model due to condition (A1). As a consequence, identification and estimation of $\zeta_j(x)$, for any specific value $x \in \mathbb{R}^T$, is impossible unless continuity conditions are assumed such as

(A2) There exists a function ζ_j on the domain $\mathbb{R}^T$ to the set of all probability measures on $\mathfrak{B}(\mathbb{R})$ which is continuous with respect to the Fourier distance on its codomain; and satisfies

$$\big\{\zeta_j(X)\big\}(B) \;=\; P[Z_j \in B \mid X], \text{ a.s.},$$

for all $B \in \mathfrak{B}(\mathbb{R})$.[3]

Condition (A2) resembles the usual constraints in the setting of standard nonparametric regression where the regression function is required to be continuous under continuously distributed covariates in order to attain pointwise consistency at a fixed site. The following lemma shows that $\zeta_j(x)$ is uniquely determined for each x in the support of X.

[3]Here, the Fourier distance between two probability measures P and Q on $\mathfrak{B}(\mathbb{R})$ is defined by

$$\mathcal{F}(P, Q) \;:=\; \sup_{s \in \mathbb{R}} \big| P^{ft}(s) - Q^{ft}(s) \big|, \tag{2.2}$$

where $P^{ft}(s) := \int \exp(isx) dP(x)$ denotes the Fourier transform of P. Note that the total variation distance $\mathrm{TV}(P, Q)$ between P and Q, i.e.

$$\mathrm{TV}(P, Q) \;:=\; \sup_{B \in \mathfrak{B}(\mathbb{R})} |P(B) - Q(B)|,$$

dominates the Fourier distance $\mathcal{F}(P, Q)$. The set of all probability measures on $\mathfrak{B}(\mathbb{R})$, equipped with the Fourier distance $\mathcal{F}$, forms a complete metric space thanks to the completeness of the space $C_0(\mathbb{R})$ and Lévy's continuity theorem (e.g. Williams, 1991, section 18.1).

Lemma 2.1. *Assume two functions ζ_j and $\tilde{\zeta}_j$ which satisfy the continuity assumptions imposed on ζ_j in (A2); and*

$$\big\{\zeta_j(X)\big\}(B) \;=\; P[Z_j \in B \mid X] \;=\; \big\{\tilde{\zeta}_j(X)\big\}(B) \;\; a.s., \forall B \in \mathfrak{B}(\mathbb{R})\,.$$

Then the restrictions of ζ_j and $\tilde{\zeta}_j$ to the support $\mathcal{S}_X$ of X coincide.

With respect to the error vector $U := (U_1, \ldots, U_T)$ we provide three alternative assumptions, which we discuss thereafter. Assumption (A3) and (A3') refer to the case of $T = 2$ where $\Delta_U := U_1 - U_2$ and $\Delta\phi := \phi(X_1, A) - \phi(X_2, A)$.

(A3) ΔU and A are conditionally independent given X, and ΔU and X_2 are conditionally independent given X_1,

(A3') $\Delta\phi$ and ΔU are conditionally independent given X; and ΔU and X are independent,

(A3") U and (A, X) are independent.

Assumption (A3"), which is strongest, is used in Section 3, in which we show non-identifiability of ζ_j in a general setting, and in Section 5, in which a lower bound on the attainable convergence rates for any estimator is established. Note that the stronger the conditions the stronger the negative identification result, meaning if non-identification holds already in a restricted class of models it certainly holds more generally. Assumption (A3) which is implied by the stronger (A3") is used in Section 4, in which we provide a positive identification result. That this is a weakening of (A3") is obvious, and the relationship to timing assumptions common in empirical IO has already been discussed in the introduction. Assumption (A3') is applied when we study the properties of our estimator in Section 6.

3 Non-Identification

As outlined above, in this section we establish a general non-identification result for the distribution of random coefficients. Specifically, we focus on the question for which elements x of $\mathcal{S}_X$ the probability measure $\zeta_j(x)$ can be identified from the observed data (X_t, Y_t), $t = 1, \ldots, T$, under the Assumptions (A1), (A2) and (A3"). We provide a negative result for those x whose components are all different from one another; the set of such x will be denoted by $\mathcal{T}_X$ in Lemma 3.2. We proceed by providing a very general class of counterexamples. We consider T time periods, and specify the ϕ function to be the T-th polynomial in x, probably the most natural nonlinear specification. In this specification, A would denote the $T+1$ vector of random coefficients.

We then proceed in two steps. In the first, we show that this model for a specific random coefficient vector $A^{[0]}$ and associated distribution is observationally indistinguishable (i.e.,

generates the same distribution of observables (Y, X)) from a class of transformed model with shifted random vector $A^{[b]}$, where b is the shift parameter. In a second step, we then show that even despite this observational equivalence in terms of observables, indeed the distributions of $A^{[0]}$ and any $A^{[b]}$ is different.

To establish these results, we introduce the notation $p(x) := (1, x^1, \dots, x^T)^\dagger$, and $q(x) := (0, 1, 2x, \dots, Tx^{T-1})^\dagger$ for the polynomial and its derivative. Moreover, the make use of the following useful tool:

Lemma 3.1. *The vectors $p(x_1), \dots, p(x_T), q(x_j)$, for any $j \in \{1, \dots, T\}$, are linearly independent if and only if all $x_1, \dots, x_T$ differ from each other.*

By $H(x)$, $x = (x_1, \dots, x_T)$, we denote the linear hull of $p(x_1), \dots, p(x_T)$. The squared distance between $H(x)$ and $q(x_j)$ is called $\tau_j(x)$.

Lemma 3.2. *The function τ_j is continuous and takes on only strictly positive values on the set $\mathcal{T}_X := \bigcap_{k \neq l} \{x \in \mathbb{R}^T : x_k \neq x_l\}$.*

In order to prove the non-identification result, we may, in addition, assume that the function ϕ and the distribution of the random vector U, as well as the distribution of the covariates X, are known. A similar non-identification result has been shown in Chernozhukov et al. (2019) where apparently two different competing candidates for the function ϕ have been used – in contrast to our approach. Concretely, we impose that

$$\phi(x, A) \;=\; \sum_{t=0}^{T} A_t x^t \,. \tag{3.1}$$

Let $q_j^*(x)$ denote the orthogonal projection of $q(x_j)$ onto the orthogonal complement of $H(x)$ with respect to $\mathbb{R}^{T+1}$ as this notation has already been used in the proof of Lemma 3.2. This lemma also yields $q_j^*(x) \neq 0$ for all $x \in \mathcal{T}_X$ since $|q_j^*|^2 = \tau_j$. Then we are ready to define the random variables

$$A^{[b]} \;:=\; A^{[0]} \;+\; \sqrt{b}\,\delta\, q_j^*(X)\,, \qquad b \geq 0\,, \tag{3.2}$$

where the random variable δ is standard normal; $A^{[0]}$ is an arbitrary $(T+1)$-dimensional random vector; and $(X, A^{[0]})$ and δ are independent. Then

$$\mathcal{L}^{[b]}(A \mid X) = \mathcal{L}(A^{[b]} \mid X)\,, \qquad b \geq 0\,,$$

denote competing candidates for the conditional distribution of A given X.

First step: The conditional characteristic function of $V := \big(\phi(X_1, A), \dots, \phi(X_T, A)\big)$ given

X equals

$$\psi_{V|X}(t) \;=\; E\Big\{ \exp\Big(i\sum_{k=1}^{T} t_k\phi(X_k,A)\Big) \mid X\Big\} \;=\; E\Big\{ \exp\Big(i\sum_{l=0}^{T} A_l \sum_{k=1}^{T} t_k X_k^l\Big) \mid X\Big\}$$
$$= \;\psi_{A|X}\Big(\sum_{k=1}^{T} t_k X_k^0, \ldots, \sum_{k=1}^{T} t_k X_k^T\Big)\,,$$

for all $t \in \mathbb{R}^T$ whenever (3.1) holds true. Hence, for the candidates $\mathcal{L}^{[b]}(A \mid X)$, $b \geq 0$, it holds that

$$\psi_{V|X}^{[b]}(t) \;=\; \exp\Big(-\frac{1}{2}b\Big|\sum_{k=1}^{T} t_k p(X_k)^\dagger q_j^*(X)\Big|^2\Big)\cdot \psi_{A^{[0]}|X}\Big(\sum_{k=1}^{T} t_k p(X_k)\Big)$$
$$= \;\psi_{A^{[0]}|X}\Big(\sum_{k=1}^{T} t_k p(X_k)\Big)\,,$$

for all $t \in \mathbb{R}^T$ and $b \geq 0$ so that the conditional distributions $\mathcal{L}^{[b]}(V \mid X)$ coincide almost surely for all $b \geq 0$. We have used orthogonality of the vectors $p(X_k)$ and $q_j^*(X)$ which follows from the definition of $q_j^*(x)$. Therefore, the distribution of the observed data (X, Y) with $Y := (Y_1, \ldots, Y_T)$, are identical for all candidates ($b \geq 0$) thanks to the independence of U and (A, X), and one is unable to determine the value of b based on the distribution of the observables.

Second step: It remains to be shown that also the conditional distributions which we want to identify do not coincide for different values of b. Due to (3.1) and (3.2) we have $Z_j^{[b]} = (A^{[b]})^\dagger q(X_j)$ so that

$$\zeta_j^{[b]}(x) = \mathcal{L}\big({A^{[0]}}^\dagger q(x_j) \mid X = x\big) * \mathrm{N}(0, b\tau_j^2(x))\,, \tag{3.3}$$

where $*$ denotes convolution. Consider $\mathrm{N}(0,0)$ as the Dirac measure which is concentrated at 0. The corresponding Fourier transform equals

$$\big\{\zeta_j^{[b]}(x)\big\}^{ft}(s) \;=\; \psi_{A^{[0]}|X=x}\big(sq(x_j)\big)\cdot \exp\Big(-\frac{1}{2}bs^2\tau_j^2(x)\Big)\,, \qquad s \in \mathbb{R}\,. \tag{3.4}$$

We impose the Assumption

(A4) The random vector $A^{[0]}$ has a conditional Lebesgue density $f_{A^{[0]}|X=x}$ given $X = x$ for all $x \in \mathbb{R}^T$; moreover, we have that

$$\lim_{y\to x} \mathcal{F}\big(\mathcal{L}(A^{[0]} \mid X = x), \mathcal{L}(A^{[0]} \mid X = y)\big) \;=\; 0\,, \qquad \forall x \in \mathbb{R}^T\,.$$

In Assumption (A4), we have extended the definition of the Fourier distance in (2.2) to

probability measures on $\mathfrak{B}(\mathbb{R}^{T+1})$ in a natural way by the supremum norm distance of the Fourier transforms of both measures. Note that Assumption (A4) is satisfied in particular if $A^{[0]}$ has a Lebesgue density and $A^{[0]}$ and X are independent, which is related to the scenario considered in Evdokimov (2010). The following lemma verifies Assumption (A2) in our setting.

Lemma 3.3. *The functions $\zeta_j^{[b]}$ in (3.3) are continuous for any $b \geq 0$ with respect to the Fourier distance on the codomain under the Assumption (A4).*

Furthermore Lemma 3.2, which guarantees that $\tau_j^2(x) \neq 0$ in (3.3) and (3.4), and the equation (3.3) yield that, for all $b \neq b' > 0$, the probability measures $\zeta_j^{[b]}(x)$ and $\zeta_j^{[b']}(x)$ are different from each other for all $x \in \mathcal{S}_X \cap \mathcal{T}_X$ where we use the following result.

Lemma 3.4. *Let Q be an arbitrary probability measure on $\mathfrak{B}(\mathbb{R})$. Then the equality $Q * N(0,\alpha) = Q * N(0,\alpha')$ implies $\alpha = \alpha'$ for all $\alpha, \alpha' \in [0,\infty)$.*

Thus, we have established the following theorem about non-identification of $\zeta_j(x)$, for all $x \in \mathcal{S}_X \cap \mathcal{T}_X$, i.e., values of x for which $x_1 \neq x_2$, in the model (2.1).

Theorem 1. *In the model (2.1), fix some $j = 1, \ldots, T$; select the function ϕ as in (3.1); and impose Assumptions (A1) and (A3"). Set the random variable A equal to $A^{[b]}$ in (3.2) where the choice of $A^{[0]}$ is only restricted by Assumption (A4). Then the corresponding distributions of the observations (X,Y) coincide for all $b \geq 0$ while Assumption (A2) is satisfied for all $b \geq 0$; and $\zeta_j^{[b]}(x) \neq \zeta_j^{[b']}(x)$ holds true for all $b \neq b'$ and $x \in \mathcal{S}_X \cap \mathcal{T}_X$.*

4 Identification

To establish the fundamental identification result, we proceed again in two steps. As outlined in the introduction, in a first step we use the subpopulation for which $\Delta X = 0$ (the stayers) to obtain the distribution of $\Delta U | X = x$, for any value of x. In a second step, we then use the subpopulation for $\Delta X = h$, for h small, in conjunction with the result from the first step to obtain the desired distribution of marginal effects.

To keep the notation as transparent as possible, we assume that $T = 2$ and $j = 1$. Recall from the non-identification Theorem 1 that the function $\zeta(x)$ cannot be identified from the distribution of the data unless we confine ourselves to $x \in \mathcal{S}_X \backslash \mathcal{T}_X$, which means $\{(x_1, x_2) \in \mathcal{S}_X : x_1 = x_2\}$. In addition to the previous assumptions, we assume that:

(A5) There exists some $\rho > 0$ such that the density f_X of $X = (X_1, X_2)$ is continuous and strictly positive on the strip

$$\mathcal{S}_X^{(\rho)} := \{(x_1, x_2) \in \mathbb{R}^2 : |x_1 - x_2| \leq \rho\}.$$

Under Assumption (A5) it holds that $\mathcal{S}_X \backslash \mathcal{T}_X$ is a subset of $\mathcal{S}_X^{(\rho)}$. The smoothness condition (A4) is quantified via the Assumption

(A6) The function ϕ is twice continuously differentiable and we have

$$E\Big(\sup_{\xi \in [X_1, X_2] \cup [X_2, X_1]} \Big| \frac{\partial^j \phi}{\partial x^j}(\xi, A) \Big| \,\Big|\, X_1, X_2 \Big) \,\leq\, c_\phi \qquad \text{a.s.},$$

for $j = 1, 2$ and some constant c_ϕ. Moreover ζ_1 satisfies the Lipschitz condition

$$\mathcal{F}\big(\zeta_1(x), \zeta_1(y)\big) \,\leq\, c_\zeta |x - y| \,, \qquad \forall x, y \in \mathcal{S}_X^{(\rho)} \,,$$

for some constant $c_\zeta \in (0, \infty)$.

We also introduce the notation

$$\begin{aligned} \Delta Y \,&:=\, Y_1 - Y_2 \,=\, \Delta\phi + \Delta U \,, \\ \Delta\phi \,&:=\, \phi(X_1, A) - \phi(X_2, A) \,, \\ \Delta U \,&:=\, U_1 - U_2 \,, \\ \Delta X \,&:=\, X_1 - X_2 \,. \end{aligned} \tag{4.1}$$

Step 1: Under Assumption (A3) the corresponding conditional characteristic functions satisfy

$$\psi_{\Delta Y | X} \,=\, \psi_{\Delta\phi | X} \cdot \psi_{\Delta U | X_1} \,, \qquad \text{a.s.} \,. \tag{4.2}$$

For some $h_0 \in (0, \rho)$ let us consider the term

$$\begin{aligned} T_U(h_0, s, X_1) &:= E\big\{ \exp(is\Delta Y) \cdot 1_{\mathcal{S}_X^{(h_0)}}(X) \mid X_1 \big\} / P\big[X \in \mathcal{S}_X^{(h_0)} \mid X_1\big] \\ &= E\big\{ E\big(\exp(is\Delta Y) \cdot 1_{\mathcal{S}_X^{(h_0)}}(X) \mid X \big) \mid X_1 \big\} / P\big[X \in \mathcal{S}_X^{(h_0)} \mid X_1\big] \\ &= E\big\{ \psi_{\Delta Y | X}(s) \cdot 1_{\mathcal{S}_X^{(h_0)}}(X) \mid X_1 \big\} / P\big[X \in \mathcal{S}_X^{(h_0)} \mid X_1\big] \\ &= \psi_{\Delta U | X_1}(s) \cdot E\big\{ \psi_{\Delta\phi | X}(s) \cdot 1_{\mathcal{S}_X^{(h_0)}}(X) \mid X_1 \big\} / P\big[X \in \mathcal{S}_X^{(h_0)} \mid X_1\big] \,, \end{aligned}$$

for any $s \in \mathbb{R}$, which is directly accessible from the distribution of the observation (X, Y). Therein note that $P\big[X \in \mathcal{S}_X^{(h_0)} \mid X_1\big] > 0$ is guaranteed for any $h_0 \in (0, \rho)$ by Assumption (A5); and that we have used (4.2). By Assumption (A6) it holds that

$$\begin{aligned} \big| E\big\{ \psi_{\Delta\phi | X}(s) \cdot 1_{\mathcal{S}_X^{(h_0)}}(X) \mid X_1 \big\} \,-\, P\big[X \in \mathcal{S}_X^{(h_0)} \mid X_1\big] \big| \,&\leq\, c_\phi \, |s| \, E\big\{ |\Delta X| \cdot 1_{\mathcal{S}_X^{(h_0)}}(X) \mid X_1 \big\} \\ &\leq\, c_\phi \, |s| \, h_0 \, P\big[X \in \mathcal{S}_X^{(h_0)} \mid X_1\big] \,, \end{aligned} \tag{4.3}$$

so that

$$\left|T_U(h_0, s, X_1) - \psi_{\Delta U|X_1}(s)\right| \;\leq\; c_\phi\,|s|\,|\psi_{\Delta U|X_1}(s)|\,h_0\,,$$

and, thus, $\lim_{h_0\downarrow 0} T_U(h_0, s, X_1) = \psi_{\Delta U|X_1}(s)$ for any $s \in \mathbb{R}$ almost surely. Therefore $\psi_{\Delta U|X_1}$ and, hence, the conditional distribution of ΔU given X_1 are identified from the distribution of (X, Y).

