Estimating the Moments and the Distribution of Heterogeneous Marginal Effects Using Panel Data

Vladislav Morozov*

Job Market Paper

This version: October 31, 2023
Click here to see the latest version

Abstract

This paper considers estimation of the moments and the distribution of heterogeneous marginal effects using panel data. We impose no restrictions on the form or dimension of time-invariant heterogeneity. In this setting, we identify the mean, variance, higher-order moments, and the distribution of marginal effects using two periods of data. We propose simple nonparametric estimators for the moments and the distribution, and study their asymptotic properties. The moment estimators are consistent and asymptotically normal. For the distribution estimator, we establish consistency by developing novel results that connect the convergence of distributions to the convergence of their moments. We illustrate the methodology with an application to Engel curves for food at home. Our analysis of variance, higher moments, and the distribution of marginal effects reveals significant heterogeneity. In particular, some households have upward-sloping sections in their Engel curves for lower values of expenditures. In contrast, the average Engel curve is downward-sloping for all expenditure values, in line with the previous literature.

Keywords: heterogeneous marginal effects, panel data, nonparametric identification, moment problem

JEL: C14, C23, D12

*Department of Economics and Business, Universitat Pompeu Fabra (UPF) and Barcelona School of Economics (BSE); e-mail: vladislav.morozov@barcelonagse.eu.
Acknowledgments: I am deeply indebted to Christian Brownlees and Kirill Evdokimov for their support and guidance. I am also grateful to Stephane Bonhomme, Patrick Gagliardini, Joachim Freyberger, Jesus Gonzalo, Adam Lee, Arthur Lewbel, Oliver Linton, Geert Mesters, Katerina Petrova, Barbara Rossi, Robert Wojciechowski, Piotr Zwiernik, and the participants at the BSE Summer Forum 2023, IAAE 2023, BSE Jamboree, the 27th IPDC, EEA-ESEM 2023, and SMYE 2023.

1 Introduction

Marginal effects are a key object of interest in settings where continuous covariates are present. For example, labor supply elasticity with respect to the marginal tax rate plays a large role in the design of tax-and-benefit systems and the determination of the optimal size of the public sector. (Blundell, MaCurdy, and Meghir, 2007a; Saez, Slemrod, and Giertz, 2012).

Unobserved heterogeneity poses a challenge to nonparametric analysis of marginal effects. Under heterogeneity, interest centers on the moments and the distribution of marginal effects (Heckman, Smith, and Clements, 1997). However, analysis of the distribution of marginal effects is typically either limited to average effects (Hoderlein and Mammen, 2007, 2009; Hoderlein and White, 2012; Chernozhukov, Fernández-Val, Hoderlein, Holzmann, and Newey, 2015) or restricts unobserved heterogeneity to be scalar and to enter the model monotonically (Matzkin, 2003; Imbens and Newey, 2009; Evdokimov, 2010).

In this paper we identify and estimate the moments and the distribution of marginal effects in a class of nonparametric panel data models. Specifically, we consider a setting with time-invariant heterogeneity of unrestricted form and additive time-varying heterogeneity. The moments and the distribution are identified with two time periods of data for the subpopulation of stayers – units with the same value of the covariate in both periods. In contrast to the previous literature, the variance, higher-order moments, and the distribution of marginal effects are identified without restricting the dimension of time-invariant heterogeneity or how it affects the outcome. We propose simple estimators for the moments and the distribution of the marginal effects. Our estimators are easy to compute. In particular, our moment estimators are available in closed form and require no optimization. The distribution estimators require solving a small quadratic program. For both the moment and the distribution estimators we establish their asymptotic properties.

Specifically, we consider the following model. The continuous outcome Y_{it} is generated as

$$Y_{it} = m(X_{it}, W_{it}, \alpha_i) + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T \tag{1}$$

where m is an unknown structural function. (X_{it}, W_{it}) are observed explanatory variables, X_{it} is scalar and W_{it} may be a vector. α_i and u_{it} are unobserved heterogeneity. We do not restrict the form or the dimension of time-invariant heterogeneity α_i. In particular, α_i may be a vector, a function, or a more complex object. The idiosyncratic disturbance u_{it} satisfies $\mathbb{E}\left[u_{it}|X_{it}, W_{it}\right] = 0$. We work in a fixed effect framework and do not restrict the dependence structure between α_i and $\{(X_{it}, W_{it})\}_{t=1}^{T}$. The model may be interpreted as a generalized regression of the form $\mathbb{E}\left[Y_{it}|X_{it}, W_{it}, \alpha_i\right] = m(X_{it}, W_{it}, \alpha_i)$.

The goal of this paper is to identify and estimate the moments and the distribution of

$\partial_x m(x, w, \alpha_i)$ for a given value of (x, w). These moments and the distribution fully summarize the impact of a marginal change in the covariate X_{it}, accounting for the heterogeneity in the marginal effects induced by α_i.

Identification and estimation of the mean, variance, and higher-order moments of $\partial_x m(x, w, \alpha_i)$ exploits the panel structure of the data. First, we approximate the marginal effect $\partial_x m(x, w, \alpha_i)$ by a finite difference of the form $(m(x + h, w, \alpha_i) - m(x - h, w, \alpha_i))/2h$, $h > 0$. Second, we show that the kth moment of this finite difference are identified in model (1) for the population of near-stayers – units with $X_{i1} = x - h$ and $X_{i2} = x + h$. By taking $h \to 0$, we identify the kth moment of marginal effects for stayers as the limit of the kth moment of the finite difference. Further, the moments of interest are identified explicitly in terms of conditional moments of (Y_{i1}, Y_{i2}) and their derivatives if every component of the model is smooth in x in a suitable sense. By replacing population moments and derivatives with suitable local polynomial estimators, we obtain simple optimization-free estimators for all moments of $\partial_x m(x, w, \alpha_i)$.

Identification and estimation of the distribution of $\partial_x m(x, w, \alpha)$ builds on our results for the moments. The distribution is identified under the assumption that it is determined by its moments. We estimate the distribution of interest with a penalized sieve estimator. Specifically, the distribution is approximated with a flexible mixture. The weights of the mixture are obtained by approximately matching a finite number of estimated moments. The estimator is easy to compute and only requires solving a quadratic optimization problem. In order to obtain the properties of the distribution estimator, we develop novel theoretical results that connect the convergence of moments to the weak convergence of corresponding distributions. The objective function used is a sample version of a suitable metric that measures the distance between distributions in terms of their moments.

We characterize the asymptotic properties of our moment and distribution estimators. First, we show that the moment estimators are consistent for the moments of $\partial_x m(x, w, \alpha_i)$, and we obtain convergence rates. The convergence rates depend on the order of the moment. Specifically, the convergence rate for the kth moment matches the convergence rate of a local polynomial estimator for a kth derivative of a regression function. Second, we show that the moment estimators are asymptotically normally distributed, which allows inference on the moments of the marginal effects. Third, we establish consistency of our distribution estimator. We further obtain convergence rates if the true distribution is a finite mixture with nonparametrically modeled mixing probabilities.

As an application of our methodology, we estimate the moments and the distribution of the slopes of household-level Engel curves for food at home. Overall, our results for variance, higher-order moments and the distribution reveal significant heterogeneity in Engel curve

slopes. First, we reject marginal effect homogeneity for a broad range of expenditures. Second, for a fraction of households, their individual Engel curves have upward-sloping sections for lower values of expenditures, before becoming downwards-sloping at larger expenditures. Accordingly, Engel's law does not necessarily hold at the household level, although it holds on average. Last, the overall distribution of Engel curve slopes is approximately symmetric, light-tailed, and becomes more concentrated around the mean as expenditure rises.

Our paper is related to several strands of literature. First, we contribute to the literature on nonparametric analysis of marginal effects (Matzkin, 2003; Altonji and Matzkin, 2005; Hoderlein and Mammen, 2007, 2009; Imbens and Newey, 2009; Evdokimov, 2010; Hoderlein and White, 2012; Graham and Powell, 2012; Chernozhukov, Fernández-Val, Hahn, and Newey, 2013; Chernozhukov et al., 2015). As noted above, these papers fall into two principal groups: those focusing only on average effects and those imposing monotonicity at some stage to recover the full distribution of marginal effects. In contrast, we focus on identifying and estimating all the moments and the distribution with heterogeneity of unrestricted dimension. Accordingly, our approach is compatible with economic models which allow for multidimensional unobserved heterogeneity, whether in preferences, technology, innate abilities, or in other settings.

Second, models in the spirit of model (1) have previously been considered in Newey, Powell, and Vella (1999); Newey and Powell (2003); Chen, Dahl, and Khan (2005); Evdokimov (2010); Blundell, Horowitz, and Parey (2012), among others. We do not restrict the form or dimensions of α_i and focus on the moments and the distribution of the derivative of m.

Third, our empirical application contributes to the growing literature on nonparametric analysis of Engel curves (Banks, Blundell, and Lewbel, 1997; Blundell, Browning, and Crawford, 2003; Blundell, Chen, and Kristensen, 2007b; Imbens and Newey, 2009; Horowitz, 2011; Chen and Pouzo, 2012; Chernozhukov et al., 2015). Unlike the above papers, we focus on the distributional features of the slopes of individual Engel curves and go beyond the average slope, while allowing for unrestricted variation in preferences. Accordingly, we find that some households have upward-sloping sections in their Engel curve for lower values of expenditures, while confirming that the average Engel curve is downward-sloping for all expenditure values.

We also contribute to the literature on identifying and estimating distributions from their moments (Beran and Hall, 1992; Ormoneit and White, 1999; Wu, 2003; Mnatsakanov, 2008; Mnatsakanov and Hakobyan, 2009; Ponomareva, 2010; Wu and Yang, 2020). In contrast to these papers, we face the problem of constructing a distribution estimator based on moment estimators that converge at different nonparametric rates. We metrize weak convergence of distributions in terms of differences of moments, and use a sample version of the resulting

metric in order to construct a suitable sieve estimator for the distribution.

The rest of the paper is organized as follows. In section 2, we state our key identification results for the moments and the distribution of the marginal effects. In section 3 we propose simple estimators for these quantities. Sections 4 and 5 are devoted to the asymptotic properties of the moment and distribution estimators, respectively. Section 6 contains a simulation study. Finally, section 7 contains the empirical application. All proofs are collected in the Proof Appendix. We provide some additional results in the online Supplementary Appendix, available from the author's website

2 Identification

2.1 The Model

We begin by considering model (1) in more detail. Let the outcome Y_{it} be generated as

$$Y_{it} = m(X_{it}, \alpha_i) + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, 2 \tag{2}$$

where X_{it} is a continuous random variable, α_i is a random element of a suitable topological space $\mathbb{A}$, and $m(\cdot, \cdot)$ is an unknown function. m is twice differentiable in its first argument, and derivatives of m are bounded uniformly in both variables. For clarity of exposition, we suppress W_{it} and focus on the case of $T = 2$. Units i are independent. The structural function m is assumed to be invariant over time, ruling out unrestricted time trends. The smoothness assumption on m rules out discrete choice models.

An example of setup (2) is given by Engel curves for food at home – the example we empirically consider in section 7. Engel curves relate the household budget to the share of the budget spent on a given good. In this context X_{it} is the log total expenditure. α_i reflects time-invariant preferences, captured by a household-specific utility function. The function $m(X_{it}, \alpha_i)$ is the share of the total expenditure $\exp(X_{it})$ that a household with preferences α_i would like to spend on food at home. The observed share Y_{it} is equal to optimal share subject to idiosyncratic shocks to consumption u_{it}.

We are interested in the moments and the distribution of marginal effects $\partial_x m(x, \alpha_i)$. Together, these objects provide a complete summary of the impact of a marginal change in the covariate X_{it}, accounting for the heterogeneity in the marginal effects induced by α_i. Specifically, we consider these distributional features for the population of stayers at x – the units with $\{X_{i1} = X_{i2} = x\}$. Formally, let k be a positive integer and label the kth moment

of marginal effects for stayers at x as

$$\mu_k(x) = \mathbb{E}\left[\left(\partial_x m(x, \alpha_i)\right)^k | X_{i1} = X_{i2} = x\right].$$

The moments $\mu_k(x)$ serve a double purpose. First, they may be of interest directly. In particular, $\mu_1(x)$ is a local average response; this special case has been extensively studied in the literature (Altonji and Matzkin, 2005; Hoderlein and White, 2012; Chernozhukov et al., 2015). Moments with $k \geq 2$ may be used to compute summary statistics such as the variance of marginal effects, skewness, etc. Second, the identification argument for the distribution proceeds through moments $\mu_k(x)$. The distribution of $\partial_x m(x, \alpha_i)$ is determined by $\mu_k(x)$ for $k = 1, 2, \ldots$, as $\partial_x m(x, \alpha_i)$ is a bounded random variable under our assumptions. Such an approach is required, as the structural function m is not identified due to the unrestricted nature of α_i.

We impose the following assumption on the time-varying disturbances u_{it}:

Assumption 2.1 (Properties of u)**.** *(u_{i1}, u_{i2}) satisfies the following conditions:*

(i) *u_{it} may depend on (X_{i1}, X_{i2}), but only contemporaneously: $u_{i1}|(X_{i1}, X_{i2}) \stackrel{d}{=} u_{i1}|X_{i1}$ and $u_{i2}|(X_{i1}, X_{i2}) \stackrel{d}{=} u_{i2}|X_{i2}$.*

(ii) *Mean independence: $\mathbb{E}[u_{it}|X_{it} = x] = 0$ for all x*

(iii) *u_{i1} and u_{i2} are conditionally independent conditional on (X_{i1}, X_{i2})*

(iv) *Conditional independence of u_{it} and α_i: $u_{it} \perp \alpha_i | (X_{i1}, X_{i2})$.*

Assumption 2.1 imposes a number of conditional independence assumptions on (u_{i1}, u_{i2}). It is similar to the assumptions of Evdokimov (2010), to whom we refer for a detailed discussion. *(i)* allows the distribution of u_{it} to change with t and to depend on X_{it} in a potentially complex manner. However, *(ii)* imposes a location restriction on the distribution of u_{it}. Under assumption *(ii)* the model for Y_{it} may be viewed as the general regression model considered by Wooldridge (2010a) with $\mathbb{E}(Y_{it}|X_{it}, \alpha_i) = m(X_{it}, \alpha_i)$. Together, *(i)* and *(iii)* rule out using lagged values of Y_{it} as covariates (Hoderlein and White, 2012). Finally, under *(iv)* u_{it} and α_i may be dependent, but this dependence must flow only through (X_{i1}, X_{i2}).

We adopt a fixed effects perspective and impose no assumptions on the form and distribution of α_i. Additionally, α_i may exhibit complex dependence with (X_{i1}, X_{i2}). In the Engel curve example, preferences α_i may take form of a utility function, sampled from some unspecified probability law on a suitable space of functions. The expenditure level X_{it} may be dependent with the household preferences α_i.

We focus on the subpopulation of stayers — units with $X_{i1} = X_{i2} = x$ — for two principal reasons. First, stayers and near-stayers comprise a large proportion of the data in many applications (e.g. Graham and Powell (2012)). For example, in our application to Engel

curves for food at home, most household exhibit little-to-no variation in their expenditure levels between time periods (see fig. 6 in section 7). As such, stayers are a natural population of interest; nonparametric identification and estimation of structural parameters for stayers have been extensively studied in the literature (Evdokimov, 2010; Graham and Powell, 2012; Hoderlein and White, 2012; Chernozhukov et al., 2015). Second, the moments and the distribution of marginal effects for stayers are identified in our fixed effect setting.

2.2 Identification of Moments

We now turn to the identification of conditional moments $\mu_k(x)$ of the marginal effect $\partial_x m(x, \alpha_i)$ for the population of stayers at x, where x is ranges in some interval $I = [x_{lb}, x_{ub}]$.

In order to identify $\mu_k(x)$, consider the following intuitive argument. Let $h \neq 0$. By the mean value theorem and intermediate value theorems it holds that

$$\left(\frac{m(x+h,\alpha_i) - m(x-h,\alpha_i)}{2h}\right)^k = \left(\partial_x m(\tilde{x},\alpha_i)\right)^k, \quad \tilde{x} = \tilde{x}(h,\alpha_i) \in [x-h, x+h]. \tag{3}$$

Taking expectations conditional on the event $B_{x-h,2h} = \{X_{i1} = x-h, X_{i2} = x+h\}$, we obtain that for all $h \neq 0$

$$\Delta_k(x,h) \coloneqq \mathbb{E}\left[\left(\frac{m(x+h,\alpha_i) - m(x-h,\alpha_i)}{2h}\right)^k \middle| B_{x-h,2h}\right] = \mathbb{E}\left[\left(\partial_x m(\tilde{x},\alpha_i)\right)^k | B_{x-h,2h}\right] \tag{4}$$

The moment $\mu_k(x)$ is identified as the limit of $\Delta_k(x,h)$ as $h \to 0$ if the following two conditions hold:

(1) $\mathbb{E}[\left(\partial_x m(\tilde{x},\alpha)\right)^k | B_{x-h,2h}]$ converges to $\mu_k(x)$ as $h \to 0$.

(2) $\Delta_k(x,h)$ is identified for all $h > 0$ small enough.

For the first condition, note that $\mathbb{E}[\left(\partial_x m(\tilde{x},\alpha)\right)^k | B_{x-h,2h}]$ is approximately the kth moment of marginal effects for the near-stayers – the population with $X_{i1} = x-h$ and $X_{i2} = x+h$. If the stayers are the limit of near stayers as $h \to 0$ in a suitable sense, then (1) holds.

For the second condition, we make use of model (2). The kth conditional moment of $(Y_{i2} - Y_{i1})/(2h)$ is give by

$$\mathbb{E}\left[\left(\frac{Y_{i2} - Y_{i1}}{2h}\right)^k \middle| B_{x-h,2h}\right] = \mathbb{E}\left[\left(\frac{(m(x+h,\alpha_i) - m(x-h,\alpha_i)) + (u_{i2} - u_{i2})}{2h}\right)^k \middle| B_{x-h,2h}\right].$$

By conditional independence of u_{i2}, u_{i1} and α_i (assumption 2.1), the above is equal to

$$\sum_{j=0}^{k} \binom{k}{j} \Delta_j(x,h) \frac{\sum_{l=0}^{j} \binom{j}{l} \mathbb{E}\left[u_{i2}^l | X_{i2} = x+h\right] \mathbb{E}\left[u_{i1}^{j-l} | X_{i1} = x-h\right]}{(2h)^{k-j}}. \tag{5}$$

For $k = 1$ eq. (5) shows that $\Delta_1(x,h) = \mathbb{E}\left[(Y_{i2} - Y_{i1})/(2h)|B_{x-h,2h}\right]$ as $\mathbb{E}[(u_{i2} - u_{i1})|B_{x-h,2h}] = 0$ under assumption 2.1. Further, $\mathbb{E}\left[(Y_{i2} - Y_{i1})^k/(2h)|B_{x-h,2h}\right]$ is identified from the data for all k, $x \in I$ and $h > 0$ small enough if suitable near-stayers are present in the data. We conclude that $\Delta_1(x,h)$ may identified for all h small enough. Hoderlein and White (2012) obtain a similar result for the first moment in a somewhat more general model.

Higher-order moments $\Delta_k(x,h)$ can be identified recursively from (5) if suitable moments of u_{i2} and u_{i1} are identified. As lemma 2.1 below shows, these moments can be recovered from stayers with $\{X_{i1} = X_{i2} = x \pm h\}$, provided such stayers exist in the data.

To formalize the above logic, we impose two assumptions under which conditions (1) and (2) hold. First, we assume that the stayers are the limit of near-stayers in the following sense:

Assumption 2.2 (Continuity)**.** *(i) Let $F_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(\cdot)$ be the conditional law of α_i given $\{\boldsymbol{X}_i = \boldsymbol{x}\} \equiv \{X_{i1} = x_1, X_{i2} = x_2\}$. $F_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(\cdot)$ is well-defined for all $\boldsymbol{x}$.*[1] *(ii) $F_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(\cdot)$ is continuous in $\boldsymbol{x}$ with respect to the weak topology, that is, if $\boldsymbol{x}_n \to \boldsymbol{x}$, then $F_{\alpha|\boldsymbol{X}=\boldsymbol{x}_n}(\cdot) \Rightarrow F_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(\cdot)$*

Part *(ii)* is the substantial assumption. It assumes that stayers are not discontinuously different from near-stayers in their unobservables. A similar assumption is also made by Graham and Powell (2012) and Hoderlein and White (2012) in the context of a linear model.

Second, we assume that the covariate X_{it} is continuously distributed and that there exist stayers and near-stayers for all $h > 0$ small enough. Existence of stayers and near stayers then permits identification of the moments of (Y_{i1}, Y_{i2}) given (X_{i1}, X_{i2}) present in (5).

Assumption 2.3. *(i) Let $\boldsymbol{X}_i = (X_{i1}, X_{i2})$ and $\mathbb{X} \equiv \text{supp}(\boldsymbol{X}_i)$. $\boldsymbol{X}_i$ is continuously distributed on $\mathbb{X}$ with density $f_{\boldsymbol{X}}$. (ii) Let $I = [x_{lb}, x_{ub}]$. There exists some $\epsilon > 0$ such that the ϵ-neighborhood J of the set $\{(x,x), x \in I\} \subset \mathbb{R}^2$ is strictly contained in $\mathbb{X}$. (iii) The density $f_{\boldsymbol{X}}$ is uniformly bounded away from zero on J: $\inf_{\boldsymbol{x}\in J} f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$.*

The assumption that $\boldsymbol{X}_i$ is continuously distributed may be generalized to allow point masses of stayers at certain values of x. Such point masses may arise in settings where corner solution responses are present (Wooldridge, 2010b; Graham and Powell, 2012).

Before stating the formal results, we introduce some notation. Let $g(y_1, y_2)$ be a real-valued function of (y_1, y_2). Define

$$r_g(x_1, x_2) \coloneqq \mathbb{E}\left[g(Y_{i1}, Y_{i2})|X_{i1} = x_1, X_{i2} = x_2\right]. \tag{6}$$

For example, $r_{(y_2-y_1)^k}(x_1, x_2)$ stands for $\mathbb{E}\left[(Y_{i2} - Y_{i1})^k|X_{i1} = x_1, X_{i,2} = x_2\right]$. Further, define the following moments of $m(x, \alpha_i)$ and u_{it}:

$$\nu_{m^k}(x) \coloneqq \mathbb{E}\left[m^k(x, \alpha_i)|X_{i1} = X_{i2} = x\right],$$

[1] We refer the interested reader to Tjur (1975) for a construction of continuous disintegrations of measures.

$$\nu_{u_t^k}(x) := \mathbb{E}\left[u_{it}^k | X_{it} = x\right], \quad t = 1, 2$$
$$\nu_{(u_2-u_1)^k}(x,h) := \mathbb{E}\left[(u_{i2} - u_{i1})^k | X_{i1} = x - h, X_{i2} = x + h\right]$$
$$= \sum_{j=0}^{k} \binom{k}{j} \nu_{u_1^j}(x-h)\nu_{u_2^{k-j}}(x+h). \tag{7}$$

The following lemma shows that the moments of u_{i1} and u_{i2} in (5) are indeed identified. Further, the lemma provides recursive expressions for all the above moments and for $\Delta_k(x,h)$. These expressions also serve as a foundation for our moment estimators.

Lemma 2.1. *Let assumptions 2.1-2.3 and the technical regularity condition C.1 in the appendix hold. Let ϵ, I, and J be as in assumption 2.3.*

(1) Let $\sup_{x_1,x_2} \mathbb{E}\left[|g(Y_{i2}, Y_{i2})||X_{i1} = x_1, X_{i2} = x_2\right] < \infty$. Then $r_g(x_1, x_2)$ is identified for all $(x_1, x_2) \in J$.

(2) Let $\sup_x \mathbb{E}\left[|u_{it}|^K | X_{it} = x\right] < \infty$ for some positive integer K. Then the moments $\nu_{m^k}(x)$, $\nu_{u_t^k}(x)$, $\nu_{(u_2-u_1)^k}(x,h)$ are identified for $x \in (x_{lb} - \epsilon, x_{ub} + \epsilon)$ and all non-negative integers $k \leq K$. In particular, if $k = 0$, $\nu_{u_1^0}(x) = \nu_{u_2^0}(x) = 1$; if $k = 1$, $\nu_{u_1^1}(x) = \nu_{u_2^1}(x) = 0$ and $\nu_{m^1}(x) = r_{y_2}(x,x)$, and if $k \geq 2$

$$\nu_{u_1^k}(x) = r_{y_1^{k-1}(y_1-y_2)}(x,x) - \sum_{j=1}^{k-1} \binom{k-1}{j} \nu_{m^j}(x)\nu_{u_1^{k-j}}(x), \tag{8}$$

$$\nu_{u_2^k}(x) = r_{y_2^{k-1}(y_2-y_1)}(x,x) - \sum_{j=1}^{k-1} \binom{k-1}{j} \nu_{m^j}(x)\nu_{u_2^{k-j}}(x), \tag{9}$$

$$\nu_{m^k}(x) = r_{y_1^{k-1}y_2}(x,x) - \sum_{j=1}^{k-1} \binom{k-1}{j-1} \nu_{m^j}(x)\nu_{u_1^{k-j}}(x). \tag{10}$$

(3) The moments of the finite difference $\Delta_k(x,h)$ are identified for $x \in I$, $0 < |h| < \epsilon$ and non-negative integers $k \leq K$ as

$$\Delta_k(x,h) = \frac{r_{(y_2-y_1)^k}(x-h, x+h)}{(2h)^k} - \sum_{j=0}^{k-1} \binom{k}{j} \Delta_j(x,h) \frac{\nu_{(u_2-u_1)^{k-j}}(x,h)}{(2h)^{k-j}}. \tag{11}$$

With lemma 2.1 in hand, we can now state a full identification result for $\mu_k(x)$.

Theorem 2.2. *Let assumptions 2.1-2.3 and the technical regularity condition C.1 in the appendix hold. Let k be a positive integer such that $\sup_x \mathbb{E}[|u_{it}|^k | X_{it} = x] < \infty$ for $t = 1, 2$. Then $\mu_k(x) = \mathbb{E}\left[(\partial_x m(x, \alpha_i))^k | X_{i1} = X_{i2} = x\right]$ is identified for each $x \in I$ as $\lim_{h \to 0} \Delta_k(x,h)$.*

Theorem 2.2 shows that it is possible to identify all moments of marginal effects, and not just average effects, while allowing for multivariate unobserved heterogeneity. In particular, it is possible to identify the variance $(\mu_2(x) - \mu_1^2(x))$ of marginal effects, provided the second moments of u_{it} are finite.

Fundamentally, the identification arguments of theorem 2.2 rest on two pillars. First, the within variation of near stayers permits us to identify suitable moments $\Delta_k(x,h)$ of the finite difference for arbitrarily small values of h. Second, continuity of the conditional law of α_i allows us to express the moments $\mu_k(x)$ for stayers as the limit of $\Delta_k(x,h)$ as $h \to 0$.

Remark 1. The identification argument stemming from eq. (3) uses a two-sided finite difference. For a given point x, it requires the existence of units with $X_{i1} = x - h$ for positive h small enough. However, such units might not exist if x is, for example, the minimal wage in a labor supply application. In this case our argument can be generalized to use a one-sided differences that only requires units with $\{X_{i1} = x, X_{i2} = x + h\}$ or $\{X_{i1} = x - h, X_{i2} = x\}$.

2.3 Identification of the Distribution

The distribution of $\partial_x m(x, \alpha_i)$ is completely characterized by its moments $\mu_k(x)$ under our assumption that $\partial_x m(x,a)$ is bounded uniformly in a and x. Thus, if $\mu_k(x)$ is identified for all positive integers k, then so is the corresponding distribution. We formally state this result.

Theorem 2.3. *Let assumptions 2.1-2.3 and the technical regularity condition C.1 in the appendix hold. Let* $\sup_x \mathbb{E}[|u_{it}|^k | X_{it} = x] < \infty$ *for all positive integers* k*. Then the distribution of* $\partial_x m(x, \alpha_i)$ *conditional on* $\{X_{i1} = X_{i2} = x\}$ *is identified for each* $x \in I$*.*

As theorem 2.3 shows, the distribution of marginal effects may be identified in the presence of unobserved heterogeneity α_i of unrestricted form. This distribution may then be used to conduct distributional analysis of impacts of small changes in the covariate X_{it}. In addition, theorem 2.3 strengthens 2.2, as it permits identification of all moments of marginal effects.

Remark 2. If u_{i1} and u_{i2} possess only a finite number of moments, the distribution of marginal effects is partially identified. The identified set can be characterized with the Chebyshev–Markov–Stieltjes inequalities (Akhiezer, 1965, corollary 2.5.4, p. 66) which provide a bound on the difference between any two cdfs with the same first $2k$ moments.

Remark 3. The result of theorem 2.3 may also be applied to u_{it}. The distribution of u_{it} conditional on $\{X_{it} = x\}$ is identified in the setting of theorem 2.3 if Carleman's condition holds for it, that is, $\sum_{k=1}^{\infty} (\nu_{u_t^{2k}}(x))^{-1/2k} = \infty$.

3 Estimation

We now turn to estimation. In line with our identification results, we first discuss estimation of moments of the marginal effects $\partial_x m(x, \alpha_i)$. We then leverage those moment estimates to obtain a suitable estimator for the distribution of $\partial_x m(x, \alpha_i)$.

3.1 Estimation of Moments

In order to provide a simple estimator of $\mu_k(x)$, we first refine the argument of the previous section. Characterizing $\mu_k(x)$ as the limit of $\Delta_k(x,h)$ as $h \to 0$ allows identification under fairly general assumptions. The expressions of lemma 2.1 can then be used to construct an estimator for $\Delta_k(x,h)$ for any $h \neq 0$. $\Delta_k(x,h)$ would in turn approximate $\mu_k(x)$ as h tends to 0. However, in practice a positive but not large value of the h would be necessary. Taking h too small may lead to unstable estimates in finite samples, while taking h too large may mean that $\Delta_k(x,h)$ is far from $\mu_k(x)$. Thus, a rule for choosing h would be necessary – the discretization parameter h becomes a tuning parameter of the problem. To avoid this choice, we establish an alternative characterization for $\mu_k(x)$ in terms of kth derivatives of certain expectations, if these derivatives exist. Such a characterization then allows us to construct straightforward estimators that do not require choosing h.

To obtain an alternative characterization of $\mu_k(x)$, consider the following intuitive argument. For $x \in I$ and h small enough (including zero), define $D_k(x,h)$ as

$$D_k(x,h) \coloneqq \mathbb{E}\left[(m(x+h,\alpha_i) - m(x-h,\alpha_i))^k | X_{i1} = x-h, X_{i2} = x+h\right]. \tag{12}$$

Note that $D_k(x,h)$ is simply $(2h)^k \Delta_k(x,h)$. Thus $D_k(x,h)$ identified for all $x \in I$ and all h small enough by lemma 2.1.

Fix $x \in I$. By the mean value theorem $(m(x+h,\alpha_i) - m(x-h,\alpha_i))^k$ is equal to $(2h)^k(\partial_x m(\tilde{x},\alpha_i))^k$ for some point $\tilde{x} = \tilde{x}(h,\alpha_i) \in [x-h, x+h]$, as in eq. (3). Returning to eq. (12), we can further add and subtract $(2h)^k \mu_k(x)$ to obtain

$$D_k(x,h) = (2h)^k \mu_k(x) + \underbrace{(2h)^k \left[\mathbb{E}\left[(\partial_x m(\tilde{x},\alpha_i))^k | X_{i1} = x-h, X_{i2} = x+h\right] - \mu_k(x)\right]}_{=:\theta(h)}.$$

Differentiating the first term on the right hand side k times with respect to h yields $2^k k! \mu_k(x)$. At the same time, suppose that (1) $\theta(h)$ is k times differentiable for h in some neighborhood of 0; (2) the kth derivative of $\theta(h)$ at $h = 0$ is equal to 0. Then we can then obtain $\mu_k(x)$ from the kth derivative of D_k with respect to h at $h = 0$ as

$$\mu_k(x) = \frac{1}{2^k k!} \partial_h^k D_k(x,0). \tag{13}$$

In order to formalize the above logic, we impose several smoothness assumptions that we formally state in the proof appendix (assumptions D.1-D.4). Informally, the assumptions require that the model (2) be sufficiently smooth in x. We impose smoothness in (x_1, x_2) on the structural function m and the conditional distribution of u_{it} and α_i given $\{X_{i1} = x_1, X_{i2} = x_2\}$. Each component is assumed to be differentiable at least k times. We also assume that u_{it} is

distributed continuously conditional on X_{it}. Together, these assumptions imply that $D_k(x,h)$, $\theta(h)$, and the moments of eqs. (6)-(7) are all differentiable k times.

The following theorem formally states our differentiation-based identification result for $\mu_k(x)$ under the additional smoothness assumptions described above. It offers an explicit expression for $\mu_k(x)$, rather than the limit-based characterization of theorem 2.2.

Theorem 3.1. *Let assumptions 2.1-2.3 and C.1 hold. Further, let the smoothness assumptions D.1-D.4 in the appendix hold with $\tau \geq k$. Let $\sup_x \mathbb{E}[|u_{it}|^k | X_{it} = x] < \infty$ for $t = 1, 2$. Then*
(1) $D_k(x,h)$ is identified for all $x \in I$ and $h \in (-\epsilon, \epsilon)$ for I and ϵ of assumption 2.3.
(2) D_k is k times differentiable in h for $h \in (-\epsilon, \epsilon)$
(3) Eq. (13) holds.

In light of theorem 3.1, an estimator for $\mu_k(x)$ can be obtained by constructing an estimator for $\partial_h^k D_k(x,0)$. Such an estimator will directly target $\mu_k(x)$, rather than $\Delta_k(x,h)$, and will not require choosing h (informally, h is automatically set to 0).

To construct an estimator for $\partial_h^k D_k(x,h)$, observe that eq. (11) can be restated in terms of $D_k(x,h)$ as follows:

$$D_k(x,h) = r_{(y_2-y_1)^k}(x-h, x+h) - \sum_{j=0}^{k-1} \binom{k}{j} D_j(x,h) \nu_{(u_2-u_1)^{k-j}}(x,h). \tag{14}$$

Correspondingly, the kth derivative of $D_k(x,h)$ with respect to h is given by

$$\begin{aligned} \partial_h^k D_k(x,h) = {} & \partial_h^k r_{(y_2-y_1)^k}(x-h, x+h) \\ & - \sum_{j=0}^{k-1} \binom{k}{j} \left[\sum_{i=0}^{k} \binom{k}{i} \left(\partial_h^i D_j(x,h)\right) \left(\partial_h^{k-i} \nu_{(u_2-u_1)^{(k-j)}}(x,h)\right) \right]. \end{aligned} \tag{15}$$

$\partial_h^k D_k(x,h)$ depends on the kth derivative of the conditional expectation of $(Y_{i2} - Y_{i1})^k$. It also depends on all the derivatives of order at most k of $\nu_{(u_2-u_1)^j}(x,h)$ and $D_j(x,h)$ for $j < k$. These derivatives may be obtained by differentiating eqs. (7) and (14), respectively.

The estimator $\hat{\mu}_k(x)$ is now obtained by replacing all population objects in eq. (15) by suitable sample analogs. Algorithm 1 formally defines $\hat{\mu}_k(x)$. The required derivatives are obtained by differentiating equations (6), (7), (14), and the expressions of lemma 2.1, and replacing the population objects by sample analogs in the resulting equations.

Intuitively, estimation can be broken down into three steps. In the first step we estimate conditional moments of (Y_{i2}, Y_{i1}) and their derivatives. Formally, we estimate r_g and all of its derivatives up to order k, where the functions g form the set $\{(y_2-y_1)^j, y_1^{j-1}(y_1 - y_2), y_2^{j-1}(y_2-y_1), y_1^{j-1}y_2, j \in 1, 2, \ldots, k\}$. We propose estimating these objects simultaneously

Algorithm 1: Estimation of $\mu_k(x)$ for $k = 1, \ldots, K$

For $g \in \{(y_2 - y_1)^j, y_1^{j-1}(y_1 - y_2), y_2^{j-1}(y_2 - y_1), y_1^{j-1} y_2, j \in 1, 2, \ldots, k\}$ let $\widehat{\partial_h^l r_g(x-h, x+h)|_{h=0}}$ and $\widehat{\partial_h^l r_g(x \pm h, x \pm h)|_{h=0}}$ be the local polynomial estimators of order q of $\partial_h^l r_g(x-h, x+h)|_{h=0}$ and $\partial_h^l r_g(x \pm h, x \pm h)|_{h=0}$ for $l = 0, 1, \ldots K$ (see the Implementation Appendix)

Set for $l = 0, 1, \ldots, K$

$$\begin{aligned}
\widehat{\partial_h^l \nu_{u_1^1}(x-h)|_{h=0}} = \widehat{\hat{\nu}_{u_2^1}(x+h)|_{h=0}} &= 0, \\
\widehat{\partial_h^l D_1(x,0)} &= \widehat{\partial_h^l r_{(y_2-y_1)}(x-h, x+h)|_{h=0}}, \\
\widehat{\partial_h^l \nu_m(x \pm h)|_{h=0}} &= \widehat{r_{y_2}(x \pm h, x \pm h)|_{h=0}}
\end{aligned}$$

Set

$$\hat{\mu}_1(x) = \frac{1}{2}\widehat{\partial_h D_1(x,0)} \quad ,$$

if $K \geq 2$ **then**

  Set $k = 2$ and **while** $k \leq K$ **do**

    Estimate the lth derivative ($l = 0, 1, \ldots, K$) of the kth moment of u_1 using eq. (8)

$$\widehat{\partial_h^l \nu_{u_1^k}(x-h)|_{h=0}} = \widehat{\partial_h^l r_{y_1^{k-1}(y_1-y_2)}(x-h, x-h)|_{h=0}} - \sum_{j=1}^{k-1} \binom{k-1}{j} \left[\sum_{i=0}^{l} \binom{l}{i} \left(\widehat{\partial_h^i \nu_{m^j}(x-h)|_{h=0}} \right) \left(\widehat{\partial_h^{l-i} \nu_{u_1^{k-j}}(x-h)|_{h=0}} \right) \right]$$

    Estimate the lth derivative ($l = 0, 1, \ldots, K$) of the kth moment of u_2 using eq. (9)

$$\widehat{\partial_h^l \nu_{u_2^k}(x+h)|_{h=0}} = \widehat{\partial_h^l r_{y_2^{k-1}(y_2-y_1)}(x+h, x+h)|_{h=0}} - \sum_{j=1}^{k-1} \binom{k-1}{j} \left[\sum_{i=0}^{l} \binom{l}{i} \left(\widehat{\partial_h^i \nu_{m^j}(x+h)|_{h=0}} \right) \left(\widehat{\partial_h^{l-i} \nu_{u_2^{k-j}}(x+h)|_{h=0}} \right) \right]$$

    Estimate the lth derivative ($l = 0, 1, \ldots, K$) of $\nu_{(u_2-u_1)^k}(x,h)$ at $h = 0$:

$$\widehat{\partial_h^l \nu_{(u_2-u_1)^k}(x,0)} = \sum_{j=0}^{k} \binom{k}{j} \left[\sum_{i=0}^{l} \binom{l}{i} \left(\widehat{\partial_h^i \nu_{u_1^j}(x-h)|_{h=0}} \right) \left(\widehat{\partial_h^{l-i} \nu_{u_2^{p-j}}(x+h)|_{h=0}} \right) \right].$$

    Estimate the lth derivative ($l = 0, 1, \ldots, K$) of $D_k(x,h)$ at $h = 0$ using eq. (14):

$$\widehat{\partial_h^l D_k(x,0)} = \widehat{\partial_h^l r_{(y_2-y_1)^k}(x-h, x+h)|_{h=0}} - \sum_{j=0}^{k-1} \binom{k}{j} \left[\sum_{i=0}^{l} \binom{l}{i} \left(\widehat{\partial_h^i D_j(x,h)} \right) \left(\widehat{\partial_h^{l-i} \nu_{(u_2-u_1)^{(k-j)}}(x,h)} \right) \right].$$

    Set the moment estimator

$$\hat{\mu}_k(x) = \widehat{\partial_h^k D_k(x,0)}.$$

    if $k<K$ **then**

      Estimate the lth derivative ($l = 0, 1, \ldots, K$) of the kth moment of m using eq. (10)

$$\widehat{\partial_h^l \nu_{m^k}(x \pm h)|_{h=0}} = \widehat{r_{y_1^{k-1} y_2}(x \pm h, x \pm h)|_{h=0}} - \sum_{j=1}^{p-1} \binom{k-1}{j-1} \left[\sum_{i=0}^{l} \binom{l}{i} \left(\widehat{\partial_h^i \nu_{m^j}(x \pm h)|_{h=0}} \right) \left(\widehat{\partial_h^{l-i} \nu_{u_1^{p-j}}(x \pm h)|_{h=0}} \right) \right].$$

      Set $k = k + 1$

    end

  end

end

by running a local polynomial regression of order q (LP(q)) of $g(Y_{i1}, Y_{i,2})$ on $(X_{i1}, X_{i,2})$, where $q \geq k+1$, although any other nonparametric approach may be used (see Fan and Gijbels (1996) for a reference on LP estimation). An LP(q)-based approach is practically appealing, as all the necessary estimators exist in closed form and require no optimization. In the Implementation Appendix we propose a convenient approach that allows computing all the necessary derivatives for all the functions g simultaneously with only three applications of LP(q) regression. Second, we estimate the moments of u_{i1}, u_{i2} and $m(x, \alpha_i)$ and their derivatives of order up to k according to algorithm 1. By replacing population moments with sample equivalents in lemma 2.1, we obtain estimators for $\nu_{u_t^k}$ and ν_{m^k} and their derivatives. The estimated derivatives of $\nu_{u_t^k}$ are then combined to form an estimator for $\partial_h^l \nu_{(u_2-u_1)^j}(x,h)$. Third and last, estimators for $\partial_h^l D_j(x,h), l = 0, 1, \ldots, k$, $j = 1, \ldots, k$ are formed by evaluating eq. (15) recursively starting from $j = 1$ until reaching $j = k$. The estimator $\hat{\mu}_k(x)$ is obtained by dividing the estimated value for $\partial_h^k D_k(x, 0)$ by $(2^k k!)$.

The estimator $\hat{\mu}_k(x)$ is consistent and asymptotically normal, as we establish in the section 4, provided that the smoothing bandwidth of the LP(q) estimators is chosen appropriately. Suitable confidence intervals for $\mu_k(x)$ may be constructed by nonparametric bootstrap, recomputing $\hat{\mu}_k(x)$ in bootstrap samples according to algorithm 1.

Remark 4 (Choosing the bandwidth parameter of the LP estimators)**.** The moments of interest $\mu_k(x)$ are estimated most accurately when the first step estimators of moments of (Y_{i1}, Y_{i2}) are as accurate as possible, as algorithm 1 implies. In turn, the precision of the first-step LP estimators is determined by the corresponding smoothing bandwidth. The MSE optimal bandwidth(s) may be obtained with the data-driven method of Charnigo and Srinivasan (2015), who propose a generalized C_p approach targeted at nonparametric estimation of derivatives of multivariate functions.

Remark 5 (Estimation based on moments of the finite difference)**.** Alternative moment estimators may be formed by estimating $\Delta_k(x,h)$ using a sample version of eq. (11). There are two key differences between the estimator $\hat{\mu}_k(x)$ proposed in this section and an estimator targeting $\Delta_k(x,h)$. First, estimating $\Delta_k(x,h)$ requires choosing a positive value for h that controls the bias-variance trade-off described at the beginning of this section. This difficulty is not present in $\hat{\mu}_k(x)$. Second, $\hat{\mu}_k(x)$ imposes somewhat stronger smoothness properties. Theorem 3.1 requires every component of the model to be differentiable at least k times in x. Further, at least $k+3$ derivatives are needed to establish convergence rates and asymptotic normality of $\hat{\mu}_k(x)$ (see section 4). In contrast, regardless of k, only 3 derivatives are needed to show consistency and asymptotic normality of an estimator based on $\Delta_k(x,h)$, as we formally show in the Supplementary Appendix.

3.2 Estimation of the Distribution

We now turn to estimating the distribution function $F_0(v|x)$ of marginal effects. We outline the general approach, define our estimators, and offer some discussion. Technical details and asymptotic properties of the estimators are deferred to section 5.