Step 2: Start out by writing $\mathcal{S}_X^{(h_1,h_2)} := \mathcal{S}_X^{(h_2)} \backslash \mathcal{S}_X^{(h_1)}$ for some $\rho > h_2 > h_1 > 0$. Then we consider the term

$$\begin{aligned}
T_Z &:= T_Z(x, h_1, h_2, h_3, s, X_1) \\
&:= E\big\{\psi_{\Delta U|X_1}(-s/\Delta X)\exp(is\Delta Y/\Delta X)1_{\mathcal{S}_X^{(h_1,h_2)}}(X)1_{[0,h_3]}(|X_1 - x|) \mid X_1\big\} \\
&\quad /E\big\{\left|\psi_{\Delta U|X_1}(s/\Delta X)\right|^2 1_{\mathcal{S}_X^{(h_1,h_2)}}(X)1_{[0,h_3]}(|X_1 - x|) \mid X_1\big\} \\
&= E\big\{\left|\psi_{\Delta U|X_1}(s/\Delta X)\right|^2 \psi_{\Delta\phi|X}(s/\Delta X)1_{\mathcal{S}_X^{(h_1,h_2)}}(X)1_{[0,h_3]}(|X_1 - x|) \mid X_1\big\} \\
&\quad /E\big\{\left|\psi_{\Delta U|X_1}(s/\Delta X)\right|^2 1_{\mathcal{S}_X^{(h_1,h_2)}}(X)1_{[0,h_3]}(|X_1 - x_1|) \mid X_1\big\}\,,
\end{aligned}$$

for some $h_3 > 0$ and any fixed $x = (x_1, x_2)$ with $x_1 = x_2$, which is directly accessible from the distribution of (X, Y) as $\psi_{\Delta U|X_1}$ has already been identified. Again we have used (4.2). Combining Assumption (A5) with the Assumption

(A7) The characteristic function $\psi_{\Delta U|X_1}$ vanishes nowhere almost surely.

we may ensure that, with probability one, the denominator of the term T_Z does not vanish for any s. Assumption (A6) and Taylor approximation yield that

$$\Delta\phi \;=\; Z_1 \cdot \Delta X \;+\; \mathcal{R}\,, \tag{4.4}$$

where the random remainder term $\mathcal{R}$ satisfies

$$\left|\mathcal{R}\right| \;\leq\; \frac{1}{2}c_\phi(\Delta X)^2 \qquad \text{a.s.}\,.$$

It follows from there that, on the event $\big\{X \in \mathcal{S}_X^{(h_1,h_2)}\big\} \cap \{|X_1 - x_1| \leq h_3\}$, we have that

$$\left|\psi_{\Delta\phi|X}(s/\Delta X) - \{\zeta_1(x)\}^{ft}(s)\right| \;\leq\; \big(c_\phi|s|/2\big)\,h_2 \;+\; c_\zeta\,(2h_3 + h_2)\,,$$

using Assumption (A6) so that

$$\lim_{h_2\downarrow 0} T_Z(x, h_1, h_2, h_3, s, X_1) \;=\; \{\zeta_1(x)\}^{ft}(s)\,,$$

for all $s \in \mathbb{R}$ a.s. where we arrange that $h_1 = h_2/2$ and $h_3 = h_2$ to calculate the limit. Note

that $x \in \bigcap_{h_2>0} \mathcal{S}_X^{(h_2/2,h_2)}$. Therefore $\zeta_1(x)$ is identified, and we can summarize the identification result in the following theorem.

Theorem 2. *Under the Assumptions (A1), (A2), (A3) and (A5)–(A7), $\zeta_1(x)$ is identified in the model (2.1) for any $x = (x_1, x_2) \in \mathbb{R}^2$ with $x_1 = x_2$ from the distribution of the observations (X, Y).*

Remark 1. The model (2.1) may be generalized to the setting of multiple regressors, i.e. one observes the i.i.d. data $(X_{k,t}, X'_{k,t}, Y_{k,t})$, $k = 1, \ldots, n$, $t = 1, 2$, where

$$Y_{k,t} \;=\; \phi(X_{k,t}, X'_{k,t}, A_k) + U_{k,t}\,.$$

Then we modify the definition

$$Z_j := \frac{\partial \phi}{\partial x}(x, x', A)\mid_{x=X_j, x'=X'_j}\,,$$

and that of ζ_j accordingly. Let us assume that $(X_{1,1}, X'_{1,1}, X_{1,2}, X'_{1,2})$ has a four dimensional Lebesgue density, which is continuous and strictly positive. Also impose additional Lipschitz conditions on ζ_1 and its partial derivatives with respect to the bivariate component (x, x') in Assumption (A6). Then Theorem 2 can be extended to identify $\zeta_1(x, x')$ at any $(x, x') = (x_1, x_2, x'_1, x'_2)$ with $x_1 = x_2$ and $x'_1 = x'_2$. For any unitary matrix U define

$$\tilde{\phi}(x, y, a) \;:=\; \phi(U^T(x, y)^T, a)\,.$$

Then use the above arguments to identify the conditional distribution of

$$\frac{\partial \tilde{\phi}}{\partial x}(W_1, W'_1, A) \;=\; U_{1,1} \cdot \frac{\partial \phi}{\partial x}(X_1, X'_1, A) + U_{1,2} \cdot \frac{\partial \phi}{\partial x'}(X_1, X'_1, A)\,,$$

given $W_1 = W_2$ and $W'_1 = W'_2$ based on the data Z_j and $(W_j, W'_j)^T = U(X_j, X'_j)^T$ for $j = 1, 2$. That opens the perspective to identify any directional derivative of ϕ at $x_1 = x_2 = x$ and $x'_1 = x'_2 = x'$ and, hence, the gradient of ϕ under appropriate smoothness conditions on ϕ and ζ_1.

Remark 2. If there are more time periods, it is also possible to allow for a time trend. Specifically, we allow for a linear time trend which modifies the structural function ϕ by adding the same the structural function in each time period. More formally, the model takes the form

$$Y_{k,t} = \phi_0\left(X_{k,t}, A_k\right) + \phi_1\left(X_{k,t}, A_k\right) t + U_{kt}, \quad t = 1, \ldots, T, \quad k = 1, \ldots, n, \tag{4.5}$$

where ϕ_0 and ϕ_1 satisfy analogous conditions to before. To identify this model, we require

$T = 4$. Since

$$\begin{aligned} Y_{1,2} - Y_{1,1} &= U_{1,2} - U_{1,1} + \phi_1(X_{1,1}, A_1)\,, \\ Y_{1,4} - Y_{1,3} &= U_{1,4} - U_{1,3} + \phi_1(X_{1,3}, A_1)\,, \end{aligned}$$

holds on the event $\{X_{1,1} = X_{1,2}, X_{1,3} = X_{1,4}\}$ we are able to identify the conditional distribution of $\partial_x \phi_1(x, A) \mid_{x=X_{1,1}}$ given $X_{1,1} = x$ at $x = \lambda \cdot (1,1,1,1)$, $\lambda \in \mathbb{R}$, by the arguments from section 4 under the given assumptions. Moreover

$$\begin{aligned} 2Y_{1,1} - Y_{1,2} &= 2U_{1,1} - U_{1,2} + \phi_0(X_{1,1}, A_1)\,, \\ 4Y_{1,3} - 3Y_{1,4} &= 4U_{1,3} - 3U_{1,4} + \phi_0(X_{1,3}, A_1)\,, \end{aligned}$$

holds on $\{X_{1,1} = X_{1,2}, X_{1,3} = X_{1,4}\}$ again so that the conditional distribution of $\partial_x \phi_0(x, A) \mid_{x=X_{1,1}}$ given $X_{1,1} = x$ at $x = \lambda \cdot (1,1,1,1)$, $\lambda \in \mathbb{R}$, is identified as well. Note that continuity conditions analogous to Assumption (A6) have to be imposed on both ϕ_0 and ϕ_1.

Remark 3. Our framework may be extended to allow for additional covariates, denoted in the following by S_t. The main motivation to do so stems typically from the objective to simply control for these variables; their influence is typically of lesser interest. Due to the curse of dimensionality, it is impractical to let them enter in an unrestricted fashion. Hence we propose a partially linear structure, i.e.,

$$Y_{k,t} = \phi\left(X_{k,t}, A_k\right) + \gamma' S_{k,t} + U_{kt}, \quad t = 1, \ldots, T, \quad k = 1, \ldots, n, \tag{4.6}$$

where $\gamma \in R^{\dim(S_t)}$ is a fixed parameter. Constructive identification of γ is straightforwardly established by noting that, conditional on $X_{k,1} = X_{k,2} = x$, this equation is

$$Y_{k,t} = \tilde{A}_k + \gamma' S_{k,t} + U_{kt}, \quad t = 1, \ldots, T, \quad k = 1, \ldots, n, \tag{4.7}$$

where $\tilde{A}_k = \phi\left(x, A_k\right)$ is a classical, time invariant, additive "fixed effect". This implies that, for every value of x, we obtain a classical linear fixed effect model. Since the coefficient γ is invariant over x, we can then average out over x. A sample counterpart estimator to this identification argument would produce an estimator that converges at the $\dim(X)$ nonparametric regression rate (because we have to impose that $X_{k,1} = X_{k,2}$).
Finally, after forming $Y_{k,t} - \gamma' S_{k,t}$, the further analysis can proceed exactly as outlined above.

Remark 4. (Estimation) Under the stronger Assumption (A3'), the arguments outlined in Step 1 suggest the following estimator for $\psi_{\Delta U}(s)$,

$$\hat{\psi}_{\Delta U}^{(h_0)}(s) := \sum_{k=1}^{n} \exp\left(is\Delta Y_k\right) \cdot 1_{\mathcal{S}_X^{(h_0)}}(X_{k,1}, X_{k,2}) \,/\, \sum_{k=1}^{n} 1_{\mathcal{S}_X^{(h_0)}}(X_{k,1}, X_{k,2})\,, \tag{4.8}$$

based on the moment method, for some $h_0 \in (0, \rho)$ still to be selected. Note that we do not invoke full independence (A3"). By convention, put $\hat{\psi}^{(h_0)}_{\Delta U}(s)$ equal to 0 if the denominator in (4.8) vanishes.

Moreover, the quantity T_Z defined in Step 2 along with its behavior as $h \to 0$ motivates an estimator of $\{\zeta_1(x)\}^{ft}(s)$ under Assumption (A3') , namely

$$\begin{aligned}
&\hat{\psi}^{(h_0,h_1,h_2,h_3)}_{Z_1}(x;s) \\
&:= \sum_{k=1}^{n} \exp\left(is\Delta Y_k/\Delta X_k\right)\hat{\psi}^{(h_0)}_{\Delta U}(-s/\Delta X_k) \cdot 1_{\mathcal{S}_X^{(h_1,h_2)}}(X_{k,1}, X_{k,2}) \cdot K(|X_{k,1} - x_1|/h_3) \\
&\quad / \left\{\rho_n + \sum_{k=1}^{n} \left|\hat{\psi}^{(h_0)}_{\Delta U}(s/\Delta X_k)\right|^2 \cdot 1_{\mathcal{S}_X^{(h_1,h_2)}}(X_{k,1}, X_{k,2}) \cdot K(|X_{k,1} - x_1|/h_3)\right\}, \qquad (4.9)
\end{aligned}$$

for some $0 < h_0 < h_1 < h_2 < \rho$, $h_3 > 0$, some kernel function K and some ridge parameter $\rho_n > 0$ in order to prevent the denominator from getting too close to zero. This approach to heteroskedastic deconvolution is inspired by Delaigle and Meister (2007, 2008).

5 Asymptotic Lower Bound

In this section, we investigate the limits for the asymptotic performance of an arbitrary estimator under the conditions of Theorem 2. For that purpose we consider the polynomial approach (3.1) with $T = 2$ and the random vector A equals

$$A = \begin{pmatrix} X_1X_2 - (X_1 + X_2)B/2 \\ B - X_1 - X_2 \\ 1 \end{pmatrix}, \qquad (5.1)$$

where the random vector B remains to be specified. Under given $X = (X_1, X_2)$, observing

$$\begin{aligned}
Y_1 + Y_2 &= U_1 + U_2\,, \\
\Delta Y/\Delta X &= B + \Delta U/\Delta X\,, \qquad (5.2)
\end{aligned}$$

is equivalent with the observation of the data (Y_1, Y_2), i.e. the random variable (Y_1, Y_2) can be uniquely reconstructed from (5.2) and vice versa. Then $\zeta_1(x)$, at any $x = (x_1, x_2)$ with $x_1 = x_2$, equals the conditional distribution of B given $X = x$. With respect to the random vector U we impose Assumption (A3") and

(A8) $U = (U_1, U_2)$ has the bivariate Lebesgue density

$$(s,t) \mapsto 2f_{\Delta U}(s-t)f_{\Delta U}(s+t)\,,$$

where the Fourier transform of the univariate density $f_{\Delta U}$ satisfies

$$0 < c_{U,1} \leq (1+|t|^{\alpha}) \cdot \left|\psi_{\Delta U}(t)\right| \leq c_{U,2} < \infty\,, \qquad \forall t \in \mathbb{R}\,,$$

for some constants $\alpha > 0$ and $c_{U,1} < c_{U,2}$. Moreover $\psi_{\Delta U}$ is twice continuously differentiable and its derivatives satisfy

$$\sup_t (1+|t|^{\alpha+\ell}) \cdot \left|\psi^{(\ell)}_{\Delta U}(t)\right| \leq c_{U,3}\,,$$

for another constant $c_{U,3} > 0$ and $\ell = 1, 2$.

Under the Assumption (A8), $f_{\Delta U}$ is an ordinary smooth density in the terminology of Fan (1991). Moreover (A8) yields that $U_1 + U_2$ and ΔU are independent and that ΔU has the density $f_{\Delta U}$. Considering (5.2), it follows that

$$(X_{j,t}, \Delta Y_j/\Delta X_j), \qquad j = 1, \ldots, n,\ t = 1, 2\,, \tag{5.3}$$

forms a sufficient statistic for $\zeta_1(x)$ in the model in which the data $(X_{j,t}, Y_{j,t})$, $j = 1, \ldots, n$, $t = 1, 2$, are observed. Therefore we may focus on that experiment in which only the i.i.d. sample (5.3) is available.

Let us now determine the conditional distribution of B given X. Define

$$f_0(x) := c \cdot \{1 - \cos(x)\}^2/x^4\,, \qquad x \in \mathbb{R}\,,$$

with some constant $c > 0$ such that f_0 integrates to one. We introduce

$$f^{[\theta]}_{B|X}(t) := \frac{3}{4} \cdot (1+|t|)^{-4} + \frac{1}{2} f_0(t) \cdot \{1 + \theta \cdot K\big(|X-x|/\theta\big) \cdot \cos(4t)\}\,, \qquad \forall t \in \mathbb{R}\,, \tag{5.4}$$

for any $\theta \in [0, 1]$, as the competing conditional densities of B given X. Therein K denotes some continuously differentiable kernel function which is supported on $[-1, 1]$, bounded by 1 and satisfies $K(0) = 1$. As f_0^{ft} is supported on $[-2, 2]$ the function $f^{[\theta]}_{B|X}$ is a probability density indeed. Moreover we put $f^{[0]}_{B|X}(t) := 3(1+|t|)^{-4}/4 + f_0(t)/2$.

With respect to the design distribution we modify Assumption (A5) via

(A5') There exists some $\rho > 0$ such that the density f_X of $X = (X_1, X_2)$ is continuous and strictly positive on the ball around $x = (x_1, x_1)$ with the radius ρ. Moreover f_X is compactly supported.

We provide the following lower bound on the convergence rates for the estimation of the parameter θ in the model (5.4).

Theorem 3. *We impose that ϕ has the polynomial shape (3.1) with $T = 2$; that A and B obey (5.1) and (5.4), respectively; and that the Assumptions (A1), (A3"), (A5') and (A8) hold true. Then Assumption (A6) is satisfied for appropriate finite constants c_ϕ and c_ζ. For an arbitrary sequence of estimators $(\hat{\theta}_n)_n$, where θ_n is based on the i.i.d. data $(X_{j,t}, Y_{j,t})$, $j = 1, \ldots, n$, $t = 1, 2$, there exists a constant $d > 0$ such that*

$$\liminf_{n\to\infty} \sup_{\theta\in[0,1]} P_\theta^{(n)}\big[|\hat{\theta}_n - \theta|^2 \,>\, d^2 \cdot n^{-1/(2+\alpha)}\big] \,>\, 0\,.$$

6 A Conditional Parametric Estimator

In this section, our goal is to construct a parametric estimator of $\zeta_1(x)$ which attains the convergence rates outlined in Theorem 3. The parametric nature of the estimation problem is represented by the following assumption

(A9) For some fixed $x = (x_1, x_2) \in \mathbb{R}^2$ with $x_1 = x_2$, there exists a parametrization

$$\theta \in \Theta \subseteq \mathbb{R}^d,\, \theta \mapsto \zeta_1(\theta; x)\,,$$

of the admitted conditional measures $\zeta_1(x)$ for $d \geq 1$ such that

$$\inf_{\theta'\neq\theta\in\Theta} \mathcal{F}_R\big(\zeta_1(\theta'; x), \zeta_1(\theta; x)\big)/|\theta' - \theta| \,\geq\, c_p \,>\, 0\,,$$

holds true for some fixed $R \in (0, \infty)$.