We differentiate between two problems: estimating $F_0(v|x_0)$ at one fixed point x_0 and estimating the bivariate function $F_0(v|x)$ as x varies in $I = [x_{lb}, x_{ub}]$. Although the two goals are equivalent in population, they are not necessarily equivalent in finite samples. Estimating the full conditional cdf for an interval of values for x yields a cdf smooth in the conditioning argument, in line with the smoothness assumption 2.2. Further, this approach permits quick evaluation of the estimates for any value in $x \in I$, and does not require reestimating the distribution at every point separately. However, it is less flexible in x.

In both cases we approximate the distribution with a finite mixture. The mixture weights are chosen by projecting estimated moments onto the space of moments of corresponding mixtures. The key difference between the two cases lies in how the mixing probabilities are modeled. In the case of estimation at one point, the mixing probabilities are just a real vector. In contrast, in case of estimating on an interval the mixing probabilities are allowed to vary as a function of x. We approximate these functions of x using Bernstein polynomials.

Estimation of $F_0(v|x_0)$ for x_0 fixed Consider first estimating $F_0(v|x_0)$ for a fixed $x_0 \in I$.

We can model the unknown distribution function $F_0(v|x_0)$ using the following mixture approximation. Let p be a positive integer. Let $v_{1,p} < v_{2,p} < \cdots < v_{p,p}$ be the (fixed) mixture centers (see remark 6 below). Let Ψ be a smooth reference cdf; in section 5 we specify how to pick Ψ to obtain nonparametric consistency. Finally, let $\boldsymbol{\gamma}$ be a p-vector that satisfies $\sum_{i=1}^{p} \gamma_i = 1, \gamma_i \geq 0$; $\boldsymbol{\gamma}$ is the vector of mixing probabilities. We model $F_0(\cdot|x_0)$ with a finite mixture distribution with p components and mixing weights $\boldsymbol{\gamma}$ as

$$\Lambda_p(v|\boldsymbol{\gamma}) = \sum_{j=1}^{p} \gamma_j \Psi(v - v_{j,p}). \tag{16}$$

Modeling $F_0(v|x_0)$ using $\Lambda_p(v|\boldsymbol{\gamma})$ may be interpreted nonparametrically or parametrically. First, consider a nonparametric perspective. In this case there may be no value of p or $\boldsymbol{\gamma}$ such that $F_0(v|x_0) = \Lambda_p(v|\boldsymbol{\gamma})$. However, the distributions $\Lambda_p(v|\boldsymbol{\gamma})$ can approximate $F_0(v|x_0)$ arbitrarily well if p is taken large enough, and Ψ, $v_{j,p}$, and $\boldsymbol{\gamma}$ are selected as in section 5.2. In this interpretation the functions $\Lambda_p(v|\boldsymbol{\gamma})$ form increasingly complex spaces as p grows. These spaces act as sieve spaces for $F_0(v|x_0)$.[2] Alternatively, in the parametric case

[2]Such sieves are known as mixture of experts in statistics (Zeevi and Meir, 1997; Li and Barron, 1999; Norets, 2010).

$F_0(v|x_0) = \Lambda_p(v|\boldsymbol{\gamma}_0)$ for some p and some vector $\boldsymbol{\gamma}_0$, and the model is exact.

The mixture weights $\boldsymbol{\gamma}$ are selected in the same manner regardless of the interpretation of $\Lambda_p(v|\boldsymbol{\gamma})$. Let K be a positive integer and $\tilde{\mu}_k(x_0)$ be a consistent estimator for $\mu_k(x_0)$, $k = 1, \ldots, K-1$. $\tilde{\mu}_k(x_0)$ may be the estimator $\hat{\mu}_k(x)$ of section 3.1, an estimator of $\Delta_k(x, h)$, etc. Let $\lambda_N \geq 0$. Define

$$\begin{aligned}\tilde{Q}_N(\boldsymbol{\gamma}|x_0) &= \sum_{k=1}^{K-1} \frac{1}{k!} \left[\tilde{\mu}_k(x_0) - \int v^k \Lambda_p(dv|\boldsymbol{\gamma})\right]^2 \\ &\equiv \sum_{k=1}^{K-1} \frac{1}{k!} \left[\tilde{\mu}_k(x_0) - \sum_{j=1}^{p} \gamma_j \int v^k \Psi\left(d(v - v_{j,p})\right)\right]^2. \end{aligned} \tag{17}$$

The pointwise estimator at x_0 is defined as

$$\tilde{F}_N(v|x_0) = \Lambda_p(v|\tilde{\boldsymbol{\gamma}}) \tag{18}$$

where $\tilde{\boldsymbol{\gamma}}$ is defined as

$$\tilde{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}: \sum_{j=1}^{p} \gamma_j = 1, \gamma_j \geq 0}{\arg\min} \tilde{Q}_N(\boldsymbol{\gamma}|x_0) + \lambda_N \sum_{j=1}^{p} \gamma_j^2, \tag{19}$$

where we assume that $K \geq p$ or $\lambda_N > 0$. The objective function is strictly convex in $\boldsymbol{\gamma}$ on the constraint set and $\tilde{\boldsymbol{\gamma}}$ is the unique minimizer of the criterion function.

The interpretation of the estimator $\tilde{F}_N(v|x_0)$ and the choice of K reflect the interpretation of $\Lambda_p(v|\boldsymbol{\gamma})$. In the nonparametric case, $\tilde{F}_N(v|x_0)$ is a penalized sieve estimator. K and p both tend to infinity as sample size N increases. The function $\tilde{Q}_N(\boldsymbol{\gamma}|x_0)$ aims to measure the distance between the true distribution function and the sieve approximant in terms of their moments. As N increases, each $\tilde{\mu}_k(x_0)$ converges to the true $\mu_k(x_0)$; as K increases, more and more moments are matched. Thus, in the limit the problem of minimizing $\tilde{Q}_N(\boldsymbol{\gamma}|v_0)$ becomes the problem of minimizing this distance to the true distribution of interest; this property lies at the root of asymptotic properties of $\tilde{F}_N(v|x_0)$ in the nonparametric case. In the parametric case, $\tilde{F}_N(v|x_0)$ is a method of moments estimator for the mixture distribution $F_0(v|x_0)$. In this case it is sufficient to set $K = p - 1$ and $\lambda_N = 0$.

The estimation procedure itself has a projection interpretation. The estimated moments $\tilde{\mu}_k(x_0)$ may not be a valid sequence of moments in finite samples in the sense that there is no distribution corresponding to them.[3] Optimizing $\tilde{Q}_N(\boldsymbol{\gamma}|x_0)$ finds the distribution $\Lambda_p(v|\boldsymbol{\gamma})$ that best matches the first K moments.[4] The resulting estimator $\tilde{F}_N(v|x_0)$ is always a valid

[3] See ch. 10 of Schmüdgen (2017) for precise theoretical conditions for a finite vector to be a vector of moments of some distribution.

[4] An alternative method is discussed by Wu and Yang (2020) in the context of estimating the (atomic)

distribution function.

The estimator $\tilde{F}_N(v|x_0)$ has several desirable properties. First, it is easy to compute and evaluate. The problem of finding $\tilde{\boldsymbol{\gamma}}$ is effectively a simple OLS or ridge regression with a convex constraint, where the moments $\tilde{\mu}_k(x_0)$ play the role of the dependent variable. Further, K will typically be fairly small in practice (we explore values 3-8 in our simulation study and the empirical application). Second, the estimator is consistent. $\tilde{F}_N(v|x_0)$ converges uniformly to $F_0(v|x_0)$, as we show in section 5. Moreover, in the parametric case $\tilde{F}_N(v|x_0)$ converges to $F_0(v|x_0)$ in total variation.

Problem (19) is regularized in two ways. First, we restrict our estimate to be a valid distribution function by the requirement that $\gamma_j \geq 0$ and $\sum_{j=1}^{p} \gamma_j = 1$. Second, we potentially include a Tikhonov regularization term in the objective functions itself. Its inclusion ensures that the function is strictly convex even if $p > K$, allowing the functions (16) to have good approximation properties, at the price of some regularization bias.

Remark 6 (Choice of centers $v_{j,p}$)**.** The centers $\{v_{j,p}\}_{j=1}^{p}$ may be chosen based on the estimated mean and variance of marginal effects. Let $\hat{M} > 0$ be a constant such that $F_0(\hat{\mu}_1(x_0) + \hat{M}|x_0) - F_0(\hat{\mu}_1(x_0) - \hat{M}|x_0) \geq 1 - \beta + o_{a.s.}(1)$ for some fixed level β. $\hat{M}$ may be determined by Chebyshev's inequality, another inequality based on the first two moments, the 68-95-99.7 rule, or some other approach. Partition the interval $[\hat{\mu}_1(x_0) - \hat{M}, \hat{\mu}_1(x_0) + \hat{M}]$ into p equal-length subintervals and let $v_{j,p}$ be the center of the jth interval. The number of components p should be taken so that the distance between centers is not too large, for example, not exceeding some multiple of $\hat{\sigma}(x_0) \coloneqq \sqrt{\hat{\mu}_2(x_0) - \hat{\mu}_1^2(x_0)}$. As we show in section 5, such an approach allows the functions $\Lambda_p(v|\boldsymbol{\gamma})$ to approximate $F_0(v|x_0)$ arbitrarily well as $p \to \infty$. We note that in the parametric case it is possible to identify and estimate the centers $v_{j,p}$ as parameters of the problem using standard techniques of theory of finite mixtures. We do not pursue this further.

Remark 7. Support of $F_0(v|x_0)$ may itself be estimated using the moments $\mu_k(x_0)$. Kazemi, Shahdoosti, and Mnatsakanov (2017) provide estimators based on the ratio of $(k+1)$st and kth moments as $k \to \infty$.

Estimation of $F_0(v|x)$ for x ranging in I If interest lies in recovering $F_0(v|x)$ for a range of values of x, a more refined approximation is needed. We approximate $F_0(v|x)$ with a finite mixture distribution with mixing probabilities that smoothly depend on x as $x \in I = [x_{lb}, x_{ub}]$. Let Ψ be as before. Fix p_v, p_x be positive integers. Let $\boldsymbol{\gamma}$ be a $p_v \times (p_x + 1)$ matrix with

mixing distribution for a Gaussian location mixture. They propose to first project the estimated noisy moments into the moment space, and then fitting an atomic distribution that matches the moments exactly.

(j,l)th element $\gamma_{j,l}$. Let $v_{1,p_v} < v_{2,p_v} < \cdots < v_{p_v,p_v}$ be the mixture centers. Define

$$\Lambda_{p_v,p_x}(v|x,\boldsymbol{\gamma}) = \sum_{j=1}^{p_v}\left[\sum_{l=0}^{p_x}\gamma_{j,l}b_{l,p_x}(x)\right]\Psi\left(v - v_{j,p_v}\right), \tag{20}$$

where $\sum_{j=1}^{p_v}\gamma_{j,l} = 1$ for $l = 0, 1, \ldots p_x$, $\gamma_{j,l} \geq 0$ for all j, l and

$$b_{l,p_x}(x) = \binom{p_x}{l}\left(\frac{x - x_{lb}}{x_{ub} - x_{lb}}\right)^l\left(\frac{x_{ub} - x}{x_{ub} - x_{lb}}\right)^{p_x - l}.$$

Λ_{p_v,p_x} may be interpreted from a nonparametric or a semi-nonparametric perspective, similarly to Λ_p. In the first case, the distributions $\Lambda_{p_v,p_x}(v|x,\boldsymbol{\gamma})$ can approximate the bivariate function $F_0(v|x)$ arbitrarily well as $p_v, p_x \to \infty$. A full discussion is provided in section 5.2. In the semi-nonparametric case, $F_0(v|x)$ can be represented as $\sum_{j=1}^{p_v}\rho_{0,j}(x)\Psi\left(v - v_{j,p_v}\right)$ for some mixing probabilities $\rho_{0,j}(x)$, where the function $\rho_{0,j}(x) : I \to [0,1]$ satisfy $\rho_{0,j}(x) \geq 0$, $\sum_{j=1}^{p_v}\rho_{0,j}(x) = 1$ for all x. In this case $F_0(v|x)$ is a finite mixture in v.

In both cases the mixing probabilities are treated nonparametrically and approximated using Bernstein polynomials $\{b_{l,p_x}\}_{l=0}^{p_x}$ of a growing order p_x. The key advantage of such an approximation is that the approximated mixing probabilities $[\sum_{l=0}^{p_x}\gamma_{j,l}b_{l,p_x}(x)]$ are non-negative and sum to 1 under simple conditions on the coefficients $\gamma_{j,l}$.

We extend the objective function (18) to the interval case by integrating it with respect to x using some measure π. π may be the estimated distribution of stayers as $x \in I$, the Lebesgue measure or some other measure. Let K be a positive integer and $\lambda_N^I \geq 0$. Define

$$\begin{aligned}\hat{Q}_N(\boldsymbol{\gamma}) &= \int_I \sum_{k=1}^{K-1}\frac{1}{k!}\left[\tilde{\mu}_k(x) - \int t^k\Lambda_{p_v,p_x}(dv|x,\boldsymbol{\gamma})\right]^2\pi(dx) \\ &\equiv \int_I \sum_{k=1}^{K-1}\frac{1}{k!}\left[\tilde{\mu}_k(x) - \int t^k\sum_{j,l}\gamma_{j,l}b_{l,p_x}(x)\Psi\left(d(v - v_{j,p_v})\right)\right]^2\pi(dx).\end{aligned} \tag{21}$$

We define the interval estimator as

$$\hat{F}_N(v|x) = \Lambda_{p_v,p_x}(v|x,\hat{\boldsymbol{\gamma}}) \tag{22}$$

where the weights are determined as

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}:\gamma_{j,l}\geq 0, \sum_{j=1}^{p_v}\gamma_{j,l}=1\ \forall l}{\arg\min}\ \hat{Q}_N(\boldsymbol{\gamma}) + \lambda_N^I\sum_{j,l}\gamma_{j,l}^2. \tag{23}$$

The interpretation of $\hat{F}_N(v|x)$ is similar to that of $\tilde{F}_N(v|x_0)$. In the nonparametric case, $\hat{F}_N(v|x)$ is a penalized sieve estimator; K, p_v tend to infinity as N increases. In the

semi-nonparametric case, p_v is held fixed and we may take $K = p_v$. In both cases $p_x \to \infty$.

The estimator (22) shares the appealing properties of estimator (18). First, it is straightforward to compute. By interchanging the sums and the integral, we see that the problem of finding $\hat{\gamma}$ in eq. (23) is again an OLS or a ridge regression with convex constraints, though in this case the constraints are somewhat more complex to ensure that $\hat{F}_N(v|x)$ is a valid distribution for all $x \in I$. In the Implementation Appendix we offer a convenient representation of (23) as a quadratic program. Second, $\hat{F}_N(v|x)$ is consistent for $F_0(v|x)$ in the sense that $\hat{F}_N(v|x)$ is the estimated cdf, then $\int_I \sup_v |\hat{F}_N(v|x) - F_0(v|x)|\pi(dx) \to 0$. Further, we establish convergence in total variation in the semi-nonparametric case.

4 Asymptotic Properties of Moment Estimators

We now turn to the properties of the moment estimator $\hat{\mu}_k(x)$ of section 3.1. We begin by establishing the convergence rate of $\hat{\mu}_k(x)$ to $\mu_k(x)$ uniformly in x as x varies in $I = [x_{lb}, x_{ub}]$. We then establish asymptotic normality of $\hat{\mu}_k(x)$, enabling inference on $\mu_k(x)$.

The first step of constructing $\hat{\mu}_k(x)$ involves running local polynomial regressions (Y_{i1}, Y_{i2}) on (X_{i1}, X_{i2}) using a kernel ψ_{LP} (detailed expressions are provided in the Implementation Appendix). The kernel ψ_{LP} is required to satisfy the following assumption.

Assumption 4.1. *The kernel $\psi_{LP} : \mathbb{R}^2 \to \mathbb{R}$ satisfies: (i) $\psi_{LP}(v_1, v_2)$ has bounded support; (ii) $\psi_{LP}(v_1, v_2)$ is Lipschitz continuous in (v_1, v_2) on $\mathbb{R}^2$; (iii) $\int \psi_{LP}^2(v_1, v_2)dv_1dv_2 < \infty$; (iv) ψ_{LP} is a second order kernel in the sense that $\int \psi_{LP}(v_1, v_2)dv_1dv_2 = 1$ and $\int v_1\psi_{LP}(v_1, v_2)dv_1dv_2 = \int v_2\psi_{LP}(v_1, v_2)dv_1dv_2 = 0$, $\psi_{LP}(v_1, v_2) \geq 0$.*

The following theorem quantifies the rate of convergence of $\hat{\mu}_k(x)$ to $\mu_k(x)$ uniformly as x ranges through the interval I.

Theorem 4.1. *Let assumptions 2.1-2.3, 4.1, and C.1 hold. Let the smoothness assumptions D.1-D.4 in the appendix hold with $\tau \geq q + 2$ and let $q \geq k + 1$. For some $\delta > 0$ let $\sup_x \mathbb{E}\left[u_{it}^{2k+\delta}|X_{it} = x\right]$, $t = 1,2$. Let the first-step LP(q) regressions of $g(Y_{i1}, Y_{i2})$ on (X_{i1}, X_{i2}) use the kernel ψ_{LP} with a diagonal bandwidth matrix $\text{diag}\{s, s\}$ for some (common) smoothing bandwidth s, where $g \in \{(y_2 - y_1)^j, y_1^{j-1}(y_1 - y_2), y_2^{j-1}(y_2 - y_1), y_1^{j-1}y_2, j \in 1, 2, \ldots, k\}$. Let $s \to 0$ and $\log(N)/Ns^{2+2k} \to 0$. Then*

$$\sup_{x \in I} |\hat{\mu}_k(x) - \mu_k(x)| = O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2k}}} + s^{q-k+1}\right).$$

As theorem 4.1 shows, the convergence rate of $\hat{\mu}_k(x)$ is the same as the convergence rate of an LP(q) estimator for the kth derivative (see Stone (1982) and Masry (1996a)). Such a

result is unsurprising in light of eqs. (13) and (15): $\hat{\mu}_k(x)$ is based on the estimator of the kth derivative of D_k, which in turn depends on the kth derivatives of conditional expectations of (Y_{i1}, Y_{i2}) given $(X_{i1}, X_{i,2})$, as algorithm 1 shows.

The smoothness assumptions in theorem 4.1 are slightly stronger than the corresponding assumptions for identification in theorem 3.1. All components of the model are assumed to be differentiable at least $(q+2)$ times, where q is the order of the local polynomial fitted. q itself is required to satisfy $q \geq (k+1)$ in order to estimate kth derivatives. Correspondingly, the theorem assumes the existence of at least $(k+3)$ derivatives.

The optimal rate of theorem 4.1 is decreasing in moment order k for a given degree of smoothness q. If $s \sim (\log(N)/N)^{1/2(q+2)}$, then

$$\sup_{x \in I} |\hat{\mu}_k(x) - \mu_k(x)| = O_{a.s.}\Big((\log(N)/N)^{(q-k+1)/(2(q+2))}\Big).$$

Remark 8. Theorem 4.1 assumes that all the first-step LP(q) regressions use a common smoothing bandwidth s. This assumption might be relaxed as follows. Associate a bandwidth s_g to each function g and estimate the corresponding LP(q) regression using s_g. The rate result is be driven by the estimator with the highest variance and the estimator with the highest bias in sense that $\sup_{x \in I} |\hat{\mu}_k(x) - \mu_k(x)| = O_{a.s.}(\max_g (\log(N)/N s_g^{2+2k})^{1/2} + \max_g s_g^{q-k+1})$. Thus, using a common bandwidth s that satisfies $s \sim (\log(N)/N)^{1/2(q+2)}$ ensures the optimal convergence rate. Remark 4 discusses a data-driven way of selecting such a bandwidth.

The following theorem shows that $\hat{\mu}_k(x)$ is asymptotically normally distributed.

Theorem 4.2. *Let k and q be two positive integers, $q \geq k+1$. Let assumptions 2.1-2.3, 4.1, and C.1 hold. Let assumptions D.1-D.4 in the appendix hold with $\tau \geq q+2$. For some $\delta > 0$ let $\sup_x \mathbb{E}\left[u_{it}^{2k+\delta} | X_{it} = x\right]$, $t = 1,2$. Let $s \to 0$, $\log(N)/N s^{2+2k} \to 0$, $N s^{2q+4} \to 0$, $s^2 \log(N) \to 0$. There exists a positive $V_k(x)$ (characterized in the proof) such that*

$$\sqrt{N s^{2+2k}} \left(\hat{\mu}_k(x) - \mu_k(x)\right) \Rightarrow N(0, V_k(x)). \tag{24}$$

If $x_1 \neq x_2$, then $\sqrt{N s^{2+2k}} \left(\hat{\mu}_k(x_1) - \mu_k(x_1)\right)$ and $\sqrt{N s^{2+2k}} \left(\hat{\mu}_k(x_2) - \mu_k(x_2)\right)$ are asymptotically independent.

Theorem 4.2 may be used to conduct inference on $\mu_k(x)$. In particular, nonparametric bootstrap offers an attractive option. It requires recomputing $\hat{\mu}_k(x)$ in each bootstrap sample by rerunning algorithm 1. Both pointwise and uniform confidence bands may be constructed. Alternatively, a plug-in estimator for $V_k(x)$ may be used. However, we recommend against this approach due to its complexity and poor performance; see the remark after the proof of theorem 4.2.

5 Asymptotic Properties of Distribution Estimators

We now turn to the asymptotic properties of our distribution estimators (18) and (22). We study two approaches which reflect the interpretations of the estimators discussed in section 3.2. In the first case, the true distribution $F_0(v|x)$ is a finite mixture with nonparametrically modeled mixture probabilities. In this semi-nonparametric case, section 5.1 establishes a distribution identification result that is stronger than theorem 2.2 and obtains convergence rates of the estimators. In the second case, $F_0(v|x)$ belongs to a nonparametric class of functions, and estimators (18) and (22) may be interpreted as penalized sieve estimators. Sections 5.2-5.4 establish consistency of the estimators in this nonparametric case.