Therein $\mathcal{F}_R$ denotes following distance between two probability measures P and Q,

$$\mathcal{F}_R^2(P, Q) \,:=\, \int_{-R}^{R} \big|P^{ft}(t) - Q^{ft}(t)\big|^2 dt\,.$$

The specific parametrization in (5.4), which has been used to prove the lower bound in Theorem 3, satisfies Assumption (A9) when putting

$$c_p^2 \,=\, \frac{\pi}{8} \int f_0^2(t)dt\,.$$

As the estimator $\hat{\theta}$ of θ we define that $\tilde{\theta}$ which minimizes the contrast functional

$$\gamma(x; \tilde{\theta}) \,:=\, \int_{-R}^{R} \big|\hat{\psi}_{Z_1}^{(h_0,h_1,h_2,h_3)}(x; s) - \{\zeta_1(\tilde{\theta}; x)\}^{ft}(s)\big|^2 ds\,,$$

among all $\tilde{\theta} \in \Theta$ where $\hat{\psi}_{Z_1}^{(h_0,h_1,h_2,h_3)}$ is as in (4.9) and h_0, h_1, h_2 and h_3 remain to be selected.

The following theorem provides an upper bound on the estimation error of our estimator $\hat{\theta}$ under appropriate selection of the smoothing parameters. For simplicity we restrict to the uniform kernel K.

Theorem 4. *We consider the model (2.1) for $T = 2$ under the Assumptions (A1), (A3'), (A5'), (A6), (A8) and (A9). The distribution of (X_1, X_2) and the constants in the assumptions are imposed to be fixed while ϕ, θ and the distributions of A and (U_1, U_2) may move in n and d. Then, our estimator $\hat{\theta}$ of θ satisfies*

$$\left|\hat{\theta} - \theta\right|^2 \,=\, \mathcal{O}_P\big(n^{-1/(2+\alpha)}\big)\,,$$

under the selection $K = 1_{[0,1]}$, $\rho_n \asymp 1$, $h_2 = 2h_1$, $h_3 \asymp h_1$, $h_0 \asymp h_1^2$, $h_1 \asymp n^{-1/(4+2\alpha)}$.

Combining Theorem 3 and 4, it follows that our estimator $\hat{\theta}$ achieves the optimal minimax convergence rate. It is remarkable that, in spite of the parametric nature of the estimation problem, the usual square-root-asymptotics are not attainable by any estimator. In the error-free case (i.e. $\alpha = 0$), the convergence rate is $\mathcal{O}_P(n^{-1/4})$ with respect to the non-squared estimation error.

Critically we mention that the asymptotic order of h_1 in Theorem 4 depends on the parameter α from Assumption (A8), which is usually unknown. Therefore we propose a data-driven choice of h_1 (and h_0, h_2, h_3 according to Theorem 4) by splitting the sample. Precisely the estimator $\hat{\theta}$ is only based on $\lfloor qn \rfloor$ of the complete sample for some constant $q \in (0,1)$. All other observations are used to construct an empirical selector $\hat{h}_1$ of h_1 as follows: Define

$$\hat{\alpha} \,:=\, -\big(\log\big|\hat{\psi}_{\Delta U}^{(h_4)}(s_n)\big|\big) \,/\, \log s_n\,,$$

with some deterministic positive parameters h_4 and $s_n > 1$ and the estimator of $\psi_{\Delta U}$ from (4.8); and, finally,

$$\hat{h}_1 \,:=\, n^{-1/(4+2\hat{\alpha})}\,. \tag{6.1}$$

The following result suffices to show that the asymptotic upper bound from Theorem 4 is maintained when using the split-of-the-sample estimator with the plug-in selector $\hat{h}_1$ for h_1. Nevertheless a rough upper on α is required to be known in order to select the parameter γ in Theorem 5.

Theorem 5. *We impose the conditions of Theorem 4; and we choose $K = 1_{[0,1]}$, $s_n = n^\gamma$ for some $\gamma \in (0, 1/(1+2\alpha))$; and $h_4 = 1/s_n$. Then there exist some positive constants b_0 and b_1 such that the estimator $\hat{h}_1$ in (6.1) satisfies*

$$\lim_{n\to\infty} P\Big(n^{1/(4+2\alpha)} \cdot \hat{h}_1 \in [b_0, b_1]\Big) \,=\, 1\,.$$

Remark 5. Note that we estimate the parameter α under general nonparametric constraints (see Assumption (A8)), leading to the empirical bandwidth $\hat{h}_1$ in (6.1). If more restrictive parametric assumptions are imposed on the distribution of ΔU then the parameter α could also be estimated e.g. by maximum likelihood methods.

7 Simulation

For an illustration of the estimator in the univariate case, remember the panel data model in (2.1). Within this class of models, we constructed two leading specifications: a second and a third-order polynomial in the sole regressor $X_{k,t}$.

$$\begin{aligned} Y_{k,t} &= \phi\left(X_{k,t}, A_k\right) + U_{k,t} \quad \text{where:} \\ \phi\left(X_{k,t}, A_k\right) &= A_{0,k} + A_{1,k}X_{k,t} + A_{2,k}X_{k,t}^2 \qquad \text{Quadratic 1D Model} \\ \phi\left(X_{k,t}, A_k\right) &= A_{0,k} + A_{1,k}X_{k,t} + A_{2,k}X_{k,t}^2 + A_{3,k}X_{k,t}^3 \qquad \text{Cubic 1D Model} \end{aligned}$$

where, for all $k = 1, \ldots, n$ and $t = 1, \ldots, T$:

- $A_{j,k} \sim \mathcal{N}(0, .5) \qquad \forall \quad j \in \{0, 1, 2, 3\}$
- $X_{k,t} \sim .5 + e_x \qquad e_x \sim \mathcal{N}(0, .5)$
- $U_{k,t} \sim e_v \qquad e_v \sim \text{Laplace}(0, .1)$

Since this is a univariate case, we can simply nonparametrically estimate the distribution of the conditional characteristic functions by using our estimator from Equation (4.9).

We select a proper α to optimize our results, and determine the bandwidths in the following way: $h_1 = n^{-1/(4+\alpha)}, h_2 = 2h_1, h_3 = h_1, h_0 = h_1^2$, as suggested by Theorem 4. While these are the asymptotically most efficient bandwidths, there may be better bandwidths in practical application. The restrictions that the bandwidths must obey imply that $0 < h_1 < h_2 < \rho$ and $h_3 > 0$.

We will compute the values of μ and σ to minimize the Euclidean distance between $\widehat{\phi}_Z(s, x)$ and the characteristic normal distribution.

$$\phi_{\Delta Z}(s, x) = \exp(i\mu s - \sigma^2 s^2/2)$$

7.1 Results in the Baseline Specification

The specifications outlined above have easily represented true values. These are given by:

$$Z_{k,t} := \frac{\partial \phi}{\partial x}(x, A)|_{x=X_{k,t}} = A_{1,k} + 2A_{2,k}X_{k,t} \qquad \text{Quadratic 1D Model}$$
$$Z_{k,t} := \frac{\partial \phi}{\partial x}(x, A)|_{x=X_{k,t}} = A_{1,k} + 2A_{2,k}X_{k,t} + 3A_{3,k}X_{k,t}^2 \qquad \text{Cubic 1D Model}$$

To display the true model, we use an oracle kernel density estimator that uses the (in the real world unobserved) values of $Z_{k.t}$. Figures 2 and 3 show the results comparing our estimator to the true distribution estimated by such an oracle kernel density estimator.

Start out by considering Figure 2: The blue line in the left two graphs corresponds to the true mean, resp., standard deviation, of the conditional marginal effects. The left two graphs display moreover the estimated conditional means, resp. standard deviations, for each value of x, and the corresponding estimation uncertainty as given by bootstrap 95% confidence bands. As is evident, the estimated means track the true values very closely, while the standard deviations perform (as expected) worse, yet still deliver a quite satisfactory fit.

On the right are two contour graphs showing first a contour plot of the true conditional density of the marginal treatment effects along with the conditional means, as estimated using again an oracle kernel density estimator, and secondly an estimate of the conditional densities estimated using our method. As before, our estimator for the density of marginal effects matches the true distribution of the marginal effect very closely.

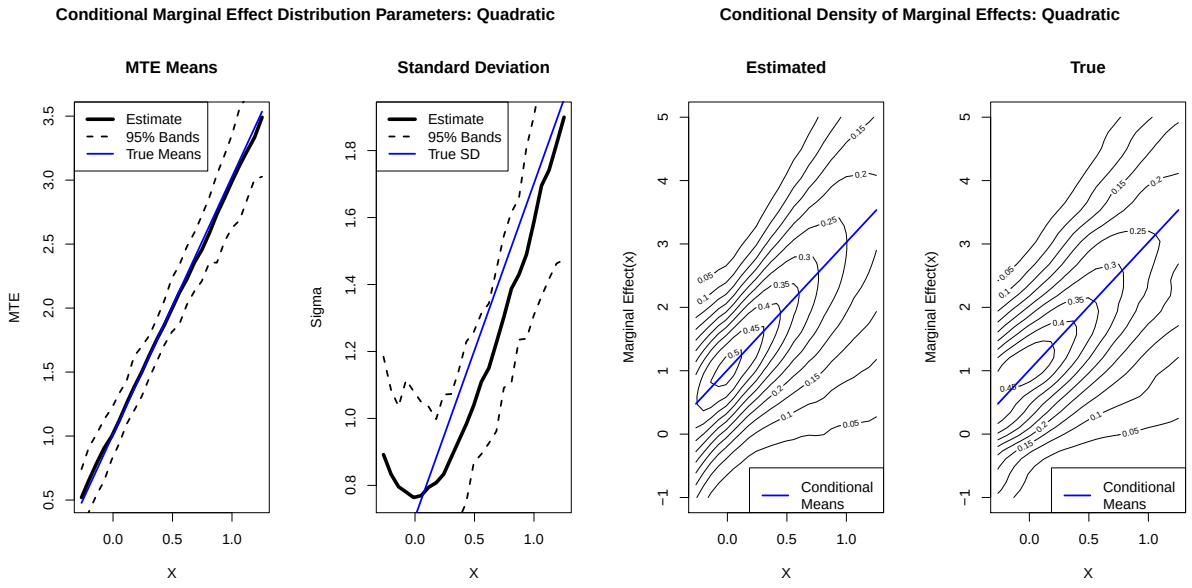

Figure 2: Estimates of quadratic 1D model using: $\alpha = 2$ and N = 10,000.The black line is our estimate. The dotted lines are our 95% confidence bands estimated with 100 bootstraps. The blue line is the true means and standard deviation we are trying to estimate.

Figure 3 then repeats the exercise for the cubic model and obtains similar, if slightly worse, performance, which is to expected given the slightly more complex model.

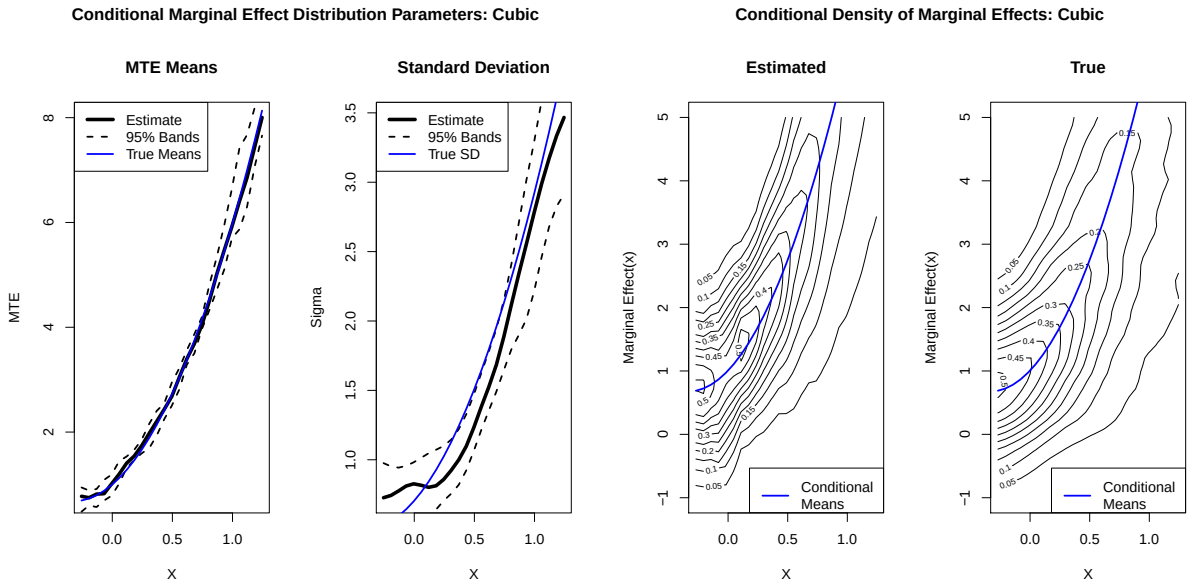

Figure 3: Estimates of cubic 1D model using: $\alpha = 2$ and N = 10,000. The dotted lines are our 95% confidence bands estimated with 100 bootstraps. The blue line is the true means and standard deviation we are trying to estimate.

We also include an estimate of the quantiles of marginal effects in Figure 4, using our approach. This is done by inferring the quantiles from the conditional normal density, for which we have estimates of μ and σ for each value of X.

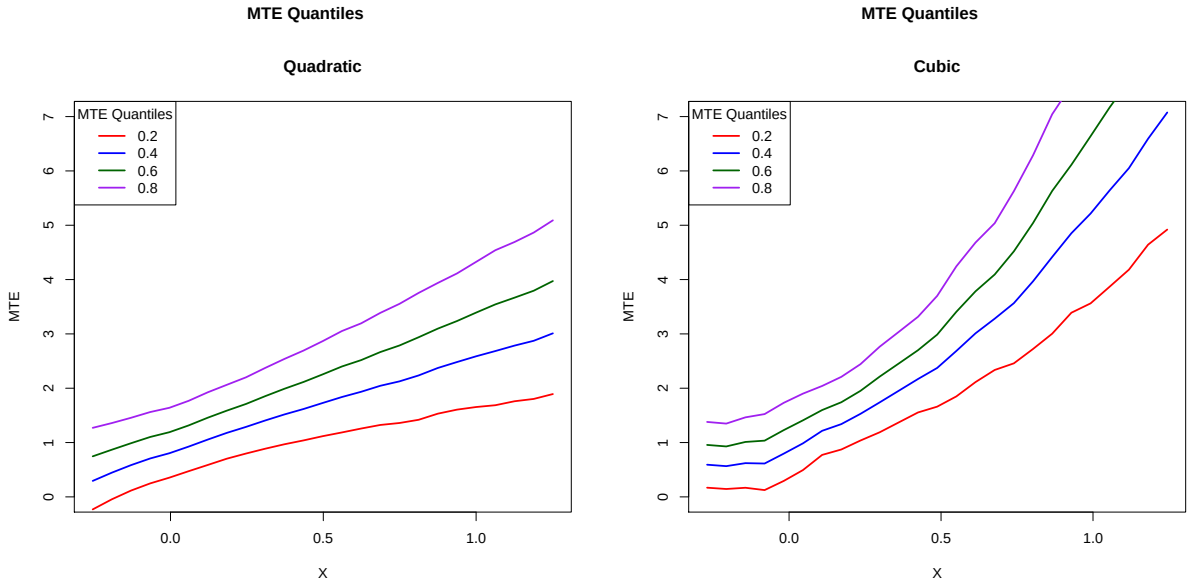

Figure 4: Estimates of Quantile Effects.

Note that these are conditional densities of marginal effects, so the most dense regions are on the boundaries where the standard deviation is the lowest, even though most of the data are near the mean of X. We can also estimate the joint densities of $Z := \frac{\partial \phi}{\partial x}(x, A)$ and X, by multiplying our estimate of the conditional density with the density of X, $f(x)$. We estimate the density of X using a kernel density estimation function. The resulting joint densities are displayed in Figure 5 below:

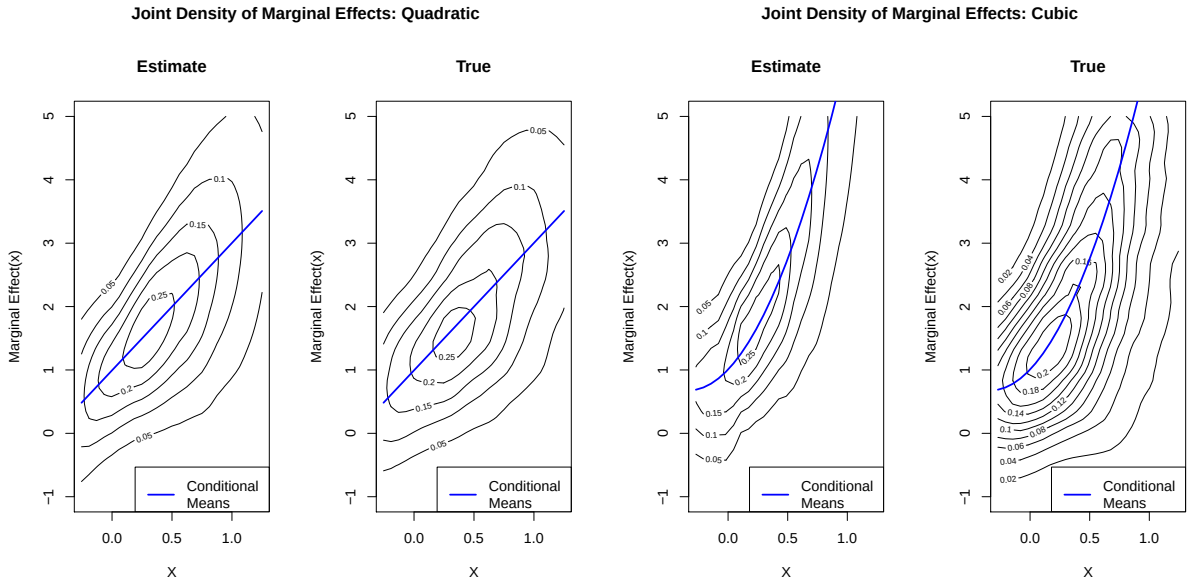

Figure 5: Estimates of joint distribution of the quadratic 1D model on the left and of the cubic 1D model on the right using the same parameters as above.