5.1 Properties in the Semi-nonparametric Case

We first consider the situation in which $F_0(v|x)$ is a finite mixture in v for all x in the interval $I = [x_{lb}, x_{ub}]$. Formally, we impose the following assumption:

Assumption 5.1. *(i) There exists a finite integer $\bar{p}_v$, real numbers $v_{1,\bar{p}_v} < v_{2,\bar{p}_v} < \ldots v_{\bar{p}_v,\bar{p}_v}$, and functions $\rho_j(x) : I \to [0,1]$, $j = 1, \ldots, \bar{p}_v$ such that $F_0(v|x) = \sum_{j=1}^{\bar{p}_v} \rho_{0,j}(x)\Psi(v - v_{j,\bar{p}_v})$, $\rho_{0,j}(x) \geq 0$, and $\sum_{j=1}^{\bar{p}_v} \rho_{0,j}(x) = 1$ for all $x \in I$. (ii) $\rho_{0,j}(x)$ is twice continuously differentiable in x for $x \in I$. (iii) $\Psi(v)$ is continuously differentiable in v.*

Under assumption 5.1 the true distribution $F_0(v|x)$ is a $\bar{p}_v$-component mixture with centers $v_{j,\bar{p}_v}$ and mixing probabilities $\rho_{0,j}(x)$. We impose no functional form restrictions on the mixing probabilities. However, they are required to be twice differentiable in x. For simplicity, the centers $v_{j,\bar{p}_v}$ are assumed to be known. Such an assumption may be relaxed at the price of using more moments in identification and estimation. We assume that the centers $v_{j,\bar{p}_v}$ are independent of $x \in I$. This assumption is not restrictive, as the interval I may be taken to be arbitrarily short or to be a singleton.

A finite collection of moments is sufficient to identify $F_0(v|x)$ under assumption 5.1. The distribution $F_0(v|x)$ is determined by the $\bar{p}_v$-vector of mixing probabilities $\rho_{0,j}(x)$. As the following theorem shows, these probabilities are identified for all $x \in I$ using the moments of orders $1, 2, \ldots, \bar{p}_v - 1$. In turn, these moments are identified if $\sup_x \mathbb{E}[|u_{it}|^{\bar{p}_v - 1}|X_{it} = x] < \infty$, as theorem 2.2 shows. This distribution identification result is stronger than theorem 2.3, which uses an infinite number of moments to identify the distribution.

Theorem 5.1. *Let assumptions 2.1-2.3, 5.1, and C.1 hold. Let $\sup_x \mathbb{E}[|u_{it}|^{\bar{p}_v - 1}|X_{it} = x] < \infty$ for $t = 1, 2$. Then the true mixing probabilities $(\rho_{0,\bar{p}_v}(x), \ldots, \rho_{0,\bar{p}_v}(x))$ are identified.*

A practical consequence of theorem 5.1 is that the first $\bar{p}_v - 1$ moments are sufficient to estimate the distribution. We may take K equal to $\bar{p}_v - 1$ in the objective functions $\tilde{Q}_N(\boldsymbol{\gamma}|x_0)$ and $\hat{Q}_N(\boldsymbol{\gamma})$ of eqs. (17) and (21). $\tilde{Q}_N(\boldsymbol{\gamma}|x_0)$ and $\hat{Q}_N(\boldsymbol{\gamma})$ are then strictly convex on the constraint set. Regularization is superfluous, and λ_N and λ_N^I can be set to 0, as in section 3.

The following theorem establishes consistency and convergence rates of estimators (18) and (22) under assumption 5.1. Let $\tilde{\mu}_k(x)$ be some consistent estimators of $\mu_k(x)$ for $k \leq \bar{p}_v - 1$. Let $d_{TV}(F, G)$ be the total variation distance between two distributions F and G.

Theorem 5.2. *Let assumptions 2.1-2.3, 5.1, and C.1 hold. let* $K = \bar{p}_v - 1$

(1) Let the moment estimators $\tilde{\mu}(x)$ *satisfy* $|\tilde{\mu}_k(x_0) - \mu_k(x_0)| = O_{a.s.}(\delta_{k,N}^{(x_0)})$*, where* $\delta_{k,N}^{(x_0)}$ *is a deterministic sequence that satisfies* $\delta_{k,N}^{(x_0)} = o(1)$*. Let* $\lambda_N = 0$*. Then the estimator* $\tilde{F}_N(\cdot|x_0)$ *of eq. (18) satisfies*

$$d_{TV}(\tilde{F}_N(\cdot|x_0), F_0(\cdot|x_0)) = O_{a.s.}\left(\max_{k=1,\ldots,\bar{p}_v-1} (\delta_{k,N}^{(x_0)})^{1/2}\right).$$

(2) Let the moment estimators $\tilde{\mu}(x)$ *satisfy* $\sup_{x\in I}|\tilde{\mu}_k(x) - \mu_k(x)| = O_{a.s.}(\delta_{k,N})$*, where* $\delta_{k,N}$ *is a deterministic sequence that satisfies* $\delta_{k,N} = o(1)$*. Let* $\lambda_N^I = 0$ *and* $p_x = p_x(N)$ *be a non-decreasing sequence such that* $p_x \to \infty$*. Then the estimator* $\hat{F}_N(\cdot|\cdot)$ *of eq. (22) satisfies*

$$\int d_{TV}(\hat{F}_N(\cdot|x) - F_0(\cdot|x))\pi(dx) = O_{a.s.}\left(\max\left\{p_x^{-1}, \max_{k=1,\ldots,\bar{p}_v-1} \delta_{k,N}^{1/2}\right\}\right).$$

The convergence rates of $\tilde{F}_N(\cdot|x_0)$ and $\hat{F}_N(\cdot|\cdot)$ are driven by the convergence rates of moment estimators $\{\tilde{\mu}(x)\}_{k=1}^{\bar{p}_v-1}$ and (for $\hat{F}_N(\cdot|\cdot)$) the order p_x of the Bernstein approximation.

The moment estimators influence the convergence rates of $\tilde{F}_N(\cdot|x_0)$ and $\hat{F}_N(\cdot|\cdot)$ through the slowest-converging moment. Typically, this will be the highest-order moment used. A special case arises if we use the moment estimators of algorithm 1. Suppose that a common order q is used for all the local polynomial estimators, and the optimal bandwidth s is used. Then by theorem 4.1 it holds that

$$\int d_{TV}(\hat{F}_N(\cdot|x) - F_0(\cdot|x))\pi(dx) = O_{a.s.}\left(\max\left\{p_x^{-1}, \left(\frac{\log(N)}{N}\right)^{(q-\bar{p}_v+2)/(4(q+2))}\right\}\right).$$

The convergence rate of the interval estimator $\hat{F}_N$ is also driven by p_x, the order of the Bernstein polynomials used to approximate the mixing probabilities $\{\rho_{0,j}(x)\}_{j=1}^{\bar{p}_v}$. p_x does not enter the "stochastic" component of the convergence rate. To see why, note that the optimization problem (23) may be intuitively viewed as a two-layer minimal distance procedure. First, a $\bar{p}_v$-vector of mixing probabilities is estimated to best approximate moment

estimates. Second, the coefficients of Bernstein polynomials are picked to approximate the mixing probabilities from the first step. The distance between $\hat{F}_N$ and F_0 is entirely determined by the distance in the corresponding mixing probabilities, a vector of fixed dimension. The behavior and the dimension of individual coefficients in $\hat{\boldsymbol{\gamma}}$ only matters inasmuch as the estimated mixing probabilities $[\sum_{l=0}^{p_x} \hat{\gamma}_{j,l} b_{l,p_x}(x)]$ approximate the true $\rho_{0,j}(x)$.

In light of the above, the only limits on the value of p_x are practical, as it is theoretically optimal to set $p_x = \infty$. As p_x grows, the problem becomes higher-dimensional and more challenging to solve. Further, using large values of p_x involves using polynomials of high orders, which may lead to numerical instability.

Remark 9. Theorem 5.2 is also a convergence result for densities, as both the estimators and F_0 are differentiable, and convergence in total variation is tightly linked to L^1 convergence of corresponding densities as $2d_{TV}(F,G) = \int |F'(v) - G'(v)|dv$.

5.2 Approximating a Nonparametric $F_0(v|x)$

We now adopt a nonparametric perspective and suppose that assumption 5.1 may fail to hold. In this case the functions $\Lambda_p(v|\boldsymbol{\gamma})$ and $\Lambda_{p_v,p_x}(v|x,\boldsymbol{\gamma})$ of eqs. (16) and (20) provide flexible mixture approximations to the unknown true distribution $F_0(v|x)$.

In order to formalize the approximating properties of $\Lambda_p(v|\boldsymbol{\gamma})$ and $\Lambda_{p_v,p_x}(v|x,\boldsymbol{\gamma})$ and the asymptotic properties of the estimators of section 3.2, we first impose a smoothness assumption on $F_0(v|x)$ and define suitable spaces $\mathcal{F}$ and $\mathcal{F}^I$ of distribution functions. The following assumption replaces assumption 5.1, and is considerably weaker than the latter.

Assumption 5.2. *(i) $\mathcal{F}$ is the space of distribution functions F with bounded support; each $F \in \mathcal{F}$ continuously differentiable with a bounded density. (ii) $\mathcal{F}^I$ is the space of bivariate functions $F(v|x)$ on $\mathbb{R} \times I$ such that (i) for each $x \in I$ $F(v|x)$ a cumulative distribution function; (ii) $F(v|x)$ is continuously differentiable in v for each $x \in I$; (iii) for each v the function $F(v|x)$ is twice continuously differentiable in x; (iv) $\sup_{v,x}|\partial_v F(v|x)| < \infty$; (v) the support of $F(v|x)$ is bounded uniformly in x. (iii) the true distribution $F_0(v|x)$ lies in $\mathcal{F}^I$ (which implies that $F_0(v|x) \in \mathcal{F}$ for any $x \in I$).*

We first consider the problem of approximating $F_0(v|x_0)$ at a given point x_0. The functions $\Lambda_p(v|\boldsymbol{\gamma})$ of eq. (16) can approximate any $F \in \mathcal{F}$ arbitrarily well if the corresponding parameters p, Ψ and the mixture centers are specified as we presently describe. The reference distribution function Ψ is assumed to satisfy the following assumption:

Assumption 5.3. *$\int \Psi(dt) = 1$, $\Psi(t) \geq 0$ for all t, Ψ supported on $[-1,1]$, Ψ is differentiable with a symmetric density $\psi(t)$.*

In order to choose centers, let C_M be a positive constant. Partition the interval $C_M\sqrt{\log(\log(p))}$ into p equal-length closed subintervals. Let $v_{j,p}$ be the center of the jth interval, numbered from left to right. The constant C_M may be chosen according to the method in remark 6 to ensure that the majority of the potential support of $F_0(v|x)$ is covered by mixture components.

Let $\mathcal{L}_p$ be the collection of all functions of the form

$$\mathcal{L}_p = \left\{ \Lambda_p(v|\boldsymbol{\gamma}) = \sum_{i=1}^{p} \gamma_i \Psi\left(\sigma_p^{-1}(v - v_{j,p})\right), \quad \sum_{i=1}^{p} \gamma_i = 1, \quad \gamma_i \geq 0 \right\}, \tag{25}$$

where

$$\sigma_p = \frac{1}{p\sqrt{\log(\log(p))}e^{\sqrt{\log(\log(p))}}}. \tag{26}$$

$\mathcal{L}_p$ is the pth (pointwise) sieve space, and any function $F \in \mathcal{F}$ may be approximated arbitrarily well as $p \to \infty$, as we show in the proof appendix. $\mathcal{L}_p$ is a variation of the mixture of experts sieve (Zeevi and Meir, 1997; Li and Barron, 1999; Norets, 2010).

A similar approach can be taken to approximate $F_0(v|x)$ as x varies in I. Let p_v be the number of mixture components. Let the mixture centers v_{j,p_v} be selected as above with p_v in place of p. Let p_x be the order of the Bernstein polynomials approximating the mixing probabilities. Define the (p_v, p_x)th (interval) sieve space $\mathcal{L}^I_{p_v,p_x}$ as

$$\mathcal{L}^I_{p_v,p_x} = \left\{ \Lambda_{p_v,p_x}(v|x,\boldsymbol{\gamma}) : \gamma_{j,l} \geq 0, \sum_{j=1}^{p_v} \gamma_{j,l} = 1, \ l = 0, \ldots, p_x \right\}, \tag{27}$$

where for σ_p of eq. (26) and Ψ of assumption 5.3 we define

$$\Lambda_{p_v,p_x}(v|x,\boldsymbol{\gamma}) = \sum_{j=1}^{p_v} \left[\sum_{l=0}^{p_x} \gamma_{j,l} b_{l,p_x}(x) \right] \Psi\left(\sigma_p^{-1}(v - v_{j,p_v})\right),$$

$$b_{l,p_x}(x) = \binom{p_x}{l} \left(\frac{x - x_{lb}}{x_{ub} - x_{lb}}\right)^l \left(\frac{x_{ub} - x}{x_{ub} - x_{lb}}\right)^{p_x - l}.$$

$\mathcal{L}^I_{p_v,p_x}$ is a hybrid polynomial-mixture sieve that can approximate $F_0(v|x)$ arbitrary well. The mixture weights are modeled using Bernstein polynomials on I, which yields transparent conditions on the parameters $\boldsymbol{\gamma}$ that ensure that every member of $\mathcal{L}^I_{p_v,p_x}$ is a valid distribution function for each value of the conditioning argument.

5.3 Estimation as Approximate Moment Metric Minimization

We now provide the analytical background for the choice of the objective functions (17) and (21). Functions (17) and (21) approximately measure the distance between the true distribution and the sieve approximant in terms of their moments, as noted in section 3.2. In this section, we introduce such moment distances, establish their properties, and connect them to the estimation problem of section 3.2.

Define the moment 2-metric $d_{2,\mu}$ between two conditional distribution functions $F(\cdot|x_0)$ and $G(\cdot|x_0)$ at a fixed point x_0

$$d_{2,\mu}(F(\cdot|x_0), G(\cdot|x_0)) = \left[\sum_{k=1}^{\infty} \frac{1}{k!} \left[\int t^k F(dv|x_0) - \int t^k G(dv|x_0)\right]^2\right]^{1/2}.$$

In order to measure distances between families of conditional densities $F(\cdot|x)$ and $G(\cdot|x)$ as x varies in the interval I, we define the integrated moment metric. Let π be a finite measure on I such that the Lebesgue measure on I is absolutely continuous with respect to π and define the integrated metric

$$d^{\pi}_{2,\mu}(F, G) = \int d_{2,\mu}(F(\cdot|x), G(\cdot|x))\pi(dx).$$

The moment metrics $d_{2,\mu}$ and $d^{\pi}_{2,\mu}$ connect convergence of moments to weak convergence of distributions, as the following lemma shows.

Lemma 5.3. *Let $\mathcal{F}$ and $\mathcal{F}^I$ be defined as in assumption 5.2. Then*

(1) $d_{2,\mu}(\cdot,\cdot)$ is a metric on $\mathcal{F}$. If $F_n, F \in \mathcal{F}$, then $d_{2,\mu}(F_n, F) \to 0$ implies that

$$\sup_v |F_n(v) - F(v)| \to 0.$$

(2) $d^{\pi}_{2,\mu}(\cdot,\cdot)$ is a metric on $\mathcal{F}^I$. If $F_n, F \in \mathcal{F}^I$, then $d^{\pi}_{2,\mu}(F_n, F) \to 0$ implies that

$$\int \sup_v |F_n(v|x) - F(v|x)|\pi(dx) \to 0.$$

Lemma 5.3 is the key piece of our estimation strategy. Consider first the problem of estimating $F_0(v|x_0)$ at a fixed point x_0. Since $d_{2,\mu}$ is a metric on $\mathcal{F}$, $F_0(v|x_0)$ is the unique solution of $\min_{F\in\mathcal{F}} d_{2,\mu}(F_0(\cdot|x_0), F(\cdot))$. This minimization problem is infeasible, since we only have access to noisy estimates of $\mu_k(x_0)$. Moreover, the estimates of higher-order moments are dominated by noise. We then replace $d_{2,\mu}(F_0(\cdot|x_0), F(\cdot))$ by a finite-sample version that uses a finite and growing number of estimated moments $\tilde{\mu}_k(x_0)$.

For $F \in \mathcal{F}$ we define

$$\tilde{Q}(F|x_0) = \sum_{k=1}^{K-1} \frac{1}{k!} \left[\tilde{\mu}_k(x_0) - \int t^k F(dv)\right]^2, \tag{28}$$

where $\tilde{\mu}_k(x_0)$ is some estimator of $\mu_k(x_0)$. The definition in eq. (28) agrees with that of eq. (17) for all $F \in \mathcal{L}_p$. Additionally, we define the population counterpart of $\tilde{Q}_N(F|x_0)$ as

$$Q(F|x_0) = \sum_{k=1}^{\infty} \frac{1}{k!} \left[\mu_k(x_0) - \int t^k F(dv) \right]^2 \equiv (d_{2,\mu}(F_0(\cdot|x_0), F(\cdot)))^2. \tag{29}$$

Minimizing $\tilde{Q}_N(F|x_0)$ over $\mathcal{L}_p$ approximates minimizing the distance $d_{2,\mu}$ to the target cdf $F_0(\cdot|x_0)$ over $\mathcal{F}$. The approximation becomes exact as the moment estimators converge to the true moments, the number of moments used K grows, and the sieve space size p increases.

To estimate $F_0(v|x)$ as x varies in I, we proceed similarly. We define for $F \in \mathcal{F}^I$

$$\hat{Q}_N(F) = \int_I \sum_{k=1}^{K-1} \frac{1}{k!} \left[\tilde{\mu}_k(x) - \int t^k F(dv|x) \right]^2 \pi(dx) \tag{30}$$

$$Q(F) = \int_I \sum_{k=1}^{\infty} \frac{1}{k!} \left[\mu_k(x) - \int t^k F(dv|x) \right]^2 \pi(dx) \tag{31}$$

where $\tilde{\mu}_k(x)$ is an estimator of $\mu_k(x)$, and π is a finite measure on I such that the Lebesgue measure is absolutely continuous with respect to π.

As above, the problem of minimizing $\hat{Q}_N(F)$ is approximately equivalent to minimizing the distance $d^{\pi}_{2,\mu}(F(\cdot|\cdot), F_0(\cdot|\cdot))$, though the argument is somewhat more complex than in the case of $F_0(v|x_0)$. First, minimizing $\hat{Q}_N$ over $\mathcal{L}^I_{p_v,p_x}$ approximately corresponds to minimizing Q over $\mathcal{F}^I$. Second, $\min_{\mathcal{F}^I} Q(F)$ is equal to 0, and it is uniquely attained at $F = F_0(v|x)$.[5] Last, by Jensen's inequality the function Q upper bounds the integral metric as $[d^{\pi}_{2,\mu}(F, F_0)]^2 \leq \int d^2_{2,\mu}(F(\cdot|x), F_0(\cdot|x))\pi(dx) \equiv Q(F)$.

5.4 Consistency of Nonparametric Distribution Estimators

Before stating a consistency results for estimators (18) and (22), we impose a technical condition on the moment estimators $\tilde{\mu}_k(x)$. Let $\{\hat{\mu}_k(x)\}_{k=1}^{\infty}$ be some sequence of estimators $\mu_k(x)$ that are consistent in the sense that $\sup_{x \in I} |\hat{\mu}_k(x) - \mu_k(x)| = O_{a.s.}(\delta_{k,N})$ for some deterministic $\delta_{k,N} = o(1)$. Define the moment estimators $\tilde{\mu}_k(x) = \max\{\min\{\hat{\mu}_k(x), C^k_{\mu}(k!)^{1/4}\}, -C^k_{\mu}(k!)^{1/4}\}$ where $C_{\mu} > 0$ is constant such that $\tilde{\mu}_k(x)$ remains consistent in the sense that

$$\sup_{x \in I} |\tilde{\mu}_k(x) - \mu_k(x)| = O_{a.s.}(\delta_{k,N}). \tag{32}$$

[5] To see this, let $Q(F) = 0$. Then $d_{2,\mu}(F(\cdot|x), F_0(\cdot|x)) = 0$ for π-almost all x, and thus for Lebesgue-almost all x (by the absolute continuity requirement). $d_{2,\mu}(F(\cdot|x), F_0(\cdot|x))$ is a continuous function of x by definition of $\mathcal{F}^I$. We conclude that $d_{2,\mu}(F(\cdot|x), F_0(\cdot|x)) = 0$ for all $x \in I$. By lemma 5.3 we conclude that $F = F_0$.

Two remarks are in order. First, a suitable C_μ always exists under assumption 5.2. The support of $F_0(v|x)$ is bounded uniformly in x, and thus the moments $\mu_k(x)$ may grow at most exponentially in k.[6] Second, the trimmed nature of $\tilde{\mu}_k(x)$ is a purely technical assumption. C_μ may be taken arbitrarily large, ensuring that $\tilde{\mu}_k(x) = \hat{\mu}_k(x)$ for any finite collection of k.

We now state a consistency result for estimators (18) and (22).

Theorem 5.4. *Let assumptions 2.1-2.3, 4.1, 5.2-5.3, and C.1 hold. Let the moment estimators $\tilde{\mu}_k(x)$ satisfy (32). Let $K{=}K(N)$ be a non-decreasing sequence such that $K \to \infty$.*

(1) Let $p{=}p(N)$ satisfy $p \to \infty$, $\log(p) = o(\sqrt{K})$ and $\log(p) = o(\delta_{k,N}^{1/k})$ for all $k = 1, 2, \ldots$ for $\delta_{k,N}$ of (32). Let $\lambda_N \to 0$. Then the estimator $\tilde{F}_N(\cdot|x_0)$ of eq. (18) satisfies

$$\sup_{v \in \mathbb{R}} \left| \tilde{F}_N(v|x_0) - F_0(v|x_0) \right| \xrightarrow{a.s.} 0.$$

(2) Let $p_v{=}p_v(N)$ satisfy $p_v \to \infty$, $\log(p_v) = o(\sqrt{K})$ and $\log(p_v) = o(\delta_{k,N}^{1/k})$ for all $k = 1, 2, \ldots$ for $\delta_{k,N}$ of (32). Let $p_x = p_x(N)$ satisfy $p_x \to \infty$. Let $\lambda_N^I = o(p_x^{-1})$. Then the estimator $\hat{F}_N(\cdot|\cdot)$ of eq. (22) satisfies

$$\int_I \sup_{\mathbb{R}} \left| \hat{F}_N(v|x) - F_0(v|x) \right| \pi(dx) \xrightarrow{a.s.} 0.$$

Theorem 5.4 characterizes the nature of consistency of estimators (18) and (22). The estimated cdfs converge uniformly to the true cdf of interest, in line with lemma 5.3. In the case of estimation at one point x_0, $\tilde{F}_N(\cdot|x_0)$ simply converges uniformly to $F_0(\cdot|x_0)$. In the interval case, the estimated family $\hat{F}_N(\cdot|x)$ (indexed by x) converges in a L^1-L^∞ hybrid mode.