7.2 A Violation of Conditional Normality: Skewed Distribution of Effects

Next, in order to evaluate the robustness of our estimation procedure, we study the performance of our estimator in a simulation scenario which violates the conditional parametric assumption imposed for semiparametric estimation. We will assume that A comes from a mixed normal distribution.

- $A_{j,k} \sim 0.5 \cdot \mathcal{N}(0.7, 0.2) + 0.5 \cdot \mathcal{N}(-0.25, 0.1) \qquad \forall \quad j \in \{0, 1, 2, 3\}$

This function is skewed to the right, i.e., it will not exhibit symmetrical marginal effects. The results for both the cubic case and quadratic case are included below. In Figures 6 and 7, we see that our estimates of the means are still quite accurate. However, our estimates for the

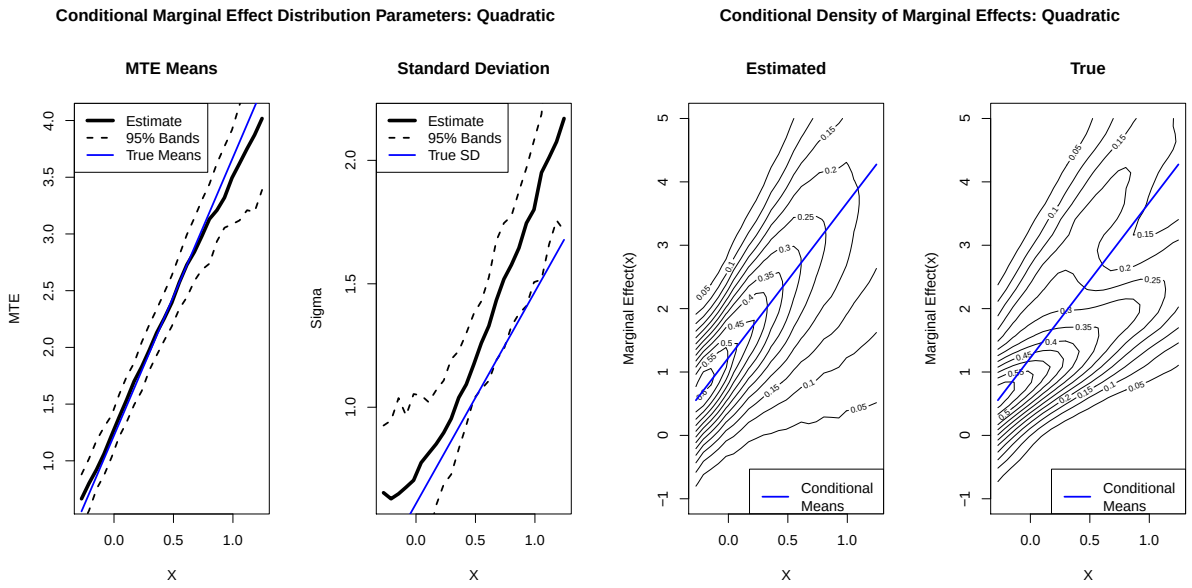

Figure 6: Estimates of quadratic 1D model using: $\alpha = 2$ and N = 10,000.The black line is our estimate. The dotted lines are our 95% confidence bands estimated with 100 bootstraps. The blue line is the true means and standard deviation we are trying to estimate.

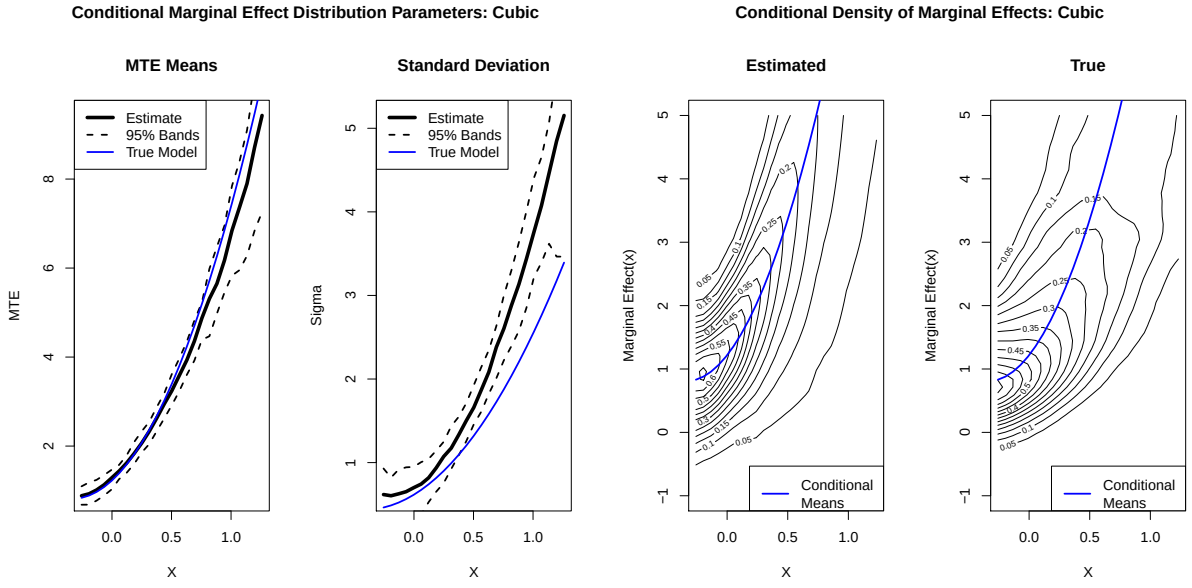

Figure 7: Estimates of cubic 1D model using: $\alpha = 2$ and N = 10,000. The dotted lines are our 95% confidence bands estimated with 100 bootstraps. The blue line is the true means and standard deviation we are trying to estimate.

standard deviation are slightly too high, since the estimated density exhibits a wider spread because of the skewed density of marginal effects.

Moreover, the joint and conditional estimated densities (see Figures 6, 7, and 8) do a reasonable job in capturing the general orientation of effects, but are unsurprisingly not fully able to capture the true model perfectly, as we (wrongly) impose normality of the conditional

distribution. Note, however, that estimated conditional means are quite close to the true results, and the overall performance appears to be reasonably robust against violations of the parametric specification.

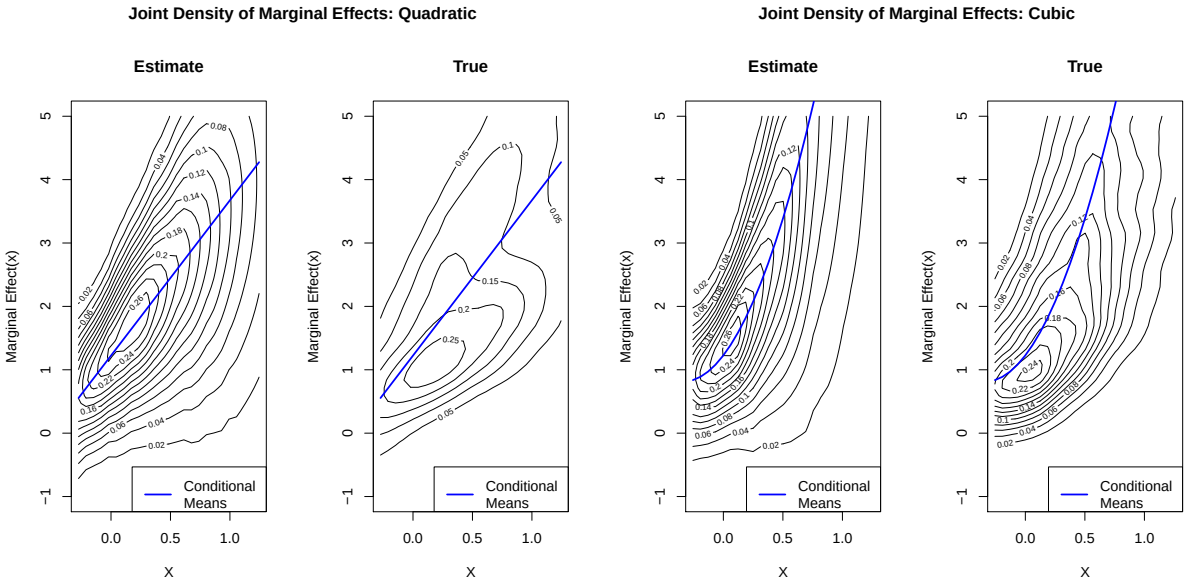

Figure 8: Estimates of joint distribution of the quadratic 1D model on the left and of the cubic 1D model on the right using the same parameters as above.

8 Empirical Application

In this section, we study the performance of our estimation procedure using real world data. We consider the estimation of the distribution of marginal effects of every additional dollar on the consumption of junk food. Because of the implied health consequences, as outlined below this question is highly policy relevant. In addition, our model is very well suited to capture differences in these marginal effects between wealthy and poor households, which are not captured at all by linear random coefficients models. This ability to exhibit differences for different wealth and income levels is crucial for the policy debate, as it is widely believed that excessive consumption of junk food is particularly prevalent at the lower end of the income distribution. As such, we hope that our estimator is able to inform this policy debate by providing a more nuanced picture of the distribution of marginal effects.

We start out with an overview of the data we use in our estimation exercise. After that, we provide a brief review of the policy debate surrounding junk food demand, especially with respect to differences in income. We then display our empirical findings which corroborate many of the suggestions put forward in the literature.

8.1 Data

8.1.1 An Overview

For our application, we use the Nielsen Scanner Dataset which is available through the Kilts Center at the University of Chicago Booth School of Business[4]. We will focus our study on the year 2014 where there are about 55,000 individuals. This is a helpful dataset for estimating demand behavior since it contains detailed information based on price and quantity of all retail purchases as well as detailed household characteristics for all consumers. The data contain a representative sample of households in the United States who use in-home scanners to record all of their purchases intended for personal, in-home use. Nielsen matches the product scanned by the household to the actual price of the store where the product was bought. Nielsen estimates that about 30% of household consumption is accounted for by these purchases.

We will call this sum over all Nielsen expenditure categories total expenditure; under additive separability of the utility function this is the relevant total outlay variable. The same variable also takes the place in derivations involving economic rationality - under additive separability, this is the relevant "income" variable, e.g., to analyze Slutsky negative semidefiniteness. For this model, we estimate the total outlay ("income") and own price elasticities and the marginal effects of an additional unit of total outlay ("income") on the demand for junk food. Nielsen aggregates millions of universal product codes (UPC) into different groups of food.

We define junk food as any food classified as potato chips, candy or carbonated beverages by Nielsen. Junk food is a good example in our situation because these items lie on one extreme of the nutrition-taste trade-off (Blaylock et al. (1999)). Junk food sacrifices almost all of its nutrition for taste. We aggregate the data to a monthly level such that period 1 is January 2014 and period 2 is February 2014. Of course, we could use different months as the time periods in our dataset as long as these periods exclude the irregular Christmas shopping period.

Prices are more precisely an aggregate price index called Stone-Lewbel (SL) cross section prices (see Lewbel (1989) and Hoderlein and Mihaleva (2008)). Generally speaking, SL prices use the fact that within a category of goods (junk food in our case), people have different tastes for the individual goods. Using standard aggregate price indices for junk food implicitly assumes that all individuals have identical Cobb Douglas preferences for all goods within this category, but SL prices allow all individuals to have heterogeneous Cobb Douglas preferences for the various commodities in this bundle. This implies that the typical approach of using aggregate price indices is a restrictive case of using SL prices. For this reason, SL prices should

[4]Researcher(s) own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researcher(s) and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

always be used when possible.

Total expenditure for all Nielsen goods and all junk food is aggregated each month as well. In order to get the proper expenditure, we only use households with two individuals and no children and divide expenditure by two, in order to estimate average expenditure per consumer. This is justified, as junk food is arguable a private good, and household composition effects can be expected to be negligible.

8.1.2 Limitations

There are a few concerns with the data. The data rely on participants successfully recording their purchases in their home, so they may suffer from recording error. The specific issue that we might be concerned with is that consumers may consume a good when it is purchased and will not record the purchase when they return home. Einav et al. (2010) finds that consumable goods like soft drinks, chips, or candy are likely to be consumed before getting home so are more likely to not be scanned. There are also recording errors such as when a six-pack of goods are purchased and recorded as quantity six. However, these errors only seem to have minor effects. When compared to data from grocery store recorded sales, the data in Nielsen Homescan data matched 94% of the time (Einav et al. (2010)).

Another potential source of measurement error is related to the price rather than the quantity. Individuals record their purchases by scanning the items they buy when they get home. The individuals input the quantity they purchase, and Nielsen matches it with the average price of the good at the store where they purchased it that week. This can lead to two types of errors. The first comes from the price changing in the middle of the week, though frequent changes during several weeks are less likely. The second type of error comes from not including discounts from loyalty cards. Einav et al. (2010) examines a retailer used in the Homescan data which has loyalty cards and finds that loyalty cards are used in about 75-80% of the transactions. Further, this would bias our prices and expenditure upwards. When comparing Homescan data with data from the retailer, Einav et al. (2010) finds that the prices used in the Homescan data is about 7% higher and the overall expenditure is 10% higher. On the other hand, these price measurement errors may be overestimated since some retailers do not have loyalty cards at all.

Finally, Homescan data errors are comparable to errors found in other commonly used data sets. Aguiar and Hurst (2007) finds that life-cycle pattern of household expenditures recorded in Homescan Data is consistent with those reported for food expenditures at home in Panel Study of Income Dynamics (PSID). Einav et al. (2010) finds that these issues are not more serious than those in any other consumption surveys like the Current Population Survey (CPS). Lin (2018) compares the fraction of expenditures on different categories of products in the Nielsen Homescan Data and finds the results consistent to results from the Consumer Expenditure

Survey (CES). In sum, we feel that these potential sources of measurement error may bias our results somewhat, but are unlikely to invalidate them.

8.2 Literature Review

There is a large literature on the determinants, extent and consequences of the consumption of junk food. As regards determinants, sometimes low-income propensity to consume unhealthy is attributed to the cost of healthy food (see, e.g., Drewnowski and Darmon (2005), Golan et al. (2008), and Drewnowski and Eichelsdoerfer (2010)). However, Carlson and Frazo (2012) found that junk food is cheaper on a per-calorie basis than healthier foods like fruits, vegetables, whole grains and proteins, but that the healthier foods are actually cheaper on a per-serving basis. Rider et al. (2012) found that health attributes have been found to not be associated with higher average transaction prices.

When it comes to extent and possible consequences, obesity is one of the most important health problems in the United States, as well as many other countries. Many of the junk foods we consider are high in sugar, and excess sugar consumption is strongly linked with many diet-related diseases such as diabetes, cancers and heart disease (see WHO (2015)). Obesity leads to several hundred billion dollars spent on medical costs in the US annually, about 10-27 percent of all medical costs (See Finkelstein et al. (2009) and Cawley et al. (2015)). Thus, consumption of unhealthy food, such as junk food, can have a major impact on individual well being as well as the economy at large.

Our estimator allows for a more nuanced picture of the demand patterns for junk food, and hence enables policy makers to better target policy measures on subgroups of the population. Obesity and diabetes rates are higher for low income individuals (see Drewnowski and Specter (2004) and Robbins et al. (2001)). Binkley and Golub (2011) and Chen et al. (2012) all found that low-income households consume less nutritious foods. Allcott et al. (2017) showed that even when controlling for supply side factors, high-income households have a greater demand for healthy foods. We add to this literature a more differentiated description of the distribution of marginal effects for individuals with different incomes, which crucially relies on the added flexibility that our approach warrants relative to linear random coefficients models, e.g., Graham and Powell (2012).

8.3 Income Elasticities and Marginal Effects of Income

To begin, as a building block for our model, but also to obtain naive "income" elasticities, we display the mean budget share of junk food (i.e., the proportion of Nielsen recorded junk food over all Nielsen recorded items) for each household, $\omega_{k,t}$, as the dependent variable and total log expenditure, $E_{k,t}$, as the right hand variable in the first period (denoted t). Throughout

this subsection, we control for prices by using households whose prices are in a neighborhood of the median price in period t, denoted p. Thus, the model we estimate is as follows:

$$\omega_{k,t} = \phi\left(E_{k,t}, A_{k,t}, p_t\right) + U_{k,t} \tag{8.1}$$

The associated graph is included in Figure 9. Note that budget share is decreasing with total expenditure which strengthens the idea that low-income households eat more unhealthy food than high-income households. The convex curve implies that both the marginal effect of income on consumption of junk food and the income-elasticity of demand of junk food varies across expenditure. We will use our method to estimate $Z_j(e,p) = \frac{\partial \phi}{\partial e} e$. We then follow standard

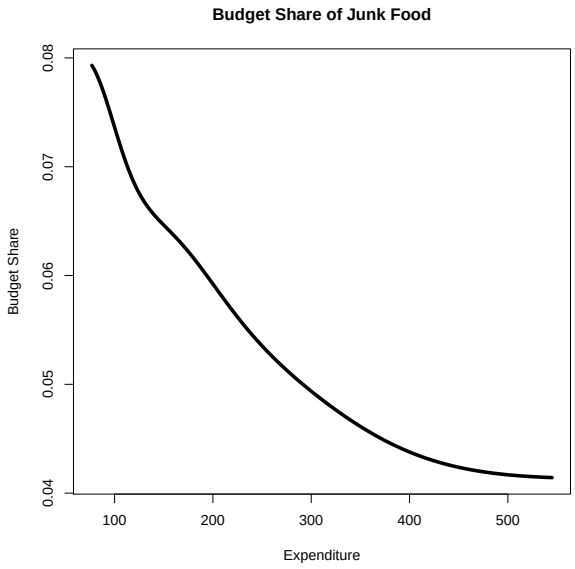

Figure 9: Nardaya-Watson kernel regression estimator of Budget Share of Junk Food based on total expenditures

arguments from Almost Ideal Demand System (AIDS) (Deaton and Muellbauer, 1980), and use equation (8.2) estimate to identify and estimate the elasticity of income, ε^d using our estimate of $Z_j(e,p)$ from equation (8.1).

$$\varepsilon_j^d(e,p) = \frac{Z_j(e,p)}{\omega_j(e,p)} + 1 \tag{8.2}$$

To utilize this for the estimation of the elasticities, we use $\omega_j(e,p)$ which, as mentioned, is estimated using Nadaraya-Watson kernel regression estimator. This allows us then to estimate the conditional density of income elasticities of demand for junk food. The means and standard deviations of the coefficients, as well as the conditional density of marginal effects, are displayed in Figure 10. The pointwise standard errors have been constructed using the naive bootstrap. Note that the income elasticities of demand decrease with expenditure, and are clearly signifi-

cantly non-linear. Thus, given an one percent increase in income, low-income individuals will increase their junk food consumption by a higher percentage than high-income individuals.

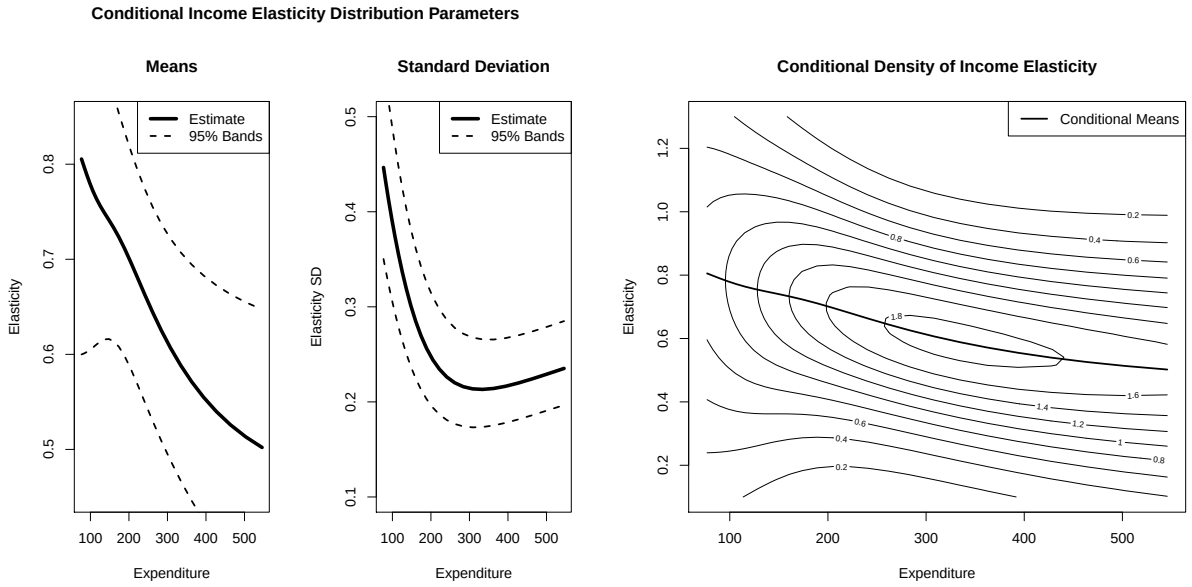

Figure 10: Estimates of Elasticity of demand using: $\alpha = 6$. For this sample, $N = 6,870$

Note that these are estimates of the conditional density of income elasticities of the demand for junk food conditioned on "income" (as discussed, actually total Nielsen goods expenditure). We can estimate the joint density by multiplying this conditional density by the distribution of total expenditure, measured using a kernel density estimation. The result of this procedure can be seen in Figure 11 where we also include estimates of the conditional quantiles of the distribution of income elasticities and income.

Furthermore, we can then use the elasticity estimates to estimate the density of marginal effects of an unit of additional income on the demand for junk food, using the following identity: Let q be the quantity of junk food consumed. Consider

$$\varepsilon_d(e,p) = \frac{\partial \log(q)}{\partial \log(e)} = \frac{\partial q}{\partial e}\frac{e}{q} = \frac{\partial q}{\partial e}\frac{p}{\omega(e,p)} \tag{8.3}$$

Since we control for own price and keep it constant, we can normalize price to be equal to one for computational ease. Thus, we can estimate the marginal effect of an additional dollar on consumption of junk food, $\frac{\partial q}{\partial e}$. The result of this analysis is displayed in Figure 12, along with the quantile of these marginal effects. The effects follow the same trend as the income elasticities of demand, but the difference between low-income individuals and high-income individuals is more pronounced.

To understand this graph better we show, in Figure 13, the estimated density of marginal effects of income on consumption of junk food for different groups based on their income

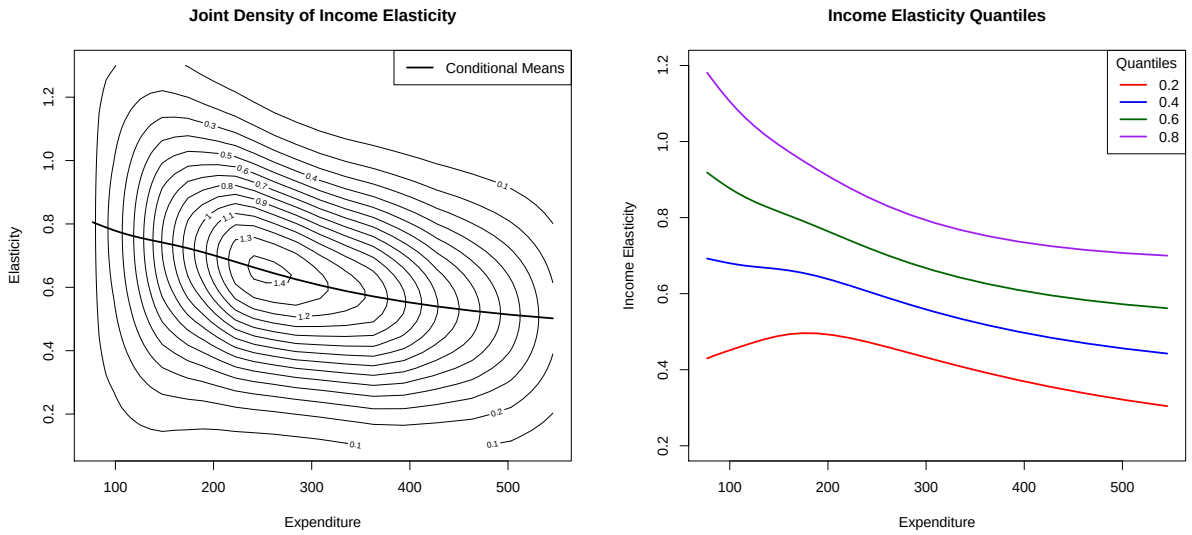

Figure 11: Joint distributions are calculated by multiplying the conditional distribution by the distribution of expenditure.

quantile. Specifically, we graph the distribution of marginal effects for those at the .2, .4, .6 and .8 quantiles of the income distribution. To illustrate this point, consider the following example. In our example, low income individuals have income elasticities of about 0.8 and high income individuals have income elasticities of about 0.5. Consider that low income budget share of junk food is 0.08 while high income budget share of junk food is about 0.04. If we plug these values into equation (12), for low income individuals we obtain $0.8 = \frac{\partial q}{\partial e}\frac{1}{0.08}$ so that the marginal effect is $\frac{\partial q}{\partial e} \cong 0.064$. For high income individuals, $\frac{\partial q}{\partial e} \cong 0.02$. Thus, while the income elasticity of low income individuals is on average only 50% higher than the elasticity of high income individuals, the marginal effect of income on quantity of junk food consumed of poor individuals is more than twice as high compared to their high income counterparts. In other words, for every dollar they spend on Nielsen goods, they consume more than twice the quantity of junk food.

Remember that these densities of marginal effects are conditional on total expenditure ("income"). To estimate the joint density, as before we multiply the estimate of the conditional density by a kernel density estimate of total expenditure ("income"). The results for the joint density of marginal effects are found in Figure 14, along with the density of marginal effects for those in the .2, .4, .6 and .8 quantiles of the "income" distribution. As is to be expected, this re-weighting results in the 0.6 quantile of the income distribution to deliver the density with largest values, rather than the edge case of the 0.8 quantile as is the case with the conditional densities.

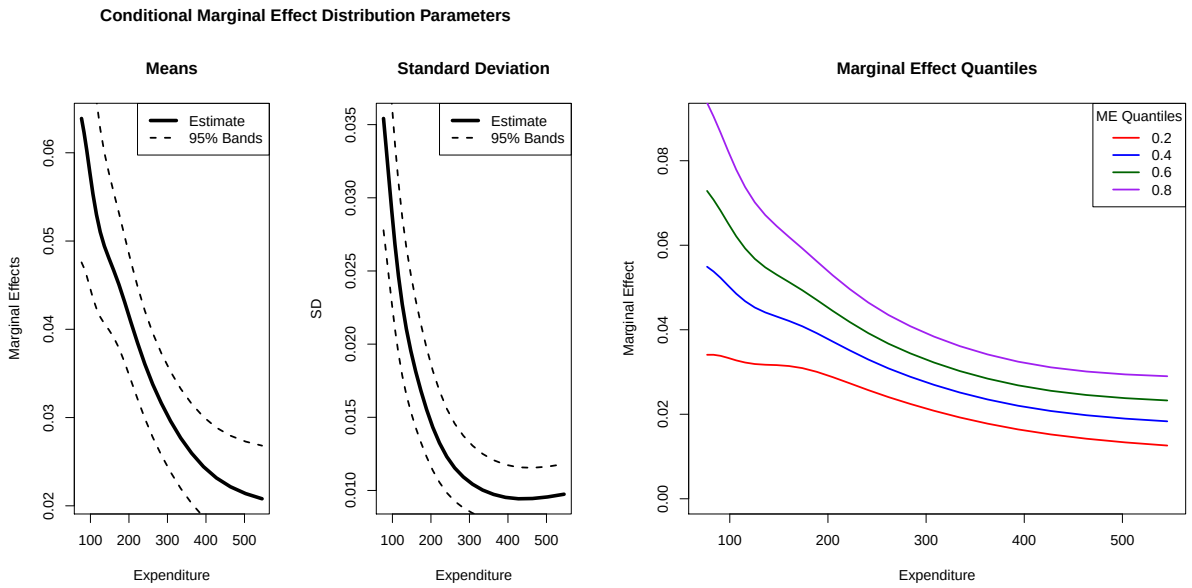

Figure 12: Estimates of the marginal effect of an additional dollar of expenditure on junk food using: $\alpha = 6$. For this sample, $N = 6,870$

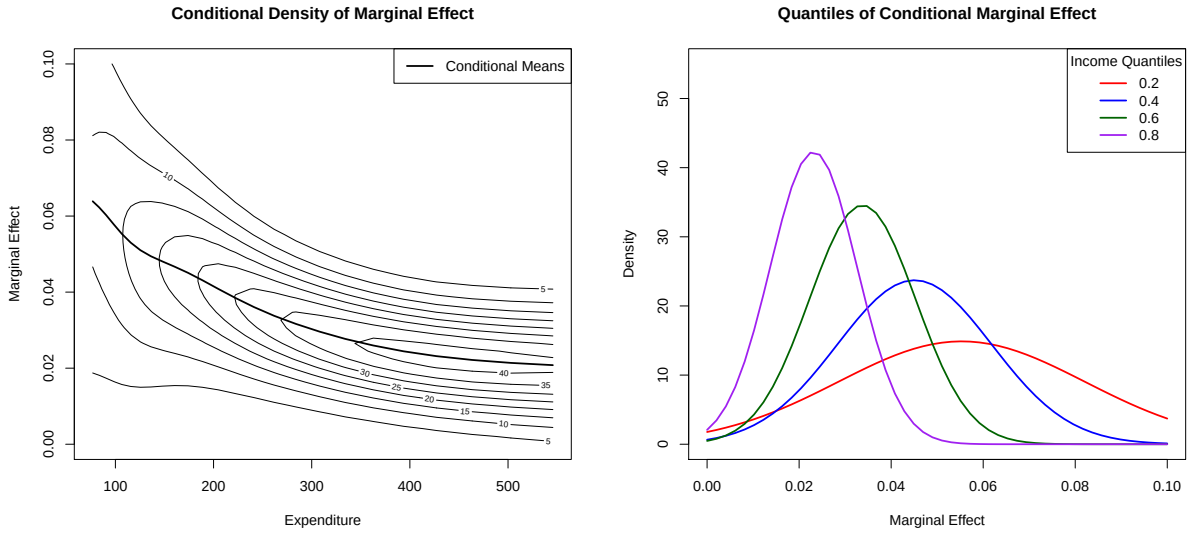

Figure 13: Conditional density and different expenditure quantiles of the estimates of marginal effect.

Finally, note that a naive estimator could be based on an estimated derivative of the budget share graph in Figure 9. However, we expect these estimates to be biased because they do not account for the endogeneity stemming from the correlation between the high dimensional unobservables and income. The results are included below in Figure 15, which exhibit significant differences from our previous estimates.

Additional results with a different method to control for prices can be found in the appendix.

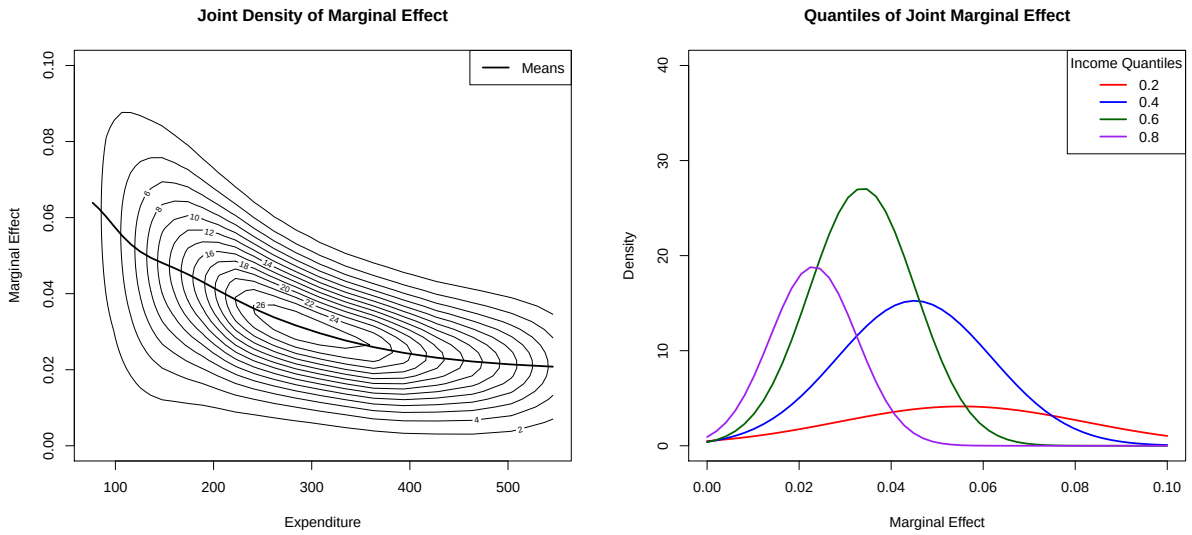

Figure 14: Joint distributions are calculated by multiplying the conditional distribution by the distribution of expenditure

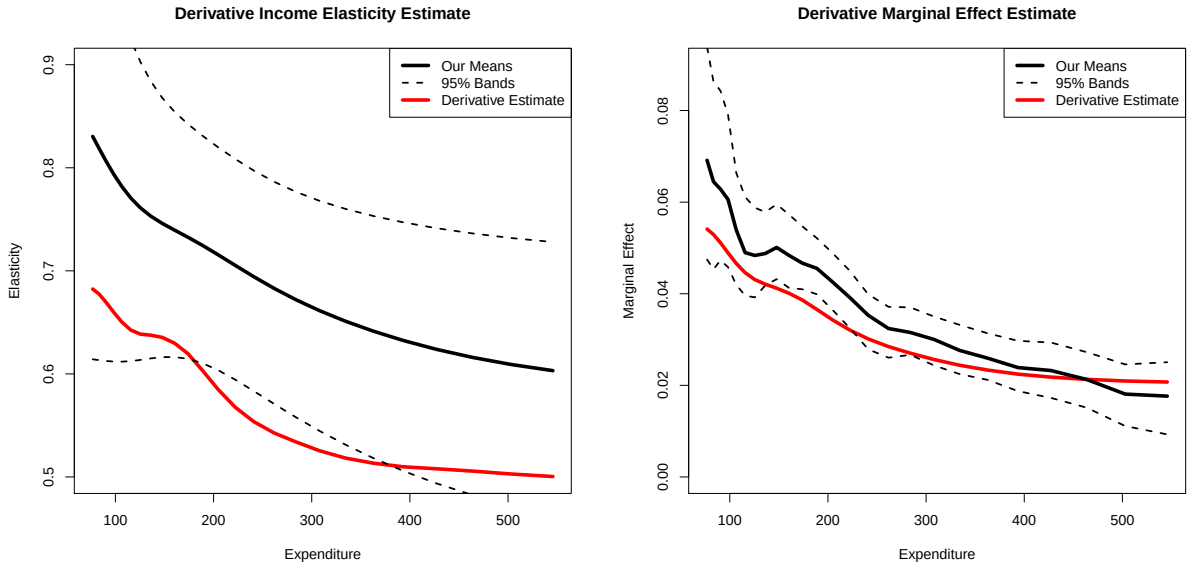

Figure 15: Mean and 95% bands of the mean of our estimates of income elasticity and marginal effect estimates compared to a nonparametric estimate of the derivative of the budget share graph.