Conditions of theorem 5.4 can be broadly split into three groups standard in penalized sieve estimation literature (e.g. Chen and Pouzo (2012)). The first group deals with the complexity of sieve spaces. As $p \to \infty$ or $p_v, p_x \to \infty$, the sieve spaces are able to approximate $F_0(v|x)$ arbitrary well. The sample optimization problem can return a solution that is arbitrarily close to $F_0(v|x)$. The second group concerns the conditions on the penalty imposed on the sample problems (19) and (23). In the case of estimation at one point, the nature of the sieve spaces limits the magnitude of the penalty as $\sum_{j=1}^{p} \gamma_j^2 \leq 1$ uniformly; in this case it is sufficient for $\lambda_N \to 0$ to avoid regularization bias in the limit. In the interval case, a rate condition on λ_N^I is necessary to ensure that the penalty is asymptotically negligible. The third and most complex group deals with convergence of the sample functions $\tilde{Q}_N$ and $\hat{Q}_N$ to the corresponding population functions Q. This group links together the convergence rates $\delta_{k,N}$ of the moments, K, and the mixture sizes p and p_v. As K increases without bound, the truncation error in $\hat{Q}$ incurred by using only a finite number of moment decreases. At

[6] In particular, let $\text{supp}(F_0(v|x)) = [c_\mu(x), C_\mu(x)]$. Then $C_\mu := \sup_{x \in I} \max\{|c_\mu(x)|, |C_\mu(x)|\}$ satisfies (32) (this value may be estimated, see remark 7).

the same time, the mixtures cannot expand too quickly relative to both the number K of moments used and the convergence rate of the moments. The relevant conditions are stated in a high-level form, and they can be specialized given a particular form of $\delta_{k,N}$. For example, if the moment estimators of section 3.1 are used, p and p_v may be taken polynomial in N, and K may be taken polynomial or logarithmic in N. In contrast, there is no restriction on the order p_x of Bernstein polynomials, as in the semi-nonparametric case (see theorem 5.2).

6 Monte Carlo

In this section, we assess the finite sample performance of our moment and distribution estimators with a Monte Carlo study. The data is generated as follows:

$$Y_{it} = m(X_{it}, \alpha_i) + u_{it}, \quad i = 1, \ldots, N, \quad T = 1, 2, \tag{33}$$
$$m(x, \alpha_i) = \alpha_i^{(1)} \left(0.75 x^{(\alpha_i^{(2)}-1)/\alpha_i^{(2)}} + 0.25(24-x)^{(\alpha_i^{(2)}-1)/\alpha_i^{(2)}}\right)^{\alpha_i^{(2)}/(\alpha_i^{(2)}-1)},$$

where X_{i1} is uniformly distributed on $[0, 6]$; $X_{i2} = \min\{\max\{X_{i1} + \xi_{it}, 0\}, 6\}$, ξ_{it} is Beta(2, 2)-distributed on $[-\sigma_x, \sigma_x]$ for $\sigma_x = 1$ (picked to approximately reproduce the empirical share of near-stayers). The shock u_{it} satisfies $u_{it}|(X_{i1}, X_{i2}) \sim N(0, \sigma_u^2)$; σ_u^2 chosen to match the average (across x) variance of marginal effects at x. The time-invariant vector $(\alpha_i^{(1)}, \alpha_i^{(2)})$ is drawn is follows; $\alpha_i^{(1)}|(X_{i1}, X_{i2}) = (x_1, x_2)$ is Beta(2, 2)-distributed on $[0.5, 1.5]$, $\alpha_i^{(2)}|(X_{i1}, X_{i2}) = (x_1, x_2)$ is Beta$(3 - x_1/3, 1 + x_1/3)$-distributed on the interval $[0.5, 1.5]$; $\alpha_i^{(1)}$ and $\alpha_i^{(2)}$ independent conditional on (X_{i1}, X_{i2}). The number of cross-sectional units N is taken as 7500 (approximately 50% of the cross-sectional dimension of the dataset of section 7) and $T = 2$. Cross-sectional units are iid. We draw 1000 datasets.

Specification (33) may be interpreted as a constant elasticity of substitution function with unit-specific scale $\alpha_i^{(1)}$ and elasticity of substitution $\alpha_i^{(2)}$. The support of $\alpha_i^{(2)}$ includes 1; this value corresponds to a Cobb-Douglas form in eq. (33). $\alpha_i^{(2)}$ is more likely to lie below 1 for lower values of x; m is closer to a Leontief function for such draws of $\alpha_i^{(2)}$. $\alpha_i^{(2)}$ is more likely to lie above 1 for higher values of x; m is closer to a linear function for such values of $\alpha_i^{(2)}$.

We estimate the moments $\mu_k(x)$, the distribution of marginal effects using the estimators of section 3. To estimate $\mu_k(x)$, we estimate the first stage conditional moments and their derivatives using an local polynomial estimator of order $k + 1$. The common smoothing bandwidth s is taken to be 1.2. It is selected by applying the multivariate generalized C_p criterion of Charnigo and Srinivasan (2015) for first derivatives (see remark 4). We use a 2-dimensional product Epanechnikov kernel. We report estimates for the first 4 moments in this section. Negative estimates of even-order moments are replaced by zeros. To estimate

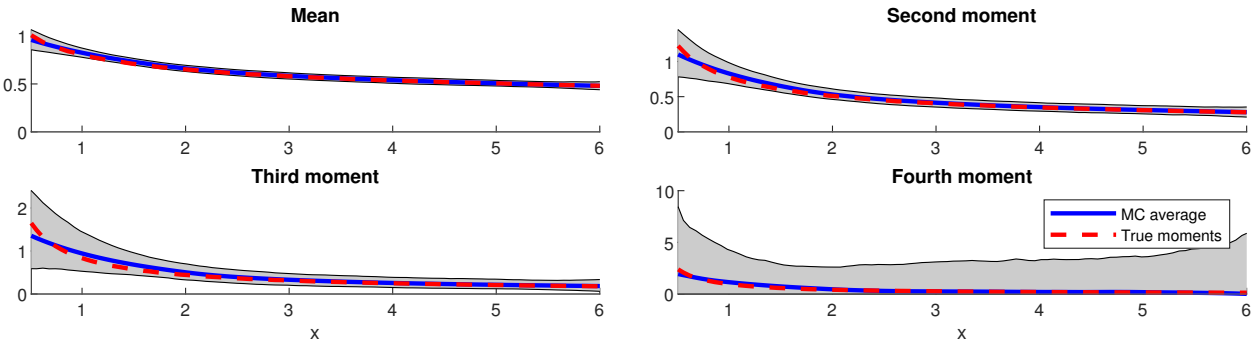

Figure 1: Simulation results for estimators of moments $\mu_k(x)$ of marginal effects, $k = 1, 2, 3, 4$. Solid line – average of estimates across Monte Carlo sample; dashed line – true population moments; shaded area – 95% Monte Carlo bands

the distribution, we fit mixtures of 3-8 components using 3-8 moments (including the zeroth moment). The locations of the mixture are located symmetrically around the estimated mean using Chebyshev's inequality (see remark 6). The reference cdf Ψ is a unit-variance Epanechnikov cdf. We only report the results for the pointwise estimator in this section, as the estimates are virtually indistinguishable from those of the interval estimator. Some further results may be found in the Supplementary Appendix.

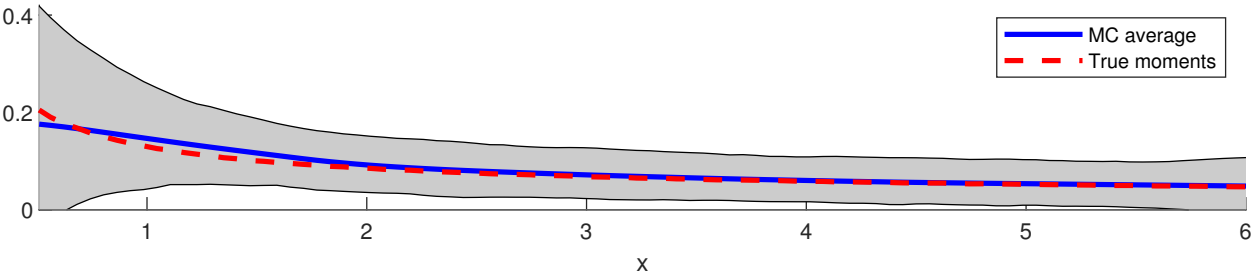

Figure 2: Simulation results for estimator of variance of marginal effects. Solid line – average of estimates across Monte Carlo sample; dashed line – true population moments; shaded area – 95% Monte Carlo bands

Simulation results for moments are graphically represented on figs. 1-2. We report the results for the the raw moments $\mu_k(x)$ and the variance $(\mu_2(x) - \mu_1^2(x))$ of marginal effects. For each moment, we plot the average of the Monte Carlo samples and 95% Monte Carlo bands, as well as the target population moments.

The moment estimators for $\mu_k(x)$ generally perform well in terms of bias and variance. The bias is generally low and the estimators track the shape of $\mu_k(x)$ well. Estimator variability is generally limited for the moments under consideration, though the Monte Carlo bands become larger relative to the estimand for $\mu_4(x)$, in line with theorem 4.1. Estimation is more challenging closer to the boundaries of the support of X.

We now turn to the results for the distribution estimators. The top panel of fig. 3 shows the average five component mixture density fitted with the first five moments, starting from

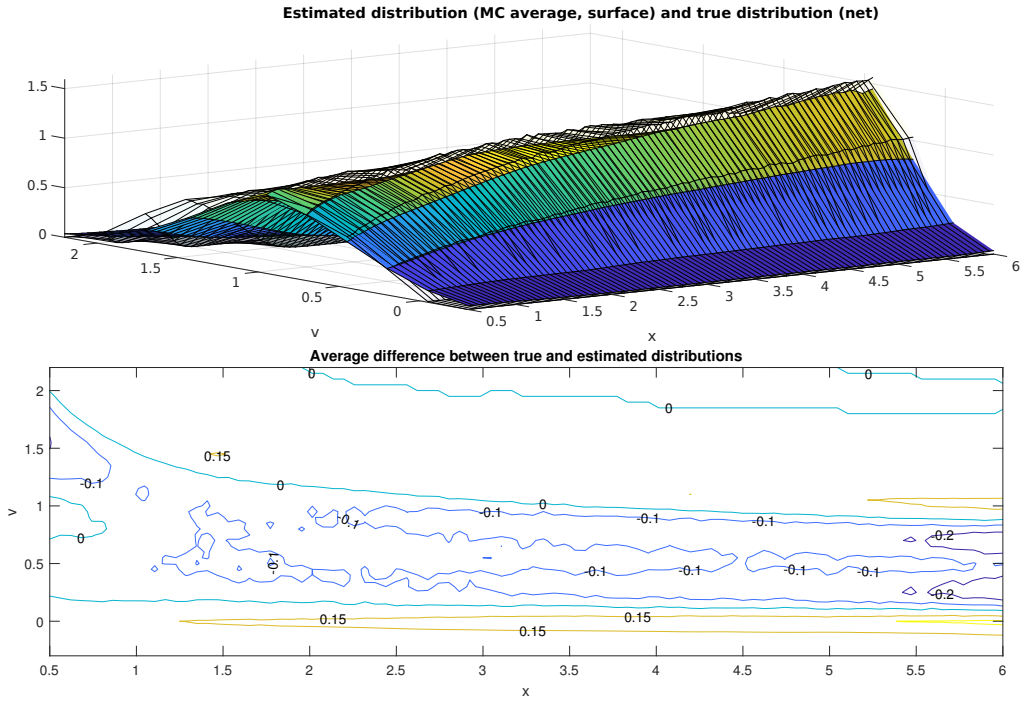

Figure 3: Simulation results for estimator of density $f(v|x)$ of marginal effects conditional on $\{X_{i1} = X_{i2} = x\}$. Top panel: solid surface – average of estimates across Monte Carlo samples, five components mixture fitted using moments 0-4; transparent net – true population distribution. Bottom panel: contours of difference between average distribution estimate and true distribution

the zeroth moment ($K = p = 5$ in eq. (17)), averaged across Monte Carlo samples. The density is obtained by differentiating the cdf estimators. The shape is contrasted with the true distribution, reported on fig. 3 as a transparent net. For clarity, we also plot the contours of the difference between the two on the bottom panel fig. 3. Contours of the true distribution and of different estimator specifications are depicted on fig. 5.

Overall, the bias-variance properties of the distribution estimators reflect those of $\hat{\mu}_k(x)$. Consider bias first. The estimates generally track the shape of the distribution well, including tail behavior and the asymmetry for lower values of x. To measure variance, we report 95% Monte Carlo bands for the distribution on fig. 4, splitting the plot at $x = 3$ to reveal the lower band. The peaks of the estimates are somewhat more variable in the regions where the variance of marginal effects is smaller, such as for x around 6 (compare fig. 2). Near-zero estimates of variance lead to very tightly concentrated distribution estimates.

The distribution estimates are fairly insensitive to the number of components and the number of moments used in estimation. On fig. 5 we report contours for different combinations of the number of components p and the number of moments K, as well as the contours of the true density. Overall, all 3-6 component mixtures using 3-6 moments approximate the distribution well. Increasing the number of moments beyond that allows the model to better

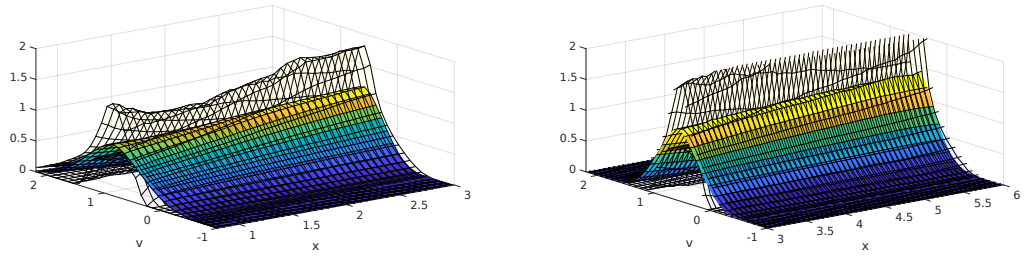

Figure 4: Simulation results for estimator of distribution of marginal effects for stayers at x. Solid surface – average of estimates across Monte Carlo samples, five components mixture fitted using moments 0-4. Transparent net – 95% Monte Carlo bands.

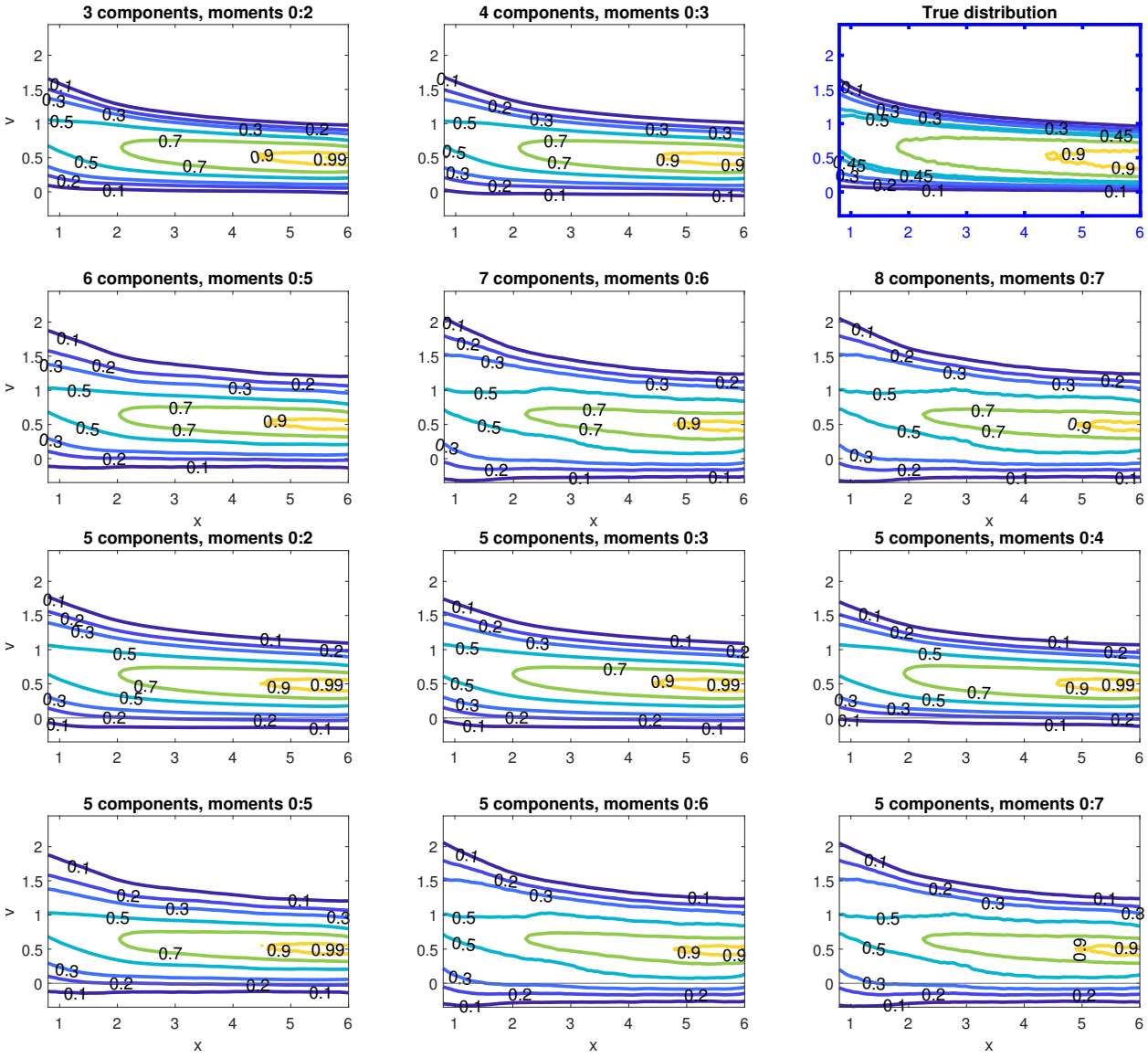

Figure 5: Sensitivity of distribution estimates to number of components (p) and number of moments (K): contour lines of averages of estimates across Monte Carlo samples. Top right panel – contours of true density. Values of density normalized to lie in the interval $[0, 1]$ for clarity

approximate features such as the peak of the true distribution, at the price of somewhat overestimating the left tail.

Finally, we note that the Supplementary Appendix contains the results for two further simulations: a simulation that recreates several key features of the empirical application of section 7, and specification (33) with $N = 15000$. The results of these additional simulation studies support the results presented in this section.

7 Empirical Application

In this section we apply our methodology to Engel curves for food at home. Specifically, we consider the first four moments and the distribution of the slopes of individual Engel curves while allowing for unrestricted variation in individual preferences. Demand analysis is a natural domain of application of our methods, as preferences over consumption are typically multidimensional and potentially related in a complex manner with determinants of demand (McFadden, 2005; Browning and Carro, 2007). The particular relevance of the Engel curve slope is that it captures information about the income effect. As Banks et al. (1997) note, fully estimating the distribution of the income effect at all points of the income distribution is key for predicting responses to tax reforms. Further, Engel curves for food at home may be used to establish purchasing-parity conversions (Almås, 2012).

Data and model The data on household expenditures on food at home and total expenditures is drawn from the 2011-2019 waves of the US Consumer Expenditure Survey (CES). The CES is a quarterly rotating panel dataset: each household is surveyed in up to 4 consecutive quarters before being replaced by a new household in the sample. To obtain a demographically homogeneous sample, we only retain the households formed by married or cohabitating couple with no children where the head of the household is between 20 and 65 years. This selection matches the demographics considered by Imbens and Newey (2009); Chen and Pouzo (2012) and Chernozhukov et al. (2015). The resulting sample contains 8132 individual households that participated in the survey at least twice. To account for price changes between 2011 and 2019, we deflate the expenditures to the level of the first quarter of 2011 by the consumer price index. We transform the dataset into a balanced panel of length $T = 2$ by treating each pair of consecutive quarters for every household as a new separate observation. The resulting dataset has 15231 cross-sectional observations. A formal description of the procedure is available in the Supplementary Appendix. Let i index the new resulting cross-section.

Let Y_{it} be the share of food at home in total expenditure in period t where $t = 1, 2$. We

assume that Y_{it} satisfies model (2) as

$$Y_{it} = m(X_{it}, \alpha_i) + u_{it}.$$

Here X_{it} is the deflated log total expenditure and α_i plays the role of time-invariant preferences. The function $m(X_{it}, \alpha_i)$ is the Engel curve. This model allows for a complex relationship between expenditures and preferences. Last, u_{it} reflects idiosyncratic shocks to consumption. We assume that $\mathbb{E}[u_{it}|X_{it}] = 0$, this assumption is found to hold for food at home by Chernozhukov et al. (2015) in their analysis of average marginal effects.

Summary statistics and the prevalence of (near-)stayers Some summary statistics of the data are presented on fig. 6. On the left panel, we plot a kernel estimate of the log expenditure in the first and the second periods; all kernel estimates use a normal kernel with the bandwidths selected by Silverman's rule of thumbs. The two distributions are virtually indistinguishable. Quarterly expenditure ranges from approximately \$400 to \$100000 2011 dollars (as log expenditure ranges from 6 to 11). The middle panel depicts a kernel estimate of the density of changes in log expenditure. There is a high density of households with small changes in expenditure. In other words, the population discussed in this paper — the near-stayers and the stayers — comprises a large part of the population represented in the data. This allows us to construct meaningful estimates for the objects of interest, as near-stayers serve as the basis of identification and estimation for the stayers. Last, on the right panel we plot a kernel estimate of the joint density of log expenditure and share of food at home in expenditure. Engel's law holds in our data – the estimates show a negative correlation between expenditure and the share going on food at home.