8.4 Own Price Elasticities

Following similar steps as above, we estimate own-price elasticities by using the budget share of junk food for each household, $\omega_{k,t}$, as the dependent variable and log of our SL price indices, $P_{k,t}$, as the right hand variable, but control again for income by selecting households with total expenditure close to the median, denoted e. Thus,

$$\omega_{k,t} = \phi\left(e, P_{k,t}, A_{k,t}\right) + U_{k,t} \tag{8.4}$$

We will use our method to estimate $\tilde{Z}_j(p,e) = \frac{\partial \phi}{\partial p}\left(p, e, A\right)|_{p=P_j}$, and use equation (8.5) to identify the elasticity of income, ε^p using our estimate of $\tilde{Z}_j(e,p)$ from equation (8.4).

$$\varepsilon_j^p(e,p) = \frac{\tilde{Z}_j(e,p)}{\omega_j(e,p)} \tag{8.5}$$

We use again the Nadaraya-Watson estimator of $\omega_j(e,p)$, now as a function of price, see Figure 16.

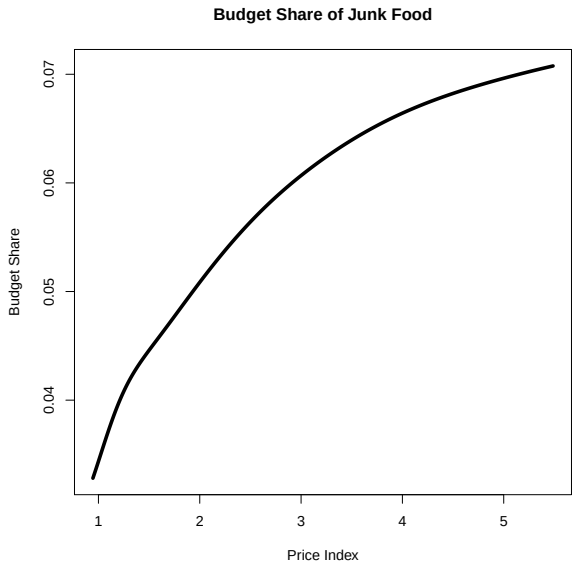

Figure 16: Nadaraya-Watson kernel regression estimator of Budget Share of Junk Food based on prices

With the estimate of budget share conditional on price, we can use our estimate of the density of $\tilde{Z}_j(e,p)$ and equation (8.5) to estimate the conditional distribution of own-price elasticities of for junk food. Below are the means and standard deviations of the coefficients as well as a contour map of the density in Figure 17, along with bootstrap standard errors. Note that own-price elasticities generally are negative and decrease with prices, i.e., increase in absolute value. Thus, given an increase of one percent in price, the reduction in demand for high-priced junk food is larger than for low-priced junk food.

Note again that these estimates are for the own-price elasticity for junk food conditional on price (and income). We can estimate the joint distribution by multiplying this conditional distribution by the density of expenditure, estimated using a kernel density estimation, see Figure 18 for the result. We also include the quantile estimates of own-price elasticities which

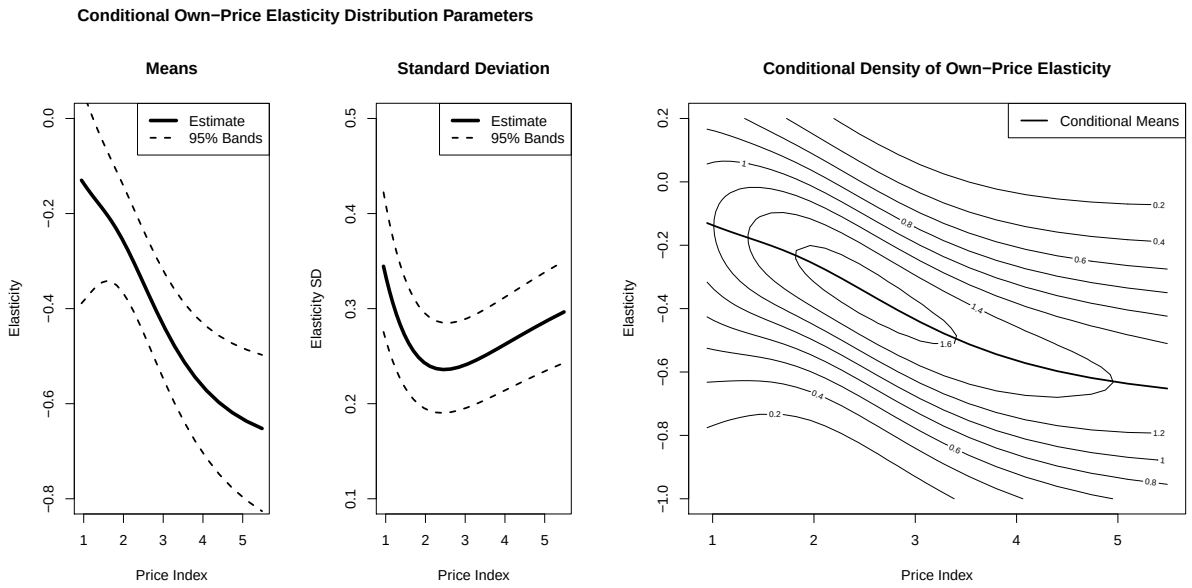

Figure 17: Estimates of Elasticity of demand using: $\alpha = 6$. For this sample, $N = 8,086$

allows to assess the difference in quantiles of consumers' own-price elasticities at different prices.

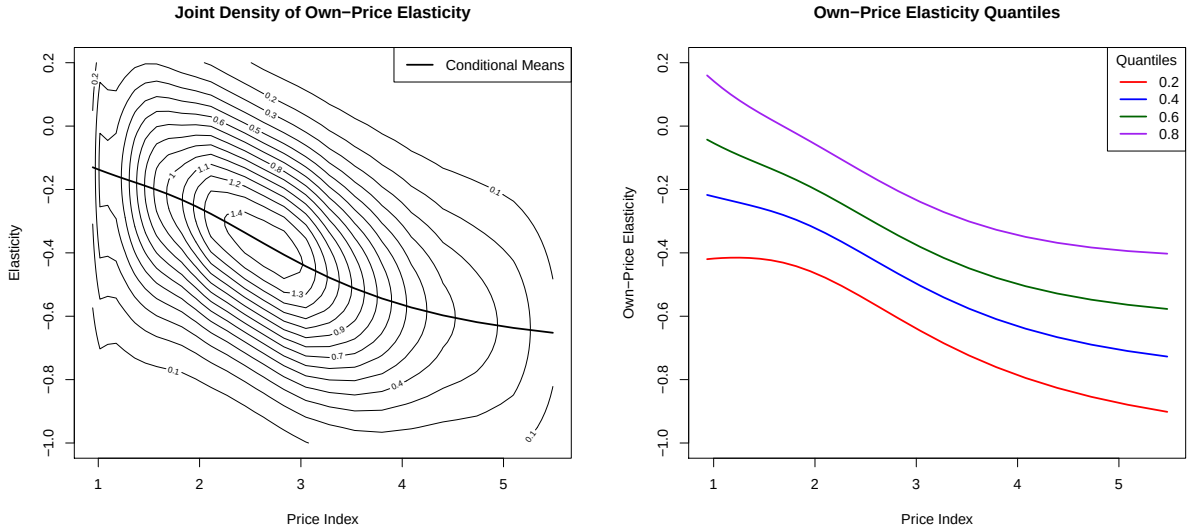

Figure 18: Joint distributions are calculated by multiplying the conditional distribution by the distribution of prices.

Finally, we compare our results again with the naive procedure that takes the derivative of the budget share regression, which differ because they do not properly account for the correlation stemming from the high dimensional correlated unobservables, see Fig 19.

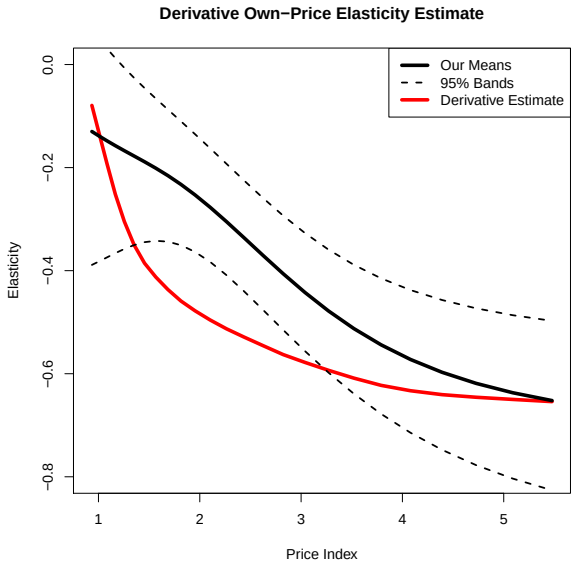

Figure 19: Mean and 95% bands of the mean of our estimates of own-price elasticity compared to a nonparametric estimate of the derivative of the budget share graph.

A Heteroskedastic Variances

In this subsection of the appendix we provide a sketch to demonstrate that the identification principles outlined in this paper can be applied to the model

$$Y_t = \phi(X_t, A) + \sigma(X_t)V_t, \qquad t = 1, 2$$

where $(V_1, V_2) \perp (X_1, X_2, A)$ and $\sigma(x) = \sqrt{\sigma^2(x)}$ is the skedastic function, and we assume for simplicity that the V_t are *iid* over time. This skedastic function specification is very common in nonparametric regression analysis. The main difference is that this model also allows for correlated unobservables A to enter the main function of interest. For simplicity, in what follows we will focus on the mean and variance of the distribution of marginal effects. As in Step 1, we identify $\sigma^2(x)$ for every x by

$$\begin{aligned} E\left[(\Delta Y)^2 | X_1 = x, \Delta X = 0\right] &= \sigma^2(x) E[(\Delta V)^2 | X_1 = x, \Delta X = 0] \\ &= \sigma^2(x) \end{aligned}$$

using that $V \perp X|A$ and the normalization $E[(\Delta V)^2] = 1$. Assuming differentiability of σ, we can thus also identify $\partial_x \sigma$. Moreover, as V_1 and V_2 are *iid*, it holds that $Var(\Delta V) = 2Var(V_1)$. Adding the standard normalization $E[V_t] = 0$, we obtain that $Var(V_1) = 0.5$.

Next, consider ΔY conditional on $X_1 = x, \Delta X = h$,

$$\begin{aligned} \Delta Y &= \phi(x + h, A) - \phi(x, A) + \sigma(x + h)V_2 - \sigma(x)V_1 \\ &\cong \partial_x \phi(x, A)h + \sigma(x + h)V_2 - \sigma(x)V_1 \end{aligned} \tag{A.1}$$

This allows to get the mean and the variance in a straigtforward fashion. Using again $V \perp (X, A)$, as well as the above normalizations, we obtain:

$$\begin{aligned} &\lim_{h\to 0} h^{-1} E\left[\Delta Y | X_1 = x, \Delta X = 0\right] = E\left[\partial_x \phi(x, A) | X_1 = x, \Delta X = 0\right], \\ &\lim_{h\to 0} h^{-2} \big\{ Var\left[\Delta Y | X_1 = x, \Delta X = 0\right] - \sigma^2(x) - h\sigma(x)\partial_x \sigma(x) \big\} \\ &\quad = Var\left[\partial_x \phi(x, A) | X_1 = x, \Delta X = 0\right] + 0.5 \big\{ \sigma(x)\partial_x^2 \sigma(x) + \big(\partial_x \sigma(x)\big)^2 \big\} \end{aligned}$$

Appropriately rearranging terms identifies $Var\left[\partial_x \phi(x, A) | X_1 = x, \Delta X = 0\right]$. Identification of the ChF follows under the same assumptions and from the realization that $\tilde{Y}_1 = Y_1/\sigma(X_1)$, $\tilde{Y}_2 = Y_2/\sigma(X_2)$ conditional on $X_1 = X_2 = x$ has the structure of Kotlarski's lemma, and hence opens a perspective to identify the ChF of $V_1 | X_1 = x, \Delta X = 0$ for any x. Using the decomposition in equation (A.1), this means that the ChF of ΔY, conditional on $X_1 = x, \Delta X = h$, and independence of A, V_1 and V_2 can be used to obtain the ChF of $\partial_x \phi(x, A)$, conditional on $X_1 = x, \Delta X = h$, using limiting arguments as in the paper.

B Proofs

Proof of Lemma 2.1: As $\mathfrak{B}(\mathbb{R})$

is generated by a countable system of sets (e.g. consider the intervals $(-\infty, q]$, $q \in \mathbb{Q}$) the uniqueness theorem for probability measures guarantees that the measures ζ_j and $\tilde{\zeta}_j$ coincide almost surely by the assumptions of the lemma. Thus the set

$$\mathcal{Z}_j \,:=\, \left\{x \in \mathbb{R}^T \,:\, \zeta_j(x) \neq \tilde{\zeta}_j(x)\right\},$$

is a $\mathcal{L}(X)$-null set; and $\mathcal{Z}_j$ is open in $\mathbb{R}^T$

thanks to the continuity of ζ_j and $\tilde{\zeta}_j$. Hence, the random vector X lies in the closed set $\mathcal{S}_X \backslash \mathcal{Z}_j$ almost surely. As $\mathcal{S}_X$ is defined as the intersection of all those closed sets in which X is located almost surely, it follows that

$$\mathcal{S}_X \,=\, \mathcal{S}_X \backslash \mathcal{Z}_j\,,$$

so that $\zeta_j(x) = \tilde{\zeta}_j(x)$ for all $x \in \mathcal{S}_X$. □

Proof of Lemma 3.1: For any $x \in \mathbb{R}$, we consider the $(T+1) \times (T+1)$-Vandermonde matrix $M(x)$ which contains $p(x_1)^\dagger, \ldots, p(x_T)^\dagger, p(x)^\dagger$ as its rows; and the matrix $N(x)$ which is obtained from $M(x)$ by replacing its last row by $q(x)$. Note that $\det N(x_j) = 0$ is equivalent to linear independence of the vectors $p(x_1), \ldots, p(x_T), q(x_j)$. Thanks to the multilinearity of the determinant and the well-known representation of determinants of Vandermonde matrices we deduce that

$$\det N(x) \,=\, \frac{d}{dx}\{\det M(x)\} \,=\, \Big(\prod_{1 \leq k < l \leq T} (x_l - x_k)\Big) \cdot \frac{d}{dx} \prod_{t=1}^{T} (x - x_t)\,.$$

Thus, $\det N(x_j)$ vanishes if and only if at least two of the $x_1, \ldots, x_T$ coincide or the polynomial $x \mapsto \prod_{t=1}^{T}(x - x_t)$ has a multiple zero at x_j. The latter claim requires at least one of the x_t for $t \neq j$ to coincide with x_j, which implies the first claim. □

Proof of Lemma 3.2: We easily recognize by definition that the vectors $p(x_1), \ldots, p(x_T), q(x_j)$

are all continuous functions in $x \in \mathbb{R}^T$. Applying a Gram-Schmidt process we obtain that

$$\begin{aligned} p_k^*(x) &= p(x_k) - \sum_{l=1}^{k-1} \big(p(x_k)^\dagger p_l^*(x)\big)\, p_l^*(x) / \big|p_l^*(x)\big|^2, \quad k = 1, \ldots, T, \\ q_j^*(x) &= q(x_j) - \sum_{l=1}^{T} \big(q(x_j)^\dagger p_l^*(x)\big)\, p_l^*(x) / \big|p_l^*(x)\big|^2, \\ \tau_j(x) &= \big|q_j^*(x)\big|^2, \end{aligned}$$

for $x \in \mathcal{X}$ so that τ_j is continuous on $\mathcal{T}_X$ as well. The positivity of τ_j is an immediate consequence of Lemma 3.1 as $\tau_j(x) = 0$ implies linear dependence between $p(x_1), \ldots, p(x_T), q(x_j)$. □

Proof of Lemma 3.3: For any $x, y \in \mathbb{R}^T$, $b \geq 0$, we deduce by the triangle inequality that

$$\begin{aligned} \mathcal{F}\big(\zeta_j^{[b]}(x), \zeta_j^{[b]}(y)\big) \leq\ & \mathcal{F}\big(\mathcal{L}(A^{[0]} \mid X = x), \mathcal{L}(A^{[0]} \mid X = y)\big) \\ & + \sup_{s \in \mathbb{R}} \big|\psi_{A^{[0]}|X=x}\big(sq(x_j)\big) - \psi_{A^{[0]}|X=x}\big(sq(y_j)\big)\big| \\ & + \sup_{s \in \mathbb{R}} \big|\psi_{A^{[0]}|X=x}\big(sq(x_j)\big)\big| \cdot \Big|\exp\Big(-\frac{1}{2}bs^2\tau_j^2(x)\Big) - \exp\Big(-\frac{1}{2}bs^2\tau_j^2(y)\Big)\Big|. \end{aligned} \tag{B.1}$$

The first term in (B.1) converges to 0 as $y \to x$ by Assumption (A4). As $A^{[0]}$ has a conditional Lebesgue density given $X = x$ it follows from the Riemann-Lebesgue lemma (see e.g. Bochner & Chandrasekharan, 1949) that $\lim_{|u| \to \infty} \psi_{A^{[0]}|X=x}(u) = 0$. Thus, for any $\varepsilon > 0$, there exists some $R > 0$ such that $\big|\psi_{A^{[0]}|X=x}(u)\big| < \varepsilon/4$ for all u with $|u| > R$. Since $|q(x)| \geq 1$ for all $x \in \mathbb{R}$ the second term in (B.1) obeys the upper bound

$$\varepsilon/2 + \sup_{|s| \leq R} \big|\psi_{A^{[0]}|X=x}\big(sq(x_j)\big) - \psi_{A^{[0]}|X=x}\big(sq(y_j)\big)\big|. \tag{B.2}$$

As the function $x \mapsto q(x)$ is continuous and any characteristic function is uniformly continuous, (B.2) is bounded from above by ε whenever $|y - x|$ is sufficiently small with respect to only ε and R. Therefore the second term tends to 0 as $y \to x$.

It remains to consider the third term in (B.1). Let ε and R be as in the previous paragraph. Then the third term is smaller or equal to

$$\varepsilon/2 + \sup_{|s| \leq R} \Big|\exp\Big(-\frac{1}{2}bs^2\tau_j^2(x)\Big) - \exp\Big(-\frac{1}{2}bs^2\tau_j^2(y)\Big)\Big|. \tag{B.3}$$

As $x \mapsto \tau_j(x)$ is continuous (see Lemma 3.2) and the exponential mapping is uniformly continuous on any bounded domain, the term (B.3) is bounded from above by ε whenever $|y - x|$ is sufficiently small with respect to ε and R. Finally we have shown that all three terms in (B.1)

converge to 0 as y tends to x. $\square$

Proof of Lemma 3.4: Applying Fourier transformation to both sides of the given equality we obtain that

$$Q^{ft}(x) \cdot \exp\Big(-\frac{1}{2}\alpha x^2\Big) \;=\; Q^{ft}(x) \cdot \exp\Big(-\frac{1}{2}\alpha'^2\Big), \quad \forall x \in \mathbb{R}\,.$$

As Q^{ft} is continuous and satisfies $Q^{ft}(0) = 1$ there exists a non-void open neighborhood of 0 in which Q^{ft} does not vanish. Therefore the functions $x \mapsto \exp\big(-\alpha x^2/2\big)$ and $x \mapsto \exp\big(-\alpha'^2/2\big)$ coincide on this neighborhood so that $\alpha = \alpha'$. $\square$

Proof of Theorem 3: Thanks to (5.4) and the compact support of f_X, which is guaranteed by Assumption (A5'), we may easily verify the first part of Assumption (A6) for some c_ϕ sufficiently large. With respect to the second part we deduce that

$$\mathcal{F}\big(\zeta_1(y), \zeta_1(z)\big) \;\le\; c_\zeta \cdot |y - z|\,,$$

for all $y, z \in \mathbb{R}$ where $c_\zeta := \|K'\|_\infty/2$. Thus Assumption (A6) holds true.

As the statistic ΔY_j, $j = 1, \ldots, n$, has been shown to be sufficient for $\zeta_1(x)$ and, hence, for the parameter θ, we may consider $P_\theta^{(n)}$ as the image measure of this statistic. Now we put $\theta_n := 3d \cdot n^{-1/(4+2\alpha)}$ so that at least one of the events $\{|\hat\theta_n - \theta_n| > d \cdot n^{-1/(4+2\alpha)}\}$ and $\{|\hat\theta_n| > d \cdot n^{-1/(4+2\alpha)}\}$ occurs. For sufficiently large n it holds that

$$\sup_{\theta\in[0,1]} P_\theta^{(n)}\big[|\hat\theta_n - \theta| \;>\; d \cdot n^{-1/(4+2\alpha)}\big] \;\ge\; \frac{1}{2} - \frac{1}{2}\,\mathrm{TV}\big(P_{\theta_n}^{(n)}, P_0^{(n)}\big)\,.$$

By standard information-theoretic arguments, we deduce that

$$\mathrm{TV}\big(P_{\theta_n}^{(n)}, P_0^{(n)}\big) \;\le\; 2\Big\{\Big(1 + E\chi^2\big(f_{B|X}^{(\theta_n)} * f_{\Delta U}(\cdot/(X_1 - X_2)), f_{B|X}^{(0)} * f_{\Delta U}(\cdot/(X_1 - X_2))\big)\Big)^n - 1\Big\}^{1/2},$$

where χ^2 stands for the χ^2-distance between two measures. By Parseval's identity, it holds that

$$\begin{aligned}
&E\chi^2\big(f_{B|X}^{(\theta_n)} * f_{\Delta U}(\cdot/(X_1 - X_2)), f_{B|X}^{(0)} * f_{\Delta U}(\cdot/(X_1 - X_2))\big) \\
&\quad\le\; \text{const.} \cdot \theta_n^2 \cdot E\,K^2\big(|X - x|/\theta_n\big) \cdot \int \big|\{f_0 \cos(4\cdot)\} * f_{\Delta U}(\cdot/(X_1 - X_2))\big|^2(t)(1 + t^4)dt \\
&\quad=\; \text{const.} \cdot \theta_n^2 \cdot \max\Big\{E\,K^2\big(|X - x|/\theta_n\big)\,|X_1 - X_2|^{-2\ell_2} \\
&\qquad\qquad \cdot \int \big|\big\{f_0^{ft}\big\}^{(\ell_1)}(t \pm 4)\big|^2\,\big|\psi_{\Delta U}^{(\ell_2)}(t/(X_1 - X_2))\big|^2 \;:\; \ell_1, \ell_2 \in \mathbb{N}_0, \ell_1 + \ell_2 \le 2\Big\} \\
&\quad=\; \mathcal{O}\big(\theta_n^{4+2\alpha}\big)\,.
\end{aligned}$$

Therefore, choosing $d > 0$ sufficiently small, we may ensure that

$$\limsup_{n\to\infty} \mathrm{TV}\big(P_{\theta_n}^{(n)}, P_0^{(n)}\big) \,<\, 1\,,$$

which completes the proof of the theorem. □

Proof of Theorem 4: Writing

$$\begin{aligned} N_0 \,&:=\, \sum_{k=1}^{n} 1_{\mathcal{S}_X^{(h_0)}}(X_{k,1}, X_{k,2})\,, \\ N_1 \,&:=\, \sum_{k=1}^{n} 1_{\mathcal{S}_X^{(h_1,h_2)}}(X_{k,1}, X_{k,2}) \cdot 1_{[0,h_3]}(|X_{k,1} - x_1|)\,, \end{aligned}$$

we introduce the events

$$\begin{aligned} \mathcal{E}_0 \,&:=\, \{N_0 \geq c \cdot nh_0\}\,, \\ \mathcal{E}_1 \,&:=\, \{N_1 \geq c \cdot nh_3(h_2 - h_1)\}\,, \end{aligned}$$

for some constant $c > 0$. By Chebyshev's inequality and Assumption (A5') we deduce that the probabilities for the complements of $\mathcal{E}_0$ and $\mathcal{E}_1$ converge to zero as n tends to infinity for $c > 0$ sufficiently small. The events $\mathcal{E}_0$ and $\mathcal{E}_1$ are contained in the σ-field σ_X which is generated by the random variables $X_{k,t}$, $k = 1, \ldots, n$, $t = 1, 2$.

Now put $\varepsilon_n := dn^{-1/(4+2\alpha)}$ for some constant $d > 0$. By Assumption (A9) the inequality

$$\int_{-R}^{R} \big|\{\zeta_1(\hat{\theta}; x)\}^{ft}(s) - \{\zeta_1(\theta; x)\}^{ft}(s)\big|^2 ds \,\geq\, c_p^2\, \varepsilon_n^2\,,$$

holds true on the event $\{|\hat{\theta} - \theta| > \varepsilon_n\}$. Then it follows from the definition of $\hat{\theta}$ that

$$\int_{-R}^{R} \big|\hat{\psi}_{Z_1}^{(h_0,h_1,h_2,h_3)}(x; s) - \{\zeta_1(\theta; x)\}^{ft}(s)\big|^2 ds \,\geq\, \frac{1}{4} c_p^2\, \varepsilon_n^2\,,$$

whenever $|\hat{\theta} - \theta| > \varepsilon_n$. Hence, by Markov's inequality, we deduce that

$$P\big[|\hat{\theta} - \theta| > \varepsilon_n\big] \,\leq\, 4c_p^{-2}\, \varepsilon_n^{-2} \cdot \int_{-R}^{R} E\, 1_{\mathcal{E}_0 \cap \mathcal{E}_1} \big|\hat{\psi}_{Z_1}^{(h_0,h_1,h_2,h_3)}(x; s) - \{\zeta_1(\theta; x)\}^{ft}(s)\big|^2 ds + 1 - P(\mathcal{E}_0 \cap \mathcal{E}_1)\,. \tag{B.4}$$

By a standard bias-variance decomposition for the conditional expectation, the Cauchy-Schwarz

inequality and Assumption (A6), we obtain that

$$E\big\{\big|\hat{\psi}_{Z_1}^{(h_0,h_1,h_2,h_3)}(x;s) - \{\zeta_1(\theta;x)\}^{ft}(s)\big|^2 \mid \sigma_X,\, \hat{\psi}_{\Delta U}^{(h_0)}\big\}$$
$$\leq\ (2\rho_n+1)/\big\{\rho_n\,+\,\hat{\Xi}_U\big\}\,+\,4\,\big\{c_\phi R h_2/2 + c_\zeta(2h_3+h_2)\big\}^2\,+\,4\,\hat{\Xi}_\Delta\,/\,\big\{\rho_n\,+\,\hat{\Xi}_U\big\}\,, \qquad \text{(B.5)}$$

for all $s \in [-R,R]$ where σ_X denotes the σ-field generated by $X_1,\ldots,X_n$; and

$$\hat{\Xi}_U\ :=\ \sum_{k=1}^{n}\big|\hat{\psi}_{\Delta U}^{(h_0)}(s/\Delta X_k)\big|^2\cdot 1_{\mathcal{S}_X^{(h_1,h_2)}}(X_{k,1},X_{k,2})\cdot 1_{[0,h_3]}(|X_{k,1}-x_1|)\,,$$
$$\hat{\Xi}_\Delta\ :=\ \sum_{k=1}^{n}\big|\hat{\psi}_{\Delta U}^{(h_0)}(s/\Delta X_k)-\psi_{\Delta U}(s/\Delta X_k)\big|^2\cdot 1_{\mathcal{S}_X^{(h_1,h_2)}}(X_{k,1},X_{k,2})\cdot 1_{[0,h_3]}(|X_{k,1}-x_1|)\,,$$
$$\Xi_U\ :=\ \sum_{k=1}^{n}\big|\psi_{\Delta U}(s/\Delta X_k)\big|^2\cdot 1_{\mathcal{S}_X^{(h_1,h_2)}}(X_{k,1},X_{k,2})\cdot 1_{[0,h_3]}(|X_{k,1}-x_1|)\,.$$

We deduce by Assumption (A6) that

$$E\big(\hat{\Xi}_\Delta \mid \sigma_X\big)\ \leq\ N_1/N_0\ +\ R^2 c_\phi^2\,\Xi_U\,h_0^2/h_1^2\,. \qquad \text{(B.6)}$$

Thus, on the event $\mathfrak{E}_3(s) := \{\hat{\Xi}_U > \Xi_U/2\}$, $|s|\leq R$, the conditional expectation of term (B.5) given σ_X obeys the upper bound

$$\mathcal{O}\big(h_2^2+h_3^2+h_0^2/h_1^2+1/\Xi_U+N_1/(\Xi_U N_0)\big)\,, \qquad \text{(B.7)}$$

where Ξ_u has the asymptotic lower bound $N_1\cdot h_1^{2\alpha}$ with uniform constants by the Assumptions (A5') and (A8). On the complement of $\mathfrak{E}_3(s)$, the conditional expectation of term (B.5) given σ_X is bounded from above by

$$\mathcal{O}(n^2)\cdot\exp\big\{-N_0(1-1/\sqrt{2}-c_\phi h_0/h_1)^2 c_{U,1}^2(1+R/h_1^\alpha)^{-2}/8\big\}\,, \qquad \text{(B.8)}$$

by Assumption (A6) and Hoeffding's inequality. Applying the expectation to the terms (B.7) and (B.8) – multiplied by $1_{\mathcal{E}_0\cap\mathcal{E}_1}$ – we conclude that the right hand side of (B.4) tends to zero if, first, the limit superior is taken with respect to $n\to\infty$ and, then, the limit $d\to\infty$ is applied. □

Proof of Theorem 5: It suffices to show the existence of some $c>0$ such that

$$\limsup_{n\to\infty}\ P\big(\big|\hat{\alpha}-\alpha\big|\ >\ c/\log n\big)\ =\ 0\,. \qquad \text{(B.9)}$$

Using that the probability of $\mathcal{E}_4$ (equivalent to the event $\mathcal{E}_0$ from the proof of Theorem 4 when replacing h_0 by h_4) converges to 1; that (4.2) holds true; and Hoeffding's inequality – condi-

tionally on σ_X – we can verify (B.9) when c is sufficiently large with respect to γ. □

C Summary Statistics

Below is the Summary Statistics for the data we used in our empirical application.

Table B1

	January 2014	February 2014
SL Price Index	0.7751	0.8069
	(0.5850)	(0.5769)
Junk Food Share	0.0567	0.0639
	(0.0596)	(0.0631)
Total Expenditure	477.96	448.33
	(325.43)	(302.09)

This table contains the mean and standard deviation (in parenthesis beneath the means) for the variables that we use in our analysis

D Application with Different Prices

Below are the results when we control for prices a little differently. Here, price is controlled such that price is centered around the .4 quantile. This serves as a robustness check on the results from our empirical application of the paper. The overall trends are consistent in both cases.

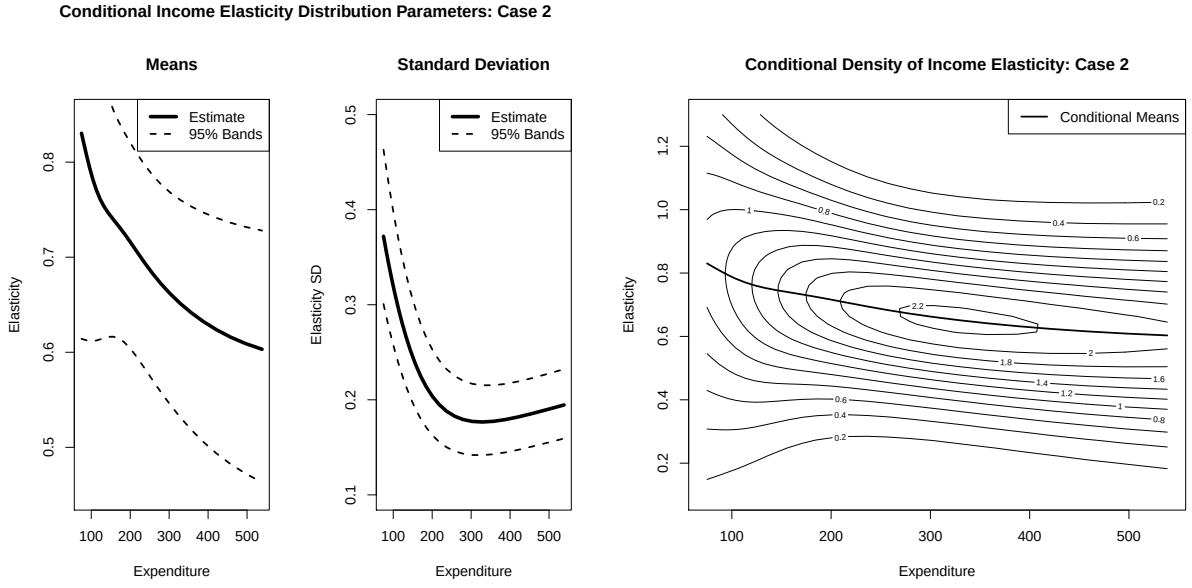

Figure 20: Estimates of Elasticity of demand using: $\alpha = 6$. For this sample, $N = 8,631$

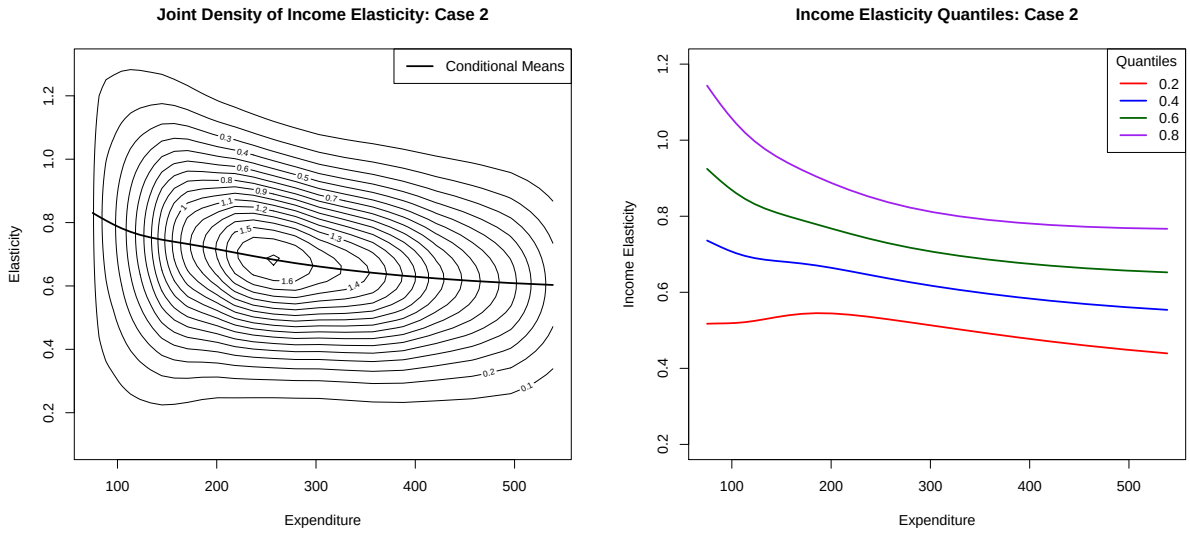

Figure 21: Joint distributions are calculated by multiplying the conditional distribution by the distribution of expenditure.