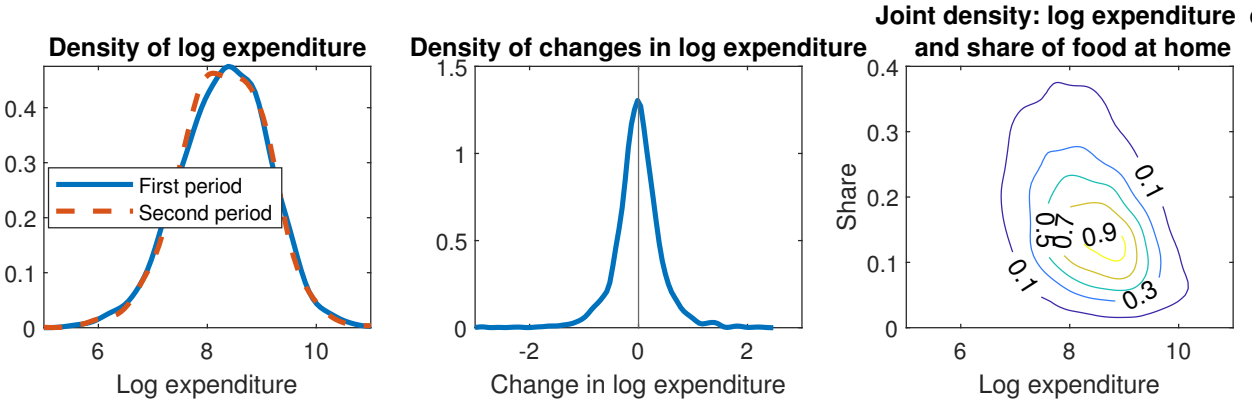

Figure 6: Summary of distributional features of log expenditure and share of food at home in expenditure. Left panel: kernel estimate of density of log expenditure in the first and second periods. Middle panel: kernel estimate of density of changes in log expenditure. Right panel: kernel estimate of joint density of log expenditure and shares. Note: values of joint density normalized to lie in the interval [0, 1] for clarity.

Methodology The moments $\mu_k(x)$ and the distribution of the Engel curve slope are estimated using the estimators of section 3 similarly to in section 6, to which we refer for a discussion. There are two differences relative to section 6. First, we report 95% pointwise confidence bands around the estimates using nonparametric bootstrap with 1000 bootstrap samples. Second, for purposes of inference, we undersmooth the moment estimators. Specifically, the C_p criterion of Charnigo and Srinivasan (2015) suggests taking the smoothing bandwidth s equal to 1.15. We reduce the value to 1 to ensure that the confidence intervals are centered at $\mu_k(x)$ by theorem 4.2.

Results: mean and standard deviation We first consider the mean and the standard deviation of the Engel curve slope. Our estimates are depicted on fig. 7. For interpretability, we report the x-axis in dollars. The estimated mean is negative for all expenditure values, indicating a downward-sloping average Engel curve. Overall, the results for the mean are consistent both with our estimates of the average Engel curve ($\nu_m(x)$, fig. 8) and the estimates of the (average) Engel curve by Blundell et al. (2007b).

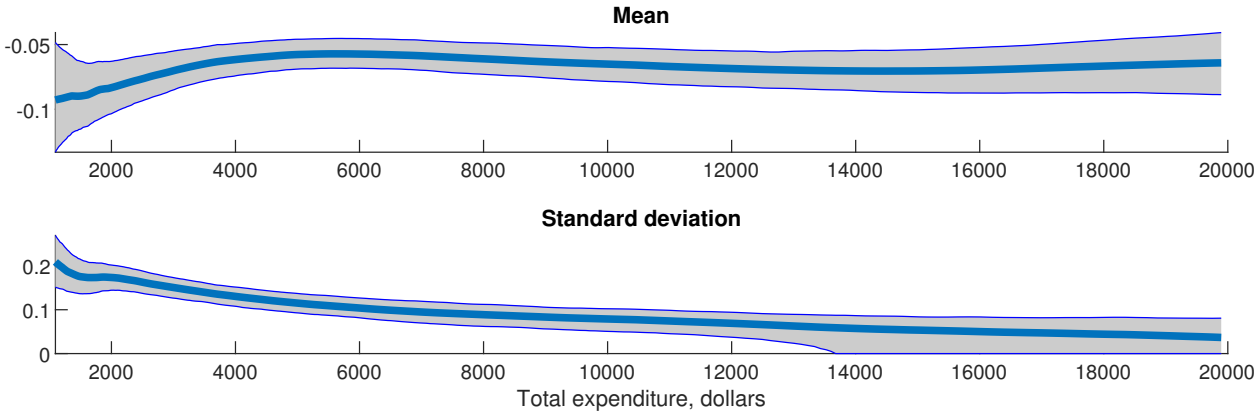

Figure 7: Mean and standard deviation of Engel curve slope ($\mu_1(x)$ and $[\mu_2(x) - \mu_1^2(x)]^{1/2}$) for food at home and 95% bootstrap pointwise confidence band. Moments estimated using log expenditure (log dollars); results reported in expenditure (dollars)

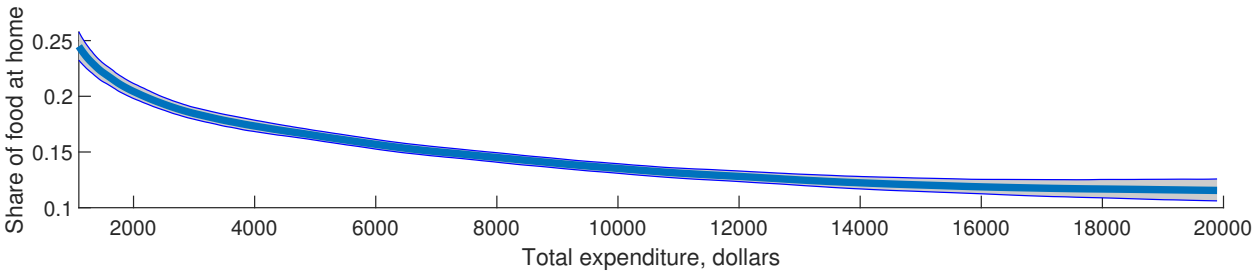

Figure 8: Average Engel curve ($\nu_m(x)$) for food at home and 95% bootstrap pointwise confidence band. Moments estimated using log expenditure (log dollars); results reported in expenditure (dollars)

Turning to the standard deviation, we reject slope homogeneity for expenditures below

$14000. Furthermore, in this region, the estimated standard deviation is similar in magnitude to the mean. This suggests that a fraction of the stayers at each point x have a positive Engel curve slope — some individual Engel curves are upward sloping at that point x (though not necessarily upward-sloping for expenditures higher or lower than x). In contrast, we do not reject homogeneity of the marginal effect for expenditures above $14000. Accordingly, heterogeneity in slopes may be ignored for these expenditure values. The average effect is sufficient to summarize the response of share of food at home to a change in expenditures.

Results: skewness and kurtosis We present estimates of the skewness and kurtosis of the Engel curve slope on figure 9. Results are plotted for expenditures for which the variance of slope is significantly different from 0. First, out estimates are consistent with a light-tailed symmetric distribution for all expenditure values. Second, the estimates are more precise for expenditures below $8000, and we split the results accordingly. This effect is driven by dividing by progressively smaller values of standard deviation as expenditure increases.

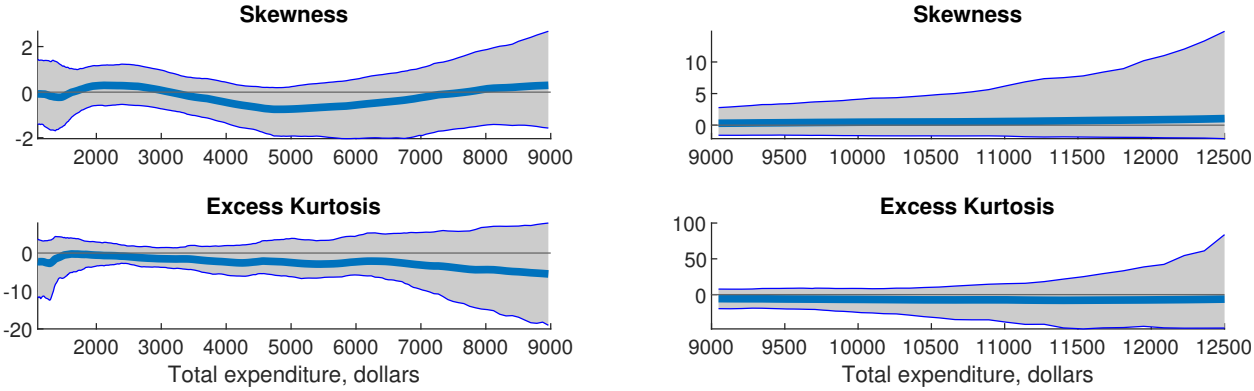

Figure 9: Skewness and excess kurtosis of Engel curve slope for stayers for food at home and 95% bootstrap pointwise confidence band. Note: moments estimated using log expenditure (log dollars); results reported in expenditure (dollars)

Distribution We now turn towards the distribution of the individual Engel curve slopes. On fig. 10 we plot a five component mixture fitted using the first five moments, starting from the zeroth moment (below we discuss robustness of the estimates). We also report 95% bands obtained by recomputing the distribution in every bootstrap sample and taking suitable quantiles of the bootstrap estimates at each value of v and expenditure. The estimated distribution reflects the properties implied by the moment estimates. The distribution is symmetric and centered around a negative value. Its variance decreases with expenditure. The estimates are more precise for expenditures below $8000, as for skewness and kurtosis.

A non-zero fraction of households has upward-sloping sections in their Engel curve for food at home. This fraction is approximately 35% for lowest expenditure values, and drops as expenditure rises. It is not significantly different from 0% for expenditure values above

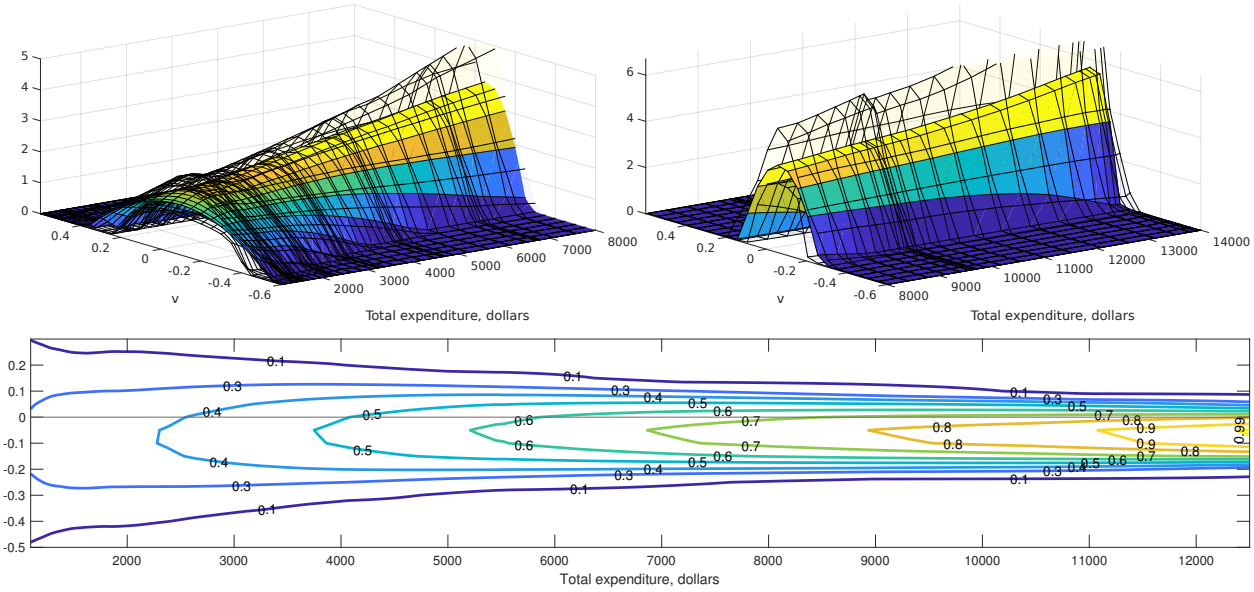

Figure 10: Distribution of Engel curve slope. Top panels: density and and 95% pointwise bootstrap intervals (transparent net). Bottom panel: contours of density; values normalized to lie in the interval [0, 1] for clarity. Results estimated using log expenditure (log dollars), reported in expenditure (dollars)

$14000. These distributional results imply that Engel's law does not necessarily hold at the household level, although it holds on average (figs. 7-8). The share of food at home may be increasing for some range of expenditures before decreasing for larger expenditure. We conjecture that this results may be driven by a combination of financial constraints and households prioritizing basic needs as their expenditures expand.

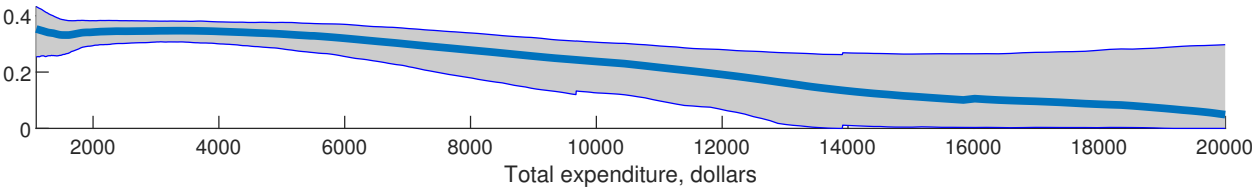

Figure 11: Share of households with a positively sloping Engel curve at given expenditure value. Shaded area: 95% bootstrap pointwise intervals. Results estimated using log expenditure (log dollars), reported in expenditure (dollars)

Distribution robustness The estimates of the distribution are robust with respect to both the number of components and the number of moments used, as was the case in our Monte Carlo study. We plot contours of the estimates for some combinations of these parameters on fig. 12. The resulting estimates are stable overall. The density displays some minor differences in shape and in the extent of the tails when using three or four components when compared to the case of using 5 components or more

Finally, we remark that the Supplementary Appendix contains a number of further results.

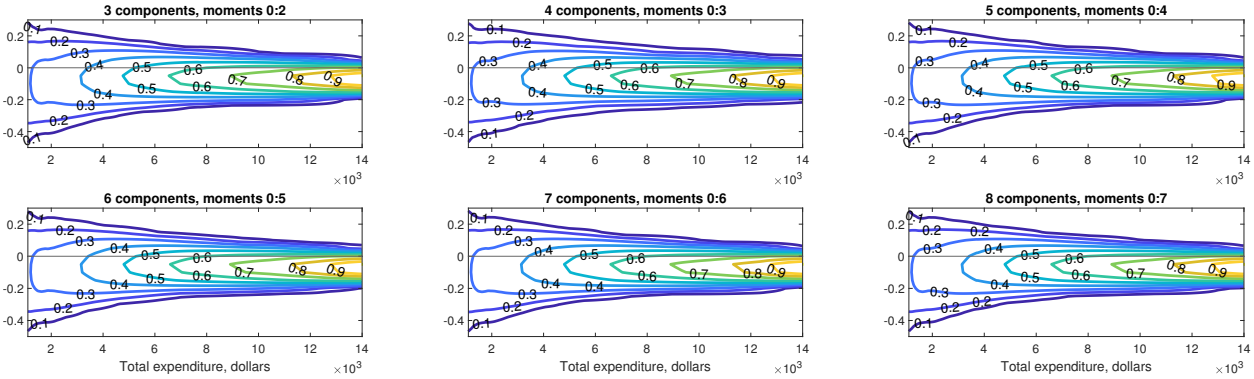

Figure 12: Sensitivity of estimated distribution of Engel curve slope to number of moments (K) and number of mixture components (p) used in estimation. Note: values of density normalized to lie in the interval [0, 1] for clarity. Results estimated using log expenditure (log dollars), but reported in expenditure (dollars)

In it, we first report estimation results for higher-order moments and moments of m and u. Second, we provide the results of applying the interval distribution estimator. Last, we provide estimates for all combinations of K and p considered. The evidence emerging from these additional results is consistent with the evidence reported here.

References

N. I. Akhiezer. *The Classical Moment Problem and Some Related Questions in Analysis*. Society for Industrial and Applied Mathematics, 1965. [Cited on page 10.]

C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Berlin Heidelberg, 3 edition, 2006. [Cited on page 50.]

I. Almås. International Income Inequality: Measuring PPP Bias by Estimating Engel Curves for Food. *American Economic Review*, 102(2):1093–1117, 2012. [Cited on page 32.]

J. G. Altonji and R. L. Matzkin. Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regerssors. *Econometrica*, 73(4):1053–1102, 2005. [Cited on pages 4 and 6.]

J. Banks, R. W. Blundell, and A. Lewbel. Quadratic Engel Curve and Consumer Demand. *Review of Economics and Statistics*, 79(4):527–539, 1997. [Cited on pages 4 and 32.]

R. Beran and P. Hall. Estimating Coefficient Distributions in Random Coefficient Regressions. *The Annals of Statistics*, 20(4):1970–1984, 1992. [Cited on page 4.]

H. J. Bierens. Kernel Estimators of Regression Functions. In *Advances in Econometrics: Fifth World Congress*, pages 99–194. 1987. [Cited on page 69.]

R. Blundell, T. MaCurdy, and C. Meghir. Labor Supply Models: Unobserved Heterogeneity, Nonparticipation and Dynamics. In *Handbook of Econometrics*, volume 6, chapter 65, pages 4667–4775. North-Holland, 2007a. [Cited on page 2.]

R. Blundell, J. L. Horowitz, and M. Parey. Measuring The Price Responsiveness of Gasoline Demand: Economic Shape Restrictions and Nonparametric Demand Estimation. *Quantitative Economics*, 3(1):29–51, 2012. [Cited on page 4.]

R. W. Blundell, M. Browning, and I. A. Crawford. Nonparametric Engel Curves and Revealed preference. *Econometrica*, 71(1):205–240, 2003. [Cited on page 4.]

R. W. Blundell, X. Chen, and D. Kristensen. Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves. *Econometrica*, 75(6):1613–1669, 2007b. [Cited on pages 4 and 34.]

M. Browning and J. M. Carro. Heterogeneity and Microeconometrics Modeling. *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Volume III*, pages 47–74, 2007. [Cited on page 32.]

C. Brownlees and V. Morozov. Unit Averaging for Heterogeneous Panels. 2022. [Cited on page 49.]

J. Bustamante. *Bernstein Operators and Their Properties*. Birkhäuser Cham, 2017. [Cited on pages 79 and 88.]

G. Chamberlain. Multivariate Regression Models for Panel Data. *Journal of Econometrics*, 18(1):5–46, 1982. [Not cited.]

R. Charnigo and C. Srinivasan. A Multivariate Generalized Cp and Surface Estimation. *Biostatistics*, 16(2):311–325, 2015. [Cited on pages 14, 28, and 34.]

S. Chen, G. B. Dahl, and S. Khan. Nonparametric Identification and Estimation of a Censored Location-Scale Regression Model. *Journal of the American Statistical Association*, 100 (469):212–221, 2005. [Cited on page 4.]

X. Chen and D. Pouzo. Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals. *Econometrica*, 80(1):277–321, 2012. [Cited on pages 4, 27, 32, and 89.]

V. Chernozhukov, I. Fernández-Val, J. Hahn, and W. K. Newey. Average and Quantile Effects in Nonseparable Panel Models. *Econometrica*, 81(2):535–580, 2013. [Cited on page 4.]

V. Chernozhukov, I. Fernández-Val, S. Hoderlein, H. Holzmann, and W. Newey. Nonparametric Identification in Panels Using Quantiles. *Journal of Econometrics*, 188(2):378–392, 2015. [Cited on pages 2, 4, 6, 7, 32, and 33.]

S. S. Dragomir. On the Ostrowski's Inequality for Riemann-Stieltjes Integral and Applications. *Journal of Applied Mathematics and Computing*, 7(3):611–627, 2000. [Cited on pages 85 and 88.]

K. Evdokimov. Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity. 2010. [Cited on pages 2, 4, 6, and 7.]

J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Springer New York, 1996. [Cited on page 14.]

B. S. Graham and J. L. Powell. Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models. *Econometrica*, 80(5): 2105–2152, 2012. [Cited on pages 4, 6, 7, and 8.]

J. J. Heckman, J. Smith, and N. Clements. Making the Most out of Programme Evaluations and Social Experiments : Accounting for Heterogeneity in Programme Impacts. *Review of Economic Studies*, 64(4):487–535, 1997. [Cited on page 2.]

S. Hoderlein and E. Mammen. Identification of Marginal Effects in Nonseparable Models without Monotonicity. *Econometrica*, 75(5):1513–1518, 2007. [Cited on pages 2 and 4.]

S. Hoderlein and E. Mammen. Identification and Estimation of Local Average Derivatives in Non-Separable Models Without Monotonicity. *The Econometrics Journal*, 12(1):1–25, 2009. [Cited on pages 2 and 4.]

S. Hoderlein and H. White. Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects. *Journal of Econometrics*, 168(2):300–314, 2012. [Cited on pages 2, 4, 6, 7, and 8.]

J. L. Horowitz. Applied Nonparametric Instrumental Variables Estimation. *Econometrica*, 79(2):347–394, 2011. [Cited on page 4.]

I. A. Ibragimov and Y. V. Linnik. *Independent and Stationary Sequences of Random Variables*.

Wolters-Noordhoff, 1971. [Cited on page 83.]

G. W. Imbens and W. K. Newey. Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity. *Econometrica*, 77(5):1481–1512, 2009. [Cited on pages 2, 4, and 32.]

A. Kazemi, H. R. Shahdoosti, and R. M. Mnatsakanov. Hausdorff moment problem: Recovery of an unknown support for a probability density function. *Journal of Inverse and Ill-Posed Problems*, 25(6):719–731, 2017. [Cited on page 17.]

J. Q. Li and A. R. Barron. Mixture Density Estimation. In S. A. Solla, T. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12 (NIPS 1999)*, pages 279–285, 1999. [Cited on pages 15 and 24.]

E. Masry. Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates. *Journal of Time Series Analysis*, 17(6):571–599, 1996a. [Cited on pages 19, 62, 63, and 68.]

E. Masry. Multivariate Regression Estimation: Local Polynomial Fitting for Time Series. *Stochastic Processes and their Applications*, 65(1):3575–3581, 1996b. [Cited on pages 68 and 69.]