The only difference of significance is that the decline of mean Elasticity of Demand does not change as much for low-income vs. high-income individuals (see Figure 20 compared to Figure 10). For example, in our base case, mean income elasticity for low-income individuals is about 0.8 and for high income individuals it is about 0.5. In our adjusted case, the income elasticity of low-income individuals is 0.8 while for high income individuals it's about 0.6. This is a minor difference and the results from these estimates easily fit in our confidence bands from our paper.

Our marginal effects estimation in this case is also very similar (see Figures 12 and 22). These results imply that these results are consistent across different prices, as long as prices are properly controlled for.

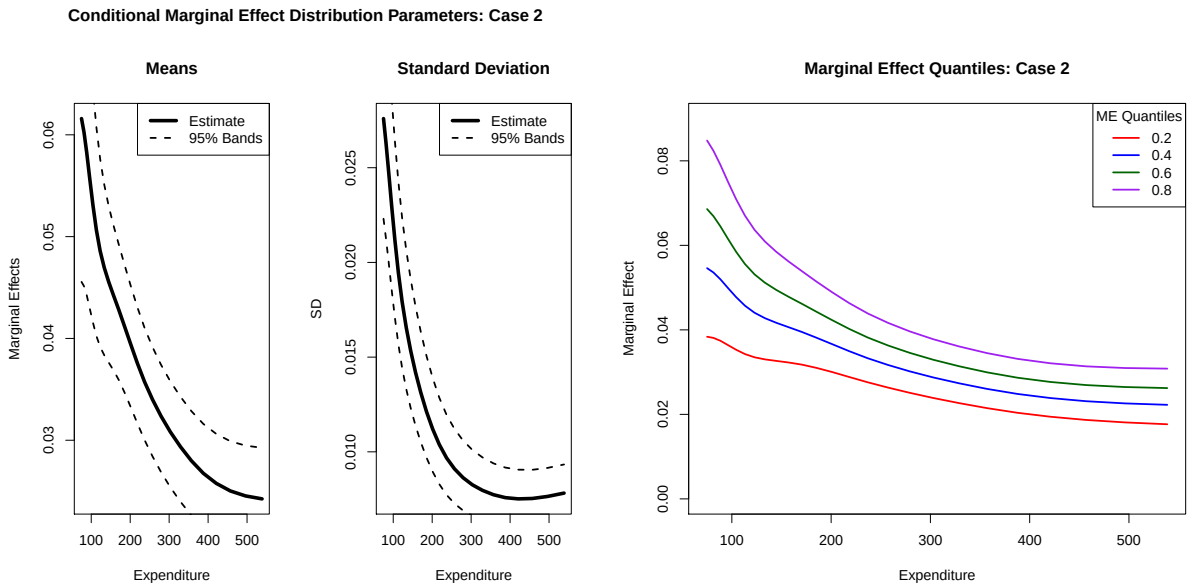

Figure 22: Estimates of the marginal effect of an additional dollar of expenditure on junk food using: $\alpha = 6$. For this sample, $N = 8,631$

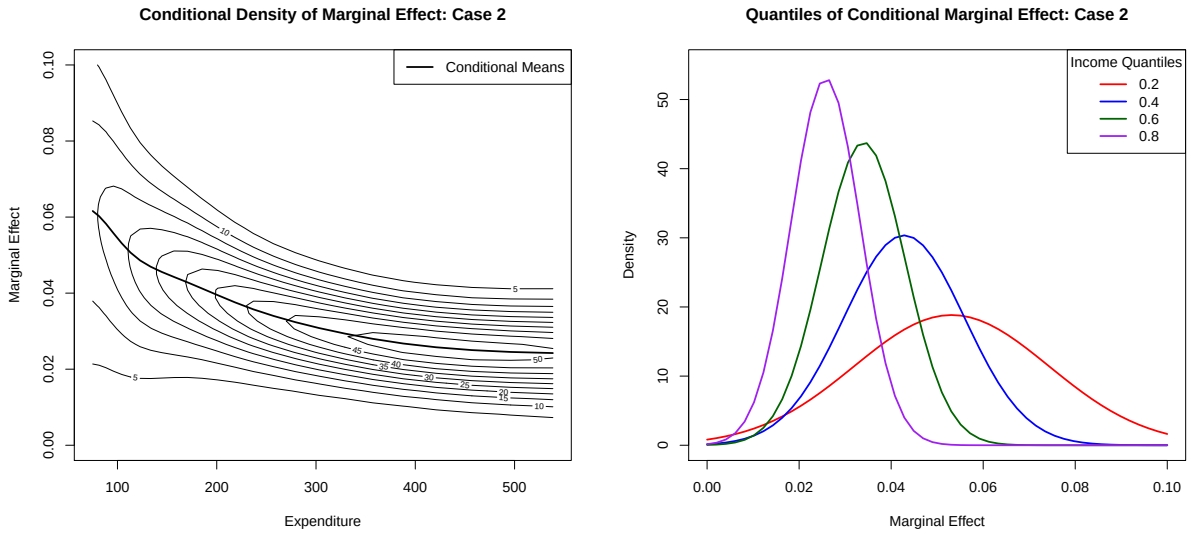

Figure 23: Conditional density and different expenditure quantiles of the estimates of marginal effect.

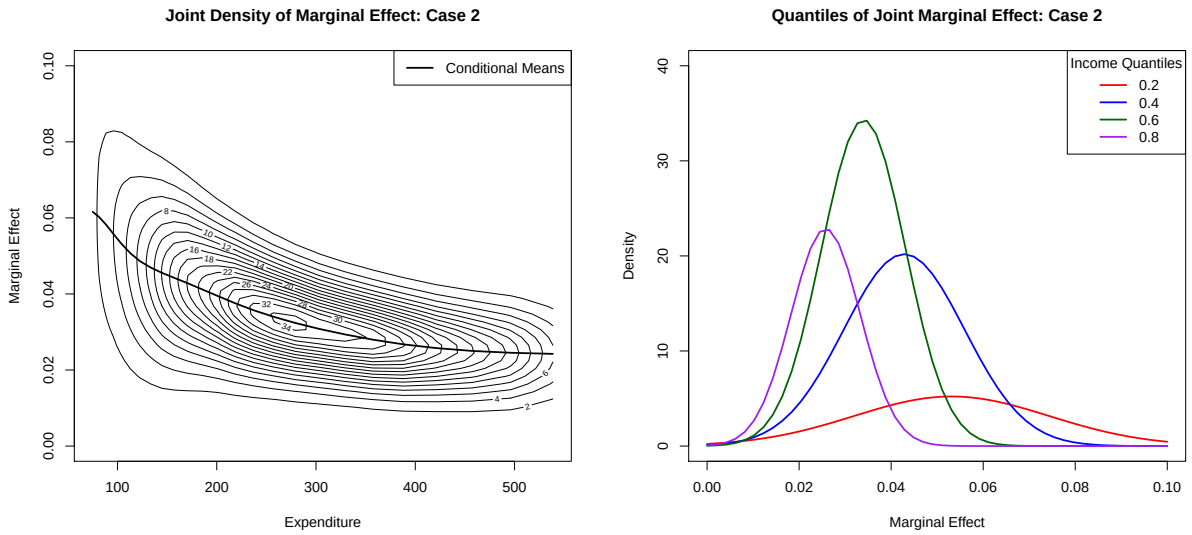

Figure 24: Joint distributions are calculated by multiplying the conditional distribution by the distribution of expenditure

References

[1] Aguiar, M., and E. Hurst (2007). Life-Cycle Prices and Production, *American Economic Review*, 97(5), 1533-1559.

[2] Allcott, H., Diamond, R., and J.-P. Dubé (2017). The Geography of Poverty and Nutrition: Food Deserts and Food Choices Across the United States, *National Bureau of Economic Research.*

[3] Altonji, J., and R. Matzkin (2005). Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors, *Econometrica*, 73, 1053–1103.

[4] Andersen, E. (1970). Asymptotic Properties of Conditional Maximum Likelihood Estimators, *Journal of the Royal Statistical Society Series B*, 32, 283-301.

[5] Arellano, M. (2003). Discrete Choice with Panel Data, *Investigaciones Economicas*, 27, 423-458.

[6] Arellano, M. and Bonhomme, S. (2012): Identifying Distributional Characteristics in Random Coefficients Panel Data Models, *Review of Economic Studies*, 79(3), 987-1020.

[7] Binkley, J. K., and A. Golub (2011). Consumer Demand for Nutrition Versus Taste in Four Major Food Categories, *Agricultural Economics*, 42(1), 65-74.

[8] Billingsley, P., *Probability and Measure, 3rd ed.*, 1995, Wiley, New York.

[9] Blaylock, J., Smallwood, D., Kassel, K., Variyam, J., and L. Aldrich (1999). Economics, Food Choices, and Nutrition, *Food Policy*, 24(2-3), 269-286.

[10] Bochner, S. and K. Chandrasekharan, *Fourier Transforms*, 1949, Princeton University Press.

[11] Carlson, A., and E. Frazão (2012). Are Healthy Foods Really More Expensive? It Depends How you Measure the Price, *United States Department of Agriculture Economic Information Bulletin*, 96.

[12] Cawley, J., Meyerhoefer, C., Biener, A., Hammer, M., and N. Wintfeld (2015). Savings in Medical Expenditures Associated with Reductions in Body Mass Index among US Adults with Obesity, by Diabetes Status, *Parmacoeconomics*, 33(7), 707-722.

[13] Chamberlain, G. (1982). Multivariate Regression Models for Panel Data, *Journal of Econometrics,* 18(1), 5 - 46.

[14] Chamberlain, G. (1984). Panel Data, in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics,* Vol. 2. New York: North Holland.

[15] Chamberlain, G. (2010). Binary Response Models for Panel Data: Identification and Information, *Econometrica*, 78(1), 159-168.

[16] Chen, S.E., Liu, J., and J. K. Binkley (2012). An Exploration of the Relationship between Income and Eating Behavior, *Agricultural and Resource Economics Review*, 41(1), 82-91.

[17] Chernozhukov, V., Fernandez-Val, I., Hahn, J., and W. Newey (2014). Identification and Estimation of Marginal Effects in Nonlinear Panel Models, MIT Working Paper.

[18] Chernozhukov, V., Fernandez-Val, I., Hoderlein, S., Holzmann, H., and W. Newey (2015). Quantile Derivatives and Panel Data, *Journal of Econometrics*, 188 (2), 378-392.

[19] Chernozhukov, V., Fernández-Val, I., and W. Newey (2019). Nonseparable multinomial choice models in cross-section and panel data. *Journal of Econometrics*. 211 (1), 104-116.

[20] Deaton, A., and J. Muellbauer (1980). An Almost Ideal Demand System, *The American Economic Review*, 70(3), 312-326.

[21] Delaigle, A., and Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *J. Amer. Statist. Assoc.* **102**, 1416–1426.

[22] Delaigle, A. and Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli* **14**, 562–579.

[23] Drewnowski, A., and N. Darmon (2005). The Economics of Obesity: Dietary Energy Density and Energy Cost, *The American Journal of Clinical Nutrition*, 82(1), 265S-273S.

[24] Drewnowski, A., and P. Eichelsdoerfer (2010). Can Low-Income Americans Afford a Healthy Diet, *Nutrition Today*, 44(6), 246.

[25] Drewnowski, A., and S. E. Spector (2004). Poverty and Obesity: the Role of Energy Density and Energy Costs, *The American Journal of Clinical Nutrition*, 79(1), 6-16.

[26] Einav, L., Leibtag, E., and A. Nevo (2010). Recording Discrepencies in Nielsen Homescan Data: Are They Present and Do They Matter? *Quantitative Marketing and Economics*, 8(2), 207-239.

[27] Evdokimov, K. (2010). Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. Princeton University, Working paper.

[28] Fan, J. (1991). On the optimal rates of convergence for non-parametric deconvolution problems. *Ann. Statist.* **19**, 1257–1272.

[29] Fernandez-Val, I., Freeman, H., and Weidner, M. (2021). Low-rank approximations of nonseparable panel models. *Econometrics Journal* **24**, C40–C77.

[30] Finkelstein, E. A., Trogdon, J. G., Cohen, J. W., and W. Dietz (2009). Annual Medical Spending Attributable to Obesity: Payer and Service-Specific Estimates, *Health Affairs*, 28(5), w822-w831.

[31] Golan, E. H., Steward, H., Kuckler, F., and D. Dong (2008). Can Low-Income Americans Afford a Healthy Diet?, *Amber Waves*, 6(5), 26-33.

[32] Graham, B., and J. Powell (2008). Identification and Estimation of 'Irregular' Correlated Random Coefficient Models, NBER Working Paper 14469.

[33] Hausman, J., Hall, B., and Z. Griliches (1984). Econometric Models for Count Data with an Application to the Patents-R&D Relationship,*Econometrica*, 52, 909-938.

[34] Hoderlein, S., and E. Mammen (2007). Identification of Marginal Effects in Nonseparable Models without Monotonicity, *Econometrica*, 75, 1513 - 1519.

[35] Hoderlein, S., and S. Mihaleva (2008). Increasing the Price Variation in a Repeated Cross Section, *Journal of Econometrics*, 247(2), 316-325.

[36] Hoderlein, S., and H. White (2012). Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects, *Journal of Econometrics*, 168 (2), 300–314 .

[37] Honore, B., and E. Kyriazidou (2000). Panel Data Discrete Choice Models with Lagged Dependent Variables, *Econometrica,* 68, 839-874.

[38] Imbens, G., and W. Newey (2009). Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity,*Econometrica*, 77, 1481-1512.

[39] Kyriazidou, E. (1997). Estimation of a Panel Data Sample Selection Model, *Econometrica*, 65, 1335-1364.

[40] Lewbel, A. (1989). Identification and Estimation of Equivalence Scales under Weak Separability, *The Review of Economic Studies*, 56(2), 311-316.

[41] Lin, X. (2018). Snap and Food Consumption Among the Elderly: a Collective Household Approach with Homescan Data, *National Bureau of Economic Research.*

[42] Manski, C. (1987). Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data, *Econometrica*, 55, 357-62.

[43] Murtazashvili, I. and J. Wooldridge (2008). Fixed Effects Instrumental Variables Estimation in Correlated Random Coefficient Panel Data Models, *Journal of Econometrics*, 142(1), 539-552.

[44] Olley, G. S. and Pakes, A. (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica*, 64, 1263-1297.

[45] Rasch. G. (1960). Probabilistic Models for some Intelligence and Attainment Tests, *Denmarks Paedagogiske Institut*, Copenhagen.

[46] Rasch, G. (1961). On the General Law and the Meaning of Measurement in Psychology,*Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. Berkeley: UC Press.

[47] Rider, J., Berck, P., and S. B. Villas-Boas (2012). Eating Healthy in Lean Times: The Relationship Between Unemployment and Grocery Purchasing Patterns. *https://www.aeaweb.org/conference/2016/retrieve.php?pdfid=1212*

[48] Robbins, J. M., Vaccarino, V., Zhange, H., and S. V. Kasl (2001). Socioeconomic Status and Type 2 Diabetes in African American and Non-Hispanic White Women and Men: Evidence from the Third National Health and Nutrition Examination Survey. *American Journal of Public Health* 91(1), 76.

[49] Williams, D., *Probability with Martingales*, 1991, Cambridge University Press.

[50] Wooldridge, J. (2002). *Econometrics of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

[51] Wooldridge, J. (2005). Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models, *The Review of Economics and Statistics*, 87, 385-390.

[52] World Health Organization (2015). Guideline: sugars intake for adults and children.