R. L. Matzkin. Nonparametric Estimation of Nonadditive Random Functions. *Econometrica*, 71(5):1339–1375, 2003. [Cited on pages 2 and 4.]

D. L. McFadden. Revealed Stochastic Preference: A Synthesis. *Economic Theory*, 26(2): 245–264, 2005. [Cited on page 32.]

R. M. Mnatsakanov. Hausdorff Moment Problem: Reconstruction of Distributions. *Statistics and Probability Letters*, 78(12):1612–1618, 2008. [Cited on page 4.]

R. M. Mnatsakanov and A. S. Hakobyan. Recovery of Distributions via Moments. In *IMS Lecture Notes - Monograph Series, Volume 57: Optimality: The Third Erich L. Lehmann Symposium*, volume 57, pages 252–265. 2009. [Cited on page 4.]

W. K. Newey and J. L. Powell. Instrumental Variable Estimation of Nonparametric Models. *Econometrica*, 71(5):1565–1578, 2003. [Cited on page 4.]

W. K. Newey, J. L. Powell, and F. Vella. Nonparametric Estimation of Triangular Simultaneous Equations Models. *Econometrica*, 67(3):565–603, 1999. [Cited on page 4.]

A. Norets. Approximation of Conditional Densities by Smooth Mixtures of Regressions. *Annals of Statistics*, 38(3):1733–1766, 2010. [Cited on pages 15 and 24.]

D. Ormoneit and H. White. An Efficient Algorithm to Compute Maximum Entropy Densities. *Econometric Reviews*, 18(2):127–140, 1999. [Cited on page 4.]

M. Ponomareva. Quantile Regression for Panel Data Models with Fixed Effects and Small T: Identification and Estimation. 2010. [Cited on page 4.]

I. E. Pritsker and R. S. Varga. Zero Distribution, the Szego Curve, and Weighted Polynomial

Approximation in the Complex Plane. *Transactions of the Americal Mathematical Society*, 349(10):4085–4105, 1997. [Cited on page 92.]

E. Saez, J. Slemrod, and S. H. Giertz. The Elasticity of Taxable Income with Respect to Marginal Tax Rates: A Critical Review. *Journal of Economic Literature*, 50(1):3–50, 2012. [Cited on page 2.]

K. Schmüdgen. *The Moment Problem*. Springer, 2017. [Cited on page 16.]

C. J. Stone. Optimal Rates of Convergence for Nonparametric Estimators. *The Annals of Statistics*, 8(6):1348–1360, 1982. [Cited on page 19.]

L. Tardella. A Note on Estimating the Diameter of a Truncated Moment Class. *Statistics and Probability Letters*, 54(2):115–124, 2001. [Cited on page 82.]

T. Tjur. A Constructive Definition of Conditional Distributions. 1975. [Cited on page 8.]

J. M. Wooldridge. Unobserved Heterogeneity and Estimation of Average Partial Effects. In *Identification and Inference for Econometric Models*, pages 27–55. 2010a. [Cited on page 6.]

J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. Number 0262232588 in MIT Press Books. The MIT Press, 2010b. [Cited on page 8.]

X. Wu. Calculation of Maximum Entropy Densities with Application to Income Distribution. *Journal of Econometrics*, 115(2):347–354, 2003. [Cited on page 4.]

Y. Wu and P. Yang. Optimal Estimation of Gaussian Mixtures via Denoised Method of Moments. *The Annals of Statistics*, 48(4):1981–2007, 2020. [Cited on pages 4 and 16.]

A. J. Zeevi and R. Meir. Density Estimation Through Convex Combinations of Densities: Approximation and Estimation Bounds. *Neural Networks*, 10(1):99–109, 1997. [Cited on pages 15 and 24.]

Appendix I: Implementation of Estimators

A Implementation of Moment Estimators

In this section we discuss some further implementation details of the moment estimator $\hat{\mu}_k(x)$ of section 3.1 in the main text. Section A.1 discusses a convenient parametrization of the first-step local polynomial (LP) regressions that estimate moments of (Y_{i1}, Y_{i2}) and their derivatives. In section A.2 we further show how to exploit the linearity of the LP estimators to run at most 3 LP regressions regardless of the order of the moment of interest. Finally, in section A.3 we discuss how the moment expressions simplify if the distribution of u_t is invariant over time.

A.1 Simple Parametrization for Local Polynomial Estimators

Step 1 of algorithm 1 requires computing local polynomial estimators for $\partial_h^l r_g(x-h, x+h)|_{h=0}$ and $\partial_h^l r_g(x \pm h, x \pm h)|_{h=0}$ for $l = 0, 1, \ldots, K$ and $g \in \mathcal{G}_K$ where

$$\mathcal{G}_K = \{(y_2 - y_1)^j, y_1^{j-1}(y_1 - y_2), y_2^{j-1}(y_2 - y_1), y_1^{j-1} y_2, j \in 1, 2, \ldots, K\}.$$

In this section we provide convenient formulas for these estimators.

Directly regressing $g(Y_{i1}, Y_{i2})$ on (X_{i1}, X_{i2}) is somewhat inconvenient for estimating the derivatives with respect to h. In principle, the derivatives of the form $\partial_h^l r_g(x-h, x+h)|_{h=0}$ and $\partial_h^l r_g(x \pm h, x \pm h)_{h=0}$ can be obtained from the partial derivatives of $r_g(x_1, x_2)$ with respect to x_1 and x_2 evaluated at $(x_1, x_2) = (x, x)$. However, the relationship between the two is given by Faà di Bruno's formula, which may be cumbersome to evaluate for $l \geq 3$.

To resolve the above issue, we propose an alternative equivalent parametrization of the regressions. This approach obviates the need for combining estimated partial derivatives according to Faà di Bruno's formula, and may be used easily in any statistical software capable of running local polynomial (LP) regressions.

Consider $r_g(x-h, x+h) \equiv \mathbb{E}[g(Y_{i1}, Y_{i2})|X_{i1} = x-h, X_{i2} = x+h]$. Define the new variables $W_{i1}^{(\Delta)} = (X_{i1} + X_{i2})/2$ and $W_{i2}^{(\Delta)} = (X_{i2} - X_{i1})/2$, and define

$$R_g^{(\Delta)}(w_1, w_2) = \mathbb{E}\left[g(Y_{i1}, Y_{i2})|W_{i1}^{(\Delta)} = w_1, W_{i2}^{(\Delta)} = w_2\right].$$

Then

$$r_g(x-h, x+h) = R_g^{(\Delta)}(0, x+h).$$

Thus, the derivative of interest can be expressed as

$$\partial_h^l r_g(x-h, x+h)|_{h=0} = \partial_{w_2}^l R_h^{(\Delta)}(w_1, w_2)|_{(w_1,w_2)=(0,x)}.$$

To estimate the derivative on the right, we regress $g(Y_{i1}, Y_{i2})$ on $(W_{i1}^{(\Delta)}, W_{i2}^{(\Delta)})$ with local polynomial regression around $(w_1, w_2) = (0, x)$. Then an estimator for $\partial_{w_2}^l R_h^{(\Delta)}(w_1, w_2)|_{(w_1,w_2)=(0,x)}$ can be directly obtained from the coefficient on $(W_{i2}^{(\Delta)})^l$ in a standard manner.

Specifically, let q be a positive integer, $q > l$. Let the $(q+1)(q+2)/2$ vector $\boldsymbol{W}_i^{(\Delta)}$ have the $(p(p+1)/2 + j + 1)$th element $(W_{i1}^{(\Delta)})^{p-j}(W_{i2}^{(\Delta)})^j$, where $j \in \{0, 1, 2, \ldots, p\}$ and $p \in \{0, 1, 2, \ldots, q\}$, and let the $N \times (q+1)(q+2)/2$ matrix $\boldsymbol{W}^{(\Delta)}$ have the ith row $\boldsymbol{W}_i^{(\Delta)}$, that is,

$$\boldsymbol{W}_i^{(\Delta)} = \left(1 \quad W_{i1}^{(\Delta)}, W_{i2}^{(\Delta)}, \left(W_{i1}^{(\Delta)}\right)^2, W_{i1}^{(\Delta)} W_{i2}^{(\Delta)}, \left(W_{i2}^{(\Delta)}\right)^2, \left(W_{i1}^{(\Delta)}\right)^3, \ldots, \left(W_{i2}^{(\Delta)}\right)^q\right),$$
$$\boldsymbol{W}^{(\Delta)} = \left(\boldsymbol{W}_1^{(\Delta)\prime}, \boldsymbol{W}_2^{(\Delta)\prime}, \ldots, \boldsymbol{W}_N^{(\Delta)\prime}\right)'.$$

Let ψ_{LP} be a kernel that satisfies assumption 4.1 and let s be a positive number. Then let $\boldsymbol{\Psi}$ be the $N \times N$ diagonal matrix with (i,i)th element given by $s^{-2}\psi_{LP}(W_{i1}^{(\Delta)}/s, (W_{i2}^{(\Delta)} - x)/s)$. In addition, let $g(\boldsymbol{Y})$ be the $N \times 1$ vector with ith element given by $g(Y_{i1}, Y_{i2})$. The LP(q) coefficient vector $\hat{\boldsymbol{\beta}}_g$ of regressing $g(Y_{i1}, Y_{i2})$ on $(W_{i1}^{(\Delta)}, W_{i2}^{(\Delta)})$ is defined as

$$\hat{\boldsymbol{\beta}}_g = \left(\boldsymbol{W}^{(\Delta)\prime}\boldsymbol{\Psi}\boldsymbol{W}^{(\Delta)}\right)^{-1}\boldsymbol{W}^{(\Delta)\prime}\boldsymbol{\Psi}g(\boldsymbol{Y}). \tag{34}$$

The LP(q) estimator for $\partial_{w_2}^l R_h^{(\Delta)}(w_1, w_2)|_{(w_1,w_2)=(0,x)}$ is then given by

$$\widehat{\partial_{w_2}^l R_h^{(\Delta)}(w_1, w_2)|_{(w_1,w_2)=(0,x)}} = l!\left(\hat{\boldsymbol{\beta}}_g\right)_{(l+1)(l+2)/2}, \tag{35}$$

where $(\hat{\boldsymbol{\beta}}_g)_j$ stands for the jth element of $\hat{\boldsymbol{\beta}}_g$.

The argument is analogous for $r_g(x \pm h, x \pm h)$. Define the new variables $W_{i1}^{(\pm)} = (X_{i2} - X_{i1})/2$ and $W_{i2}^{(\pm)} = \pm(X_{i1} + X_{i2})/2$. Define

$$R_g^{(\pm)}(w_1, w_2) = \mathbb{E}\left[g(Y_{i1}, Y_{i2})|W_{i1}^{(\pm)} = w_1, W_{i2}^{(\pm)} = w_2\right].$$

Then

$$r_g(x \pm h, x \pm h) = R_g^{(\pm)}(0, \pm x + h).$$

Thus,

$$\partial_h^l r_g(x \pm h, x \pm h)|_{h=0} = \partial_{w_2}^l R_g^{(\pm)}(w_1, w_2)|_{(w_1,w_2)=(0,\pm x)}.$$

As above, the derivative of interest is estimated by the coefficient on $(W_{i2}^{(\pm)})^l$ in the local polynomial regression of $g(Y_{i1}, Y_{i2})$ on $(W_{i1}^{(\pm)}, W_{i2}^{(\pm)})$ around $(0, \pm x)$, similarly to eq. (35).

A.2 Using Linearity of the LP(q) Estimator

An additional computational simplification is made possible by the linearity of the LP(q) estimator. Let $\widehat{\partial_h^l r_g(x-h, x+h)|_{h=0}}$ be the estimator for $\partial_h^l r_g(x-h, x+h)$ constructed as above. As the LP(q) estimator is linear, it it possible to represent $\widehat{\partial_h^l r_g(x-h, x+h)|_{h=0}}$ as

$$\widehat{\partial_h^l r_g(x-h, x+h)|_{h=0}} = \sum_{i=1}^N \mathrm{W}_i(x) g(Y_{i1}, Y_{i,2}), \tag{36}$$

where the weights $\mathrm{W}_{i,N}$ do not depend on g or (Y_{i1}, Y_{i2}). The weights $\mathrm{W}_{iN}(x)$ for all $l = 0, 1, \ldots, K$ are determined by a single matrix inversion as in eq. (34).

The estimators $\widehat{\partial_h^l r_g(x-h, x+h)|_{h=0}}$ may then be rapidly computed using eq. (36) for all $g \in \mathcal{G}_K$ once the $\mathrm{W}_i(x)$ are determined. In other words, only a single evaluation of weights is required. The same point applies to constructing estimators of $\partial_h^l r_g(x \pm h, x \pm h)$. We conclude that overall all the required estimators can be computed using a total of 3 applications of LP(q).

A.3 Exploiting Stationarity of u_{it}

If u_{it} is stationary, the expressions for moments of u_{i1} and u_{i2} of alg. 1 simplify. Let $\nu_{u^k}(x)$ be the (time-invariant) kth moment of u_{it}. Then the lth derivative of $\nu_{(u_2-u_1)^p}(x, h)$ is given by

$$\partial_h^l \nu_{(u_2-u_1)^p}(x,h) = \sum_{j=0}^p \binom{p}{j} \left[\sum_{i=0}^l \binom{l}{i} \left(\partial_h^i \nu_{u^j}(x-h)\right) \left(\partial_h^{l-i} \nu_{u^{p-j}}(x+h)\right) \right].$$

Both equations (8) and (9) identify ν_{u^k}, leading to overidentification. A simple way of combining the two possible expressions is averaging them as

$$\begin{aligned}
\nu_{u^p}(x \pm h) = &\frac{r_{y_1^{p-1}(y_1-y_2)}(x \pm h, x \pm h) + r_{y_2^{p-1}(y_2-y_1)}(x \pm h, x \pm h)}{2} \\
&- \sum_{j=1}^{p-1} \binom{p-1}{j} \nu_{m^j}(x \pm h) \nu_{u^{p-j}}(x \pm h), \\
\nu_{m^p}(x \pm h) = &\frac{r_{y_1^{p-1}y_2}(x \pm h, x \pm h) + r_{y_2^{p-1}y_1}(x \pm h, x \pm h)}{2}
\end{aligned}$$

$$- \sum_{j=1}^{p-1} \binom{p-1}{j-1} \nu_{m^j}(x \pm h) \nu_{u^{p-j}}(x \pm h).$$

The lth derivatives of the above expressions with respect to h are given by

$$\begin{aligned}
\partial_h^l \nu_{u^p}(x \pm h) = & \frac{\partial_h^l r_{y_1^{p-1}(y_1-y_2)}(x \pm h, x \pm h) + \partial_h^l r_{y_2^{p-1}(y_2-y_1)}(x \pm h, x \pm h)}{2} \\
& - \sum_{j=1}^{p-1} \binom{p-1}{j} \left[\sum_{i=0}^{l} \binom{l}{i} \left(\partial_h^i \nu_{m^j}(x \pm h) \right) \left(\partial_h^{l-i} \nu_{u^{p-j}}(x \pm h) \right) \right], \\
\partial_h^l \nu_{m^p}(x \pm h) = & \frac{\partial_h^l r_{y_1^{p-1} y_2}(x \pm h, x \pm h) + \partial_h^l r_{y_2^{p-1} y_1}(x \pm h, x \pm h)}{2} \\
& - \sum_{j=1}^{p-1} \binom{p-1}{j-1} \left[\sum_{i=0}^{l} \binom{l}{i} \left(\partial_h^i \nu_{m^j}(x \pm h) \right) \left(\partial_h^{l-i} \nu_{u^{p-j}}(x \pm h) \right) \right].
\end{aligned}$$

B Implementation of Distribution Estimators

In this section, we provide convenient representations for the optimization problems associated with the distribution estimators of section 3.2. Section B.1 discusses the case of estimating the distribution at a single point x, while section B.2 is dedicated to the case of estimating the distribution for an interval of values of x. In both cases, the estimation problem is reduced to a standard quadratic program.

B.1 Estimation at One Point

We first consider the problem of estimating $F_0(v|x_0)$ for a fixed point x_0. Recall that $F_0(\cdot|x_0)$ is approximated using mixture cdfs of the form

$$\Lambda_p(v|\boldsymbol{\gamma}) = \sum_{j=1}^{p} \gamma_j \Psi(v - v_{j,p}),$$

where p is the number of components (the dimension of the corresponding sieve space $\mathcal{L}_p$), Ψ is a cdf, and $v_{1,p} < \cdots < v_{p,p}$ are fixed centers.

We introduce some additional notation. Define

$$\boldsymbol{\Omega} = \text{diag}\{0!, 1!, \ldots, (K-1)!\}.$$

Let $\boldsymbol{M}_{\Psi,K}$ be a $K \times p_v$ matrix with (k,j)th element given by $\int v^{k-1} \Psi(d(v - v_{j,p}))dv$. Note that $\boldsymbol{M}_{\Psi,K}$ can be conveniently evaluated using formula (87) that requires only evaluating

$\int v^k \Psi(dv)$ for $k = 0, \ldots, K-1$. Further, let $\tilde{\mu}_k(x)$ be an estimator for $\mu_k(x)$ and define

$$\tilde{\boldsymbol{\mu}}_K(x) = (\tilde{\mu}_0(x), \tilde{\mu}_1(x), \ldots, \tilde{\mu}_{K-1}(x))'. \tag{37}$$

The function $\tilde{Q}_N(\boldsymbol{\gamma}|x_0)$ of eq. (17) in the main text can be represented as

$$\begin{aligned}\tilde{Q}_N(\boldsymbol{\gamma}|x_0) &= [\tilde{\boldsymbol{\mu}}_K(x_0) - \boldsymbol{M}_{\Psi,K}\boldsymbol{\gamma}]' \, \boldsymbol{\Omega} \, [\tilde{\boldsymbol{\mu}}_K(x_0) - \boldsymbol{M}_{\Psi,K}\boldsymbol{\gamma}] \\ &= \boldsymbol{\gamma}' \boldsymbol{M}'_{\Psi,K} \boldsymbol{\Omega} \boldsymbol{M}_{\Psi,K} \boldsymbol{\gamma} - 2\tilde{\boldsymbol{\mu}}_k(x_0) \boldsymbol{M}_{\Psi,K} \boldsymbol{\gamma} + \tilde{\boldsymbol{\mu}}_k(x_0)' \tilde{\boldsymbol{\mu}}_k(x_0)\end{aligned}$$

It immediately follows that the estimated weights $\tilde{\boldsymbol{\gamma}}$ can be characterized as the solution of the following quadratic program:

$$\tilde{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}: \sum_{j=1}^K \gamma_j = 1, \gamma_j \geq 0}{\arg\min} \; \boldsymbol{\gamma}'(\boldsymbol{M}'_{\Psi,K} \boldsymbol{\Omega} \boldsymbol{M}_{\Psi,K} + \lambda_N \boldsymbol{I}_p)\boldsymbol{\gamma} - 2\tilde{\boldsymbol{\mu}}_k(x_0) \boldsymbol{M}_{\Psi,K} \boldsymbol{\gamma}$$

where $\lambda_N \geq 0$ and $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. Note that the Hessian is positive definite if $K \geq p$ or $\lambda_N > 0$.

B.2 Estimation Over an Interval

We now consider the problem of estimating $F_0(v|x)$ as x range in the interval $I = [x_{lb}, x_{ub}]$. $F_0(\cdot|\cdot)$ is approximated using a finite mixture whose mixture probabilities are given by Bernstein polynomials:

$$\begin{aligned}\Lambda_{p_v,p_x}(v|x, \boldsymbol{\gamma}) &= \sum_{j=1}^{p_v} \left[\sum_{l=0}^{p_x} \gamma_{j,l} b_{l,p_x}(x) \right] \Psi\left(v - v_{j,p_v}\right), \\ b_{l,p_x}(x) &= \binom{p_x}{l} \left(\frac{x - x_{lb}}{x_{ub} - x_{lb}} \right)^l \left(\frac{x_{ub} - x}{x_{ub} - x_{lb}} \right)^{p_x - l}.\end{aligned}$$

Let the $p_v \times (p_x + 1)$ matrix $\boldsymbol{\gamma}$ have the (j,l)th element $\gamma_{j,l-1}$. Let π be a finite measure on I such that the Lebesgue measure is absolutely continuous with respect to π. Let the objective function $\hat{Q}$ be as in eq. (21):

$$\hat{Q}(\boldsymbol{\gamma}) = \int \sum_{k=0}^{K-1} \frac{1}{k!} \left[\hat{\mu}_k(x) - \int v^k \sum_{j=1}^{K} \left[\sum_{l=0}^{p_x} \gamma_{j,l} b_{l,p_x}(x) \right] \Psi(d(v - v_{j,p_v})) \right]^2 \pi(dx).$$

The estimated weights are found as in eq. (23):

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}: \gamma_{j,l} \geq 0, \sum_{j=1}^{p_v} \gamma_{j,l} = 1 \; \forall l}{\arg\min} \; \hat{Q}_N(\boldsymbol{\gamma}) + \lambda_N^I \sum_{j,l} \gamma_{j,l}^2, \quad \lambda_N \geq 0. \tag{38}$$

where $\lambda_N^I \geq 0$. The objective function is strictly convex if $K \geq p_v$ or $\lambda_N^I > 0$.

It may be challenging to estimate $\hat{\boldsymbol{\gamma}}$ using the above formulation of the problem. Every evaluation of the objective function would require evaluating the π-integrals. Further, it may be inconvenient to optimize over the matrix $\boldsymbol{\gamma}$.

However, the following proposition shows that $\hat{\boldsymbol{\gamma}}$ may be obtained by solving a quadratic problem that only requires evaluating the integrals once. We introduce some notation first. Define $\boldsymbol{\Omega} = \text{diag}\{0!, 1!, \ldots, (K-1)!\}$. Let $\boldsymbol{M}_{\Psi,K}$ be a $K \times p_v$ matrix with (k,j)th element given by $\int v^{k-1}\Psi(d(v - v_{j,p_v}))dv$. Define the matrices $\boldsymbol{H}$ and $\boldsymbol{C}$ as

$$\begin{aligned} \boldsymbol{H} &= \left((\boldsymbol{M}'_{\Psi,K}\boldsymbol{\Omega}\boldsymbol{M}_{\Psi,K}) \otimes \boldsymbol{I}_{p_x+1}\right), \\ (\boldsymbol{C})_{ij} &= \int b_{i-1,p_x}(x) b_{j-1}(p_x)(x)\pi(dx). \end{aligned} \tag{39}$$

Note that $\boldsymbol{C}$ is positive definite by definition of π and since Bernstein polynomial of order p_x form a system of (p_x+1) linearly independent functions. Correspondingly, let $\boldsymbol{C}^{1/2}$ be the (unique) positive definite matrix such that $\boldsymbol{C} = \boldsymbol{C}^{1/2}\boldsymbol{C}^{1/2}$. Let $\boldsymbol{C}^{-1/2} = (\boldsymbol{C}^{1/2})^{-1}$. Next, define $\boldsymbol{W} = (\text{vec}(\boldsymbol{I}_{p_v}) \otimes \boldsymbol{I}_{p_x+1})\, C^{-1}\, (\text{vec}(\boldsymbol{I}_{p_v})' \otimes \boldsymbol{I}_{p_x+1})$. Define $\boldsymbol{V}$ to be the $p_v(p_x+1) \times p_v(p_x+1)$ matrix with the (i,j) element given by $(\boldsymbol{V})_{ij} = \left[\sum_{k=1}^{p_v} w_{p_v(p_x+1)(k-1)+i, p_v(p_x+1)(k-1)+j}\right]$. Finally, let $\boldsymbol{h}$ be the $p_v(p_x+1)$-vector $\boldsymbol{h}$ with $((p_x+1)(j-1)+i)$th element given by

$$(\boldsymbol{h})_{(p_x+1)(j-1)+i} = \sum_{k=0}^{K+1}\sum_{l=0}^{p_x} (\boldsymbol{C}^{-1/2})_{l+1,i}\, (\boldsymbol{\Omega}\boldsymbol{M}_{\psi,K})_{k+1,j} \int \tilde{\mu}_k(x) b_{l,p_x}(x)\pi(dx), \tag{40}$$

where $j = 1, \ldots, K$ and $i = 1, \ldots, p_x$.

Proposition 1. *Let $\hat{\boldsymbol{\gamma}}$ be as in eq. (23). Then*

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{I}_{p_v} \otimes \hat{\boldsymbol{g}}')\, (\text{vec}(\boldsymbol{I}_{p_v}) \otimes \boldsymbol{I}_{p_x+1})\, \boldsymbol{C}^{-1/2},$$

where

$$\hat{\boldsymbol{g}} = \arg\min_{\boldsymbol{g}} \boldsymbol{g}'(\boldsymbol{H} + \lambda_N^I \boldsymbol{V})\boldsymbol{g} - 2\boldsymbol{h}'\boldsymbol{g} \tag{41}$$

subject to $(\imath'_K \otimes \boldsymbol{I}_{p_x+1})\,\boldsymbol{g} = \boldsymbol{C}^{1/2}\imath_{p_x+1}$ and $(\boldsymbol{I}_k \otimes \boldsymbol{C}^{-1/2})\boldsymbol{g} \geq 0$.

As proposition 1 shows, the problem of finding $\hat{\boldsymbol{\gamma}}$ may be reduced to the quadratic program (41). Solving (41) requires evaluating the π-integrals only once, to construct $\boldsymbol{H}$ and $\boldsymbol{h}$. Note that the new Hessian is positive definite if $\lambda_N^I > 0$ or $K - 1 \geq p_v$ (as implied by lemma H.1). The proof of proposition 1 may be found in section Eof the Proof Appendix.

Appendix II: Proofs of Results in the Main Text

C Identification

C.1 Proof of Lemma 2.1

Assumption C.1 (Regularity assumptions for identification)**.** *(i)* $\mathbb{X} \times \mathbb{A}$ *is equipped with the product Borel* σ*-algebra. (ii)* $\sup_{a,x} \left| \partial_x^{(l)} m(x,a) \right| < \infty$ *where* $\partial_x^{(l)} m$ *is the* l*th derivative of* $m(x,a)$ *with respect to* x *for* $l = 0, 1, 2$*. (iii)* $m(x,a)$ *is well-defined for every* $x \in \text{supp}(X_{i1}) \cup \text{supp}(X_{i2})$*. (iv)* $\partial_x^{(l)} m(x,a) < \infty$ *is a measurable function of* a *for every* x *for* $l = 0, 1, 2$ *. Further,* $\partial_x m(x,a)$ *is a continuous function of* a *with respect to the topology of* $\mathbb{A}$ *for every* x*.*

Proof of lemma 2.1. We first consider identification of $\nu_{m^k}(x)$ and $\nu_{u_1^k}(x)$ for $x \in [x_{lb} - \epsilon, x_{ub} + \epsilon]$. Observe that under assumption C.1 $\nu_{m^k}(x)$ is well-defined for all positive integers k. Let $B_{x,0} = \{X_{i1} = x, X_{i2} = x\}$. We proceed by (finite) induction on k. First, let $k = 1$. Then by assumption 2.1 $\mathbb{E}[u_{i1} | X_{i1} = x] = 0$. Further, then $\mathbb{E}[Y_{i1} | B_{x,0}] = \nu_m(x)$. Now suppose that the identification result holds for lth moments, $l = 1, \ldots, k-1$, and consider the kth moments. Consider the kth moment of Y_{i1} conditional on $B_{x,0}$.

$$\mathbb{E}[Y_{i1}^k | B_{x,0}] = \mathbb{E}\left[(m(x, \alpha_i) + u_{i1})^k | B_{x,0} \right] = \sum_{l=0}^{k} \binom{k}{l} \nu_{m^l}(x) \nu_{u_1^{k-l}}(x). \tag{42}$$

where the second equality follows by the conditional independence of α_i and u_{i1} (assumption 2.1). All terms in eq. (42) except $\nu_{m^k}(x)$ nd $\nu_{u_1^k}(x)$ are identified by the inductive assumption, as they are moments of order $< k$.

Now consider the expected value of $Y_{i1}^{k-1} Y_{i2}$ conditional on $B_{x,0}$:

$$\begin{aligned}
\mathbb{E}\left[Y_{i1}^{k-1} Y_{i2} | B_{x,0} \right] &= \sum_{l=0}^{k-1} \binom{k-1}{l} \mathbb{E}\left[m^{l+1}(x, \alpha_i) u_{i1}^{(k-1)-l} | B_{x,0} \right] \\
&\quad + \sum_{l=0}^{k-1} \binom{k-1}{l} \mathbb{E}\left[m^{l}(x, \alpha) u_{i1}^{(k-1)-l} u_{i2} | B_{x,0} \right] \\
&= \sum_{l=1}^{k} \binom{k-1}{l-1} \nu_{m^l}(x) \nu_{u_1^{k-l}}(x)
\end{aligned} \tag{43}$$

where equality follows as α_i, u_{i1}, and u_{i2} are independent and $\nu_{u_2}(x) = 0$ by assumption 2.1.

Now subtract eq. (43) from eq. (42) to obtain

$$\mathbb{E}\left[Y_{i1}^{k-1}(Y_{i1}-Y_{i2})|B_{x,0}\right]=\nu_{u_1^k}(x)+\sum_{l=1}^{k-1}\left[\binom{k}{l}-\binom{k-1}{l-1}\right]\nu_{m^l}(x)\nu_{u_1^{k-l}}(x)$$

Observe that the left hand side is $r_{y_1^{k-1}(y_2-y_1)}(x,x)$, which is identified under assumption 2.3. All the moments in the sum are of order $< k$ and thus identified by the inductive assumption. Rearranging, we obtain eq. (8). Similar logic applies to $\nu_{u_2^k}(x)$.

To obtain eq. (10), we write eq. (42) as $\nu_{m^k}(x)=r_{y_1^k}(x,x)-\sum_{l=1}^{k-1}\nu_{m^l}(x)\nu_{u_1^{k-l}}(x)-\nu_{u_1^k}(x)$. Substituting (8) and combining terms, we get eq. (10). We conclude that all kth moments are identified and writable in the required form, finishing the inductive step.

Last, eq. (11) follows from eq. (5). Observe that $\nu_{(u_2-u_1)^k}(x,h)=\sum_{j=0}^{k}\binom{k}{j}\nu_{u_1^j}(x-h)\nu_{u_2^{k-j}}(x+h)$. By part 2 of the lemma, each $\nu_{u_t^j}(x)$ term is identified for all $x\in[x_{lb}-\epsilon,x_{ub}+\epsilon]$. As $|h|<\epsilon$, identification follows. □

C.2 Proof of Theorem 2.2

Before proving theorem 2.2, we prove a number of intermediate technical results. The following lemma shows that the point $\tilde{x}=\tilde{x}(\alpha_i)$ of eq. (3) in the main text can be chosen measurably.

Lemma C.1. *Let assumptions 2.1-2.3 and C.1 hold. Let ϵ be as in assumption 2.3 and let h satisfy $|h|\in[0,\epsilon)$. Then*

(1) There exists a measurable function $\lambda_h(a):\mathbb{A}\to[-1,1]$ such that for all $a\in\mathbb{A}$

$$\frac{m(x+h,a)-m(x-h,a)}{2h}=m'(x+\lambda_h(a)h,a).$$

(2) Let k be a fixed natural number. Let $\lambda_h(a)$ be a measurable function of a such that $\lambda_h(a)\in[-1,1]$ for all $a\in\mathbb{A}$. Then there exists a measurable function $\kappa_h(a):\mathbb{A}\to[0,1]$ such that for all $a\in A$

$$\begin{aligned}&\left[\partial_x m(x_0+\lambda_h(a)h,a)\right]^k-\left[\partial_x m(x_0,a)\right]^k\\&\qquad=hk\lambda_h(a)\partial_x^2 m(x_0+\kappa_h(a)\lambda_h(a)h,a)\left[\partial_x m(x_0+\kappa_h(a)\lambda_h(a)h,a)\right]^{k-1}.\end{aligned}$$

The proof of similar to that of lemma A.1.1 in Brownlees and Morozov (2022).

Proof. Fix h and $x\in I$ and define the function $f(a,y):A\times[-1,1]\to\mathbb{R}$ as

$$f(a,y)=m(x+h,a)-m(x-h,a)-2h\partial_x m(x+yh,a)$$

We first discuss some properties of $f(a, y)$. First, $f(a, y)$ is well-defined. This holds as $m(x + yh, a)$ is well-defined for any $y \in [-1, 1]$ by the fact that $x + yh \in [x_{lb} - \epsilon, x_{ub} + \epsilon]$ and by definition of ϵ and assumption C.1. Second, $f(a, y)$ is measurable in a for any $y \in [-1, 1]$ by assumption C.1. Third, $f(a, y)$ is continuous in y on $[-1, 1]$ under assumption C.1.

Define the correspondence $\phi(a) : \mathbb{A} \to [-1, 1]$ as

$$\varphi(a) = \{y \in [-1, 1] : f(a, y) = 0\}.$$

First, $\varphi(a)$ is a measurable correspondence as $g(a, y)$ satisfies the assumptions of corollary 18.8 in Aliprantis and Border (2006). Second, $\varphi(a)$ is non-empty for every value of a by the mean value and the intermediate value theorems. Third, $\varphi(a)$ is closed for every value a since $\partial_x m(\cdot, a)$ is is continuous by assumption C.1. Then by the Kuratowski–Ryll-Nardzewski measurable selection theorem (theorem 18.3 in Aliprantis and Border (2006)) $\varphi(a)$ admits a measurable selector, which we label $\lambda_h(a)$. By construction, $\lambda_h(a)$ satisfies the requirement of the lemma.

The second assertion of the lemma is proved analogously. □

Recall the notation $B_{x-h,2h} = \{X_{i1} = x - h, X_{i2} = x + h\}$. The following lemma shows that $\mathbb{E}\left[(\partial_x m(\tilde{x}, \alpha))^k \,|B_{x-h,2h}\right]$ converges to $\mu_k(x)$. This establishes the first condition for identification outlined under eq. (4).

Lemma C.2. *Let assumption 2.1-2.3 and C.1 hold. Suppose that $\lambda_h(a)$ is a measurable function of a such that $\lambda_h(a) \in [-1, 1]$ for all $a \in A$ for each h such that $|h| < \epsilon$. Let k be a natural number. Then as $h \to 0$ it holds that*

$$\mathbb{E}\left[(\partial_x m(x + \lambda_h(\alpha_i)h, \alpha_i))^k | X_{i1} = x - h, X_{i2} = x + h\right] \to \mu_k(x), \tag{44}$$

where both conditional expectations are well-defined.

Proof. By lemma C.1, there exists a measurable function $\kappa_h(a) : \mathbb{A} \to [0, 1]$ such that for all $a \in A$

$$\begin{aligned} &[\partial_x m(x + \lambda_h(a)h, a)]^k - [\partial_x m(x, a)]^k \\ &= h\lambda_h(a)k\partial_x^2 m(x + \kappa_h(a)\lambda_h(a)h, a) \left[\partial_x m(x + \kappa_h(a)\lambda_h(a)h, a)\right]^{k-1}. \end{aligned} \tag{45}$$

Observe that both sides are bounded and measurable as functions of a by assumption C.1 and measurability of λ_h and κ_h.

Recall that $F_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(\cdot)$ denotes the law of α given $\{X_{i1} = x_1, X_{i2} = x_2\}$, which is well-

defined by assumption 2.2. Using eq. (45), we consider the expectation in question:

$$
\begin{aligned}
&\mathbb{E}\left[(\partial_x m(x+\lambda_h(\alpha_i),\alpha_i))^k | X_{i1}=x-h, X_{i2}=x+h\right] \\
&= \int (\partial_x m(x+\lambda_h(a)h,a))^k \, F_{\alpha|\boldsymbol{X}=(x-h,x+h)}(da) \\
&= \int (\partial_x m(x,a))^k F_{\alpha|\boldsymbol{X}=(x-h,x+h)}(da) \\
&\quad + hk \int \lambda_h(a)\partial_x^2 m\left(x+\kappa_h(a)\lambda_h(a)h,a\right)\left(\partial_x m(x+\kappa_h(a)\lambda_h(a)h,a)\right)^{k-1} F_{\alpha|\boldsymbol{X}=(x-h,x+h)}(da) \\
&= (I)+(II).
\end{aligned}
$$

Consider the terms (I) and (II) separately. Examine (I). $(\partial_x m(x,a))^k$ is a bounded continuous function of a by assumption C.1. Hence, by assumption 2.2 it holds that $\int(\partial_x m(x,a))^k F_{\alpha|\boldsymbol{X}=(x-h,x+h)}(da) \to \int(\partial_x m(x,a))^k F_{\alpha|\boldsymbol{X}=(x,x)}(da) \equiv \mu_k(x)$. Consider (II):

$$
\begin{aligned}
&hk\left|\int \lambda_h(a)\partial_x^2 m\left(x+\kappa_h(a)\lambda_h(a)h,a\right)\left(\partial_x m(x+\kappa_h(a)\lambda_h(a)h,a)\right)^{k-1} F_{\alpha|X_1,X_2}(da|x,x+h)\right| \\
&\leq hk \sup_{a,y}\left|\partial_x^2 m(y,a)\right| \sup_{a,y}|\partial_x m(y,a)|^{k-1} \\
&\to 0,
\end{aligned}
$$

where the last line follows by assumption C.1 and $h \to 0$. Hence $(II) \to 0$. Combining the above arguments, we obtain eq. (44). □

Proof of theorem 2.2. Fix $x \in I$. Let $|h| < \epsilon$. By lemma C.1, there exists a measurable function $\lambda_h(a) : \mathbb{A} \to [-1,1]$ such that for all $a \in \mathbb{A}$ it holds that $(2h)^{-1}(m(x+h,\alpha_i) - m\left(x-h,\alpha_i\right)) = \partial_x m(x+\lambda_h(\alpha_i)h,\alpha_i)$. Raising both sides to the kth power and taking expectations conditional on $B_{x-h,2h} = \{X_{i1}=x-h, X_{i2}=x+h\}$ we get that

$$
\Delta_k(x,h) = \mathbb{E}\left[\left(\partial_x m(x+\lambda_h(\alpha_i)h,\alpha_i)\right)^k \middle| B_{x-h,2h}\right] \tag{46}
$$

which is well-defined by assumption 2.2. By lemma 2.1 $\Delta_k(x,h)$ is identified for all $x \in I$ and h such that $|h| \in (0,\varepsilon)$. By lemma C.2, it holds that $\mathbb{E}[(\partial_x m(x+\lambda_h(\alpha)h,\alpha))^k | B_{x-h,2h}] \to \mu_k(x)$ as $h \to 0$. Together with eq. (46), this fact implies that $\lim_{h\to 0}\Delta_k(x,h) = \mu_k(x)$. Since the expression under the limit is identified for every h, so is the limit as $h \to 0$. □

C.3 Proof of Theorem 2.3

Proof of theorem 2.3. By assumption C.1, $\partial_x m(x,\alpha)$ is a bounded random variable. Its conditional moment generating function converges for all values of its argument. Since

all moments of $\partial_x m(x_0, \alpha_i)$ are identified, so is the mgf, as $\mathbb{E}\left[\exp(t\partial_x m(x_0, \alpha_i))|B_{x_0,0}\right] = \sum_{k=0}^{\infty}(k!)^{-1}t^k\mu_k(x)$. Since the mgf uniquely determines the distribution, the distribution of $\partial_x m(x, \alpha)$ for stayers is also identified. □

D Proof of Theorem 3.1

Recall that $B_{x-h,2h} = \{X_{i1} = x - h, X_{i2} = x + h\}$. Let I and ϵ be as in assumption 2.3. Let D_k be defined as in eq. (12) in the main text:

$$D_k(x, h) = \mathbb{E}\left[(m(x + h, \alpha_i) - m(x - h, \alpha_i))^p|B_{x-h,2h}\right] \tag{47}$$

Note that D_k is well-defined if assumption C.1 holds and if $\sup_{x\in I}\mathbb{E}[|u_{it}|^k|X_{it} = x] < \infty$.

We begin by stating some technical smoothness assumptions and three lemmas that establish that D_k is indeed differentiable k times under the assumptions of theorem 3.1.

D.1 Technical Assumptions

In order to prove theorem 3.1 and obtain asymptotic properties of our estimators, we impose a number of regularity assumptions on the components of model (2). There are three principal assumptions. First, we assume u_{it} and (X_{i1}, X_{i2}) are distributed continuously with differentiable densities. Second, we require that the conditional laws of u_{it} and α_i given $\boldsymbol{X}_i = \boldsymbol{x}$ are sufficiently smooth in the conditioning argument and that $m(x, a)$ is smooth in x for all $a \in \mathbb{A}$. Third, we assume that all the derivatives involved are globally bounded.

Throughout, let τ be an integer. τ is taken to satisfy $\tau \geq k$ in theorem 3.1 and $\tau \geq q + 2$ in theorems 4.1 and 4.2.

Assumption D.1. *Assumption C.1 holds. For each $a \in \mathbb{A}$ the function $m(x, a)$ is τ times continuously differentiable in x. Further, $\sup_{a,x}\left|\partial_x^l m(x, a)\right| < \infty$, $l \leq \tau$.*

Assumption D.2 (Smoothness for $\boldsymbol{X}$)**.** *$\boldsymbol{X}_i = (X_{i1}, X_{i2})$ is continuously distributed on $\mathbb{X}$ with density $f_{\boldsymbol{X}}$. $f_{\boldsymbol{X}}$ is bounded uniformly on $\mathbb{X}$. $f_{\boldsymbol{X}}$ is continuously differentiable τ times with bounded derivatives.*

Assumption D.3 (Smoothness for u)**.** *For $t = 1, 2$, conditional on $X_{it} = x$, u_{it} is continuously distributed with density $f_{u_t|X_t=x}(v)$. The density $f_{u_t|X_t=x}(v)$ is τ times continuously differentiable in both the conditioning argument x and the argument v. All derivatives of $f_{u_t|X_t=x}(v)$ up to order τ are uniformly bounded over x and v. Further, let $\partial^j_{cond} f_{u_t|X_t=x}(v)$ be the jth partial derivative with respect to the conditioning argument x. Then $\int|u_t|^j \sup_x\left|\partial^l_{cond} f_{u_t|X_t=x}\right| < \infty$ for $l, j \leq \tau$.*

Assumption D.4 (Smoothness for α)**.** *Conditional on $\boldsymbol{X}_i = \boldsymbol{x}$, the distribution of α_i is absolutely continuous with respect to some measure π_α (common for all $\boldsymbol{x}$) with density $f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}$. $f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}$ is continuously differentiable τ times in the conditional argument $\boldsymbol{x}$. Furthermore, $\int \left\|\sup_{\boldsymbol{x}} \nabla^k_{\boldsymbol{x}} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)\right\| \pi_\alpha(da) < \infty$ and $\sup_{a,\boldsymbol{x}} \left\|\nabla^k_{\boldsymbol{x}} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)\right\| < \infty$ for $k = 1, \ldots, \tau$.*

We remark that assumption D.4 strengthens the continuity property of assumption 2.2.

D.2 Supporting Lemmas

Under our assumptions, the density of the observed data $(Y_{i1}, Y_{i2}, X_{i1}, X_{i2})$ satisfies a number of regularity properties. Let $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})$, $\boldsymbol{X} = (X_{i1}, X_{i2}), \boldsymbol{x} = (x_1, x_2)$.

Lemma D.1. *2.1-2.3, C.1, D.1-D.4 hold. Then*

(1) $\boldsymbol{Y}_i$ is continuously distributed given $\boldsymbol{X}_i = \boldsymbol{x}$ for any $\boldsymbol{x} \in \mathbb{X}$.

(2) The joint pdf $f_{\boldsymbol{Y},\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x})$ of $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ is differentiable τ times in $\boldsymbol{x}$ with all partial derivatives up to order q bounded uniformly over $\boldsymbol{y}$ and $\boldsymbol{x}$.

Proof. By eq. (2), $(Y_{it}, X_{it}) = (m(X_{it}, \alpha_i) + u_{it}, X_{it})$. Let $\boldsymbol{u}_i = (u_{i1}, u_{i2})$ and $\boldsymbol{m}(\boldsymbol{X}_{it}, \alpha) = (m(X_{i1}, \alpha_i), m(X_{i2}, \alpha_i))'$, so that $\boldsymbol{Y}_i = \boldsymbol{m}(\boldsymbol{X}_i, \alpha_i) + \boldsymbol{u}_i$.

Under assumption D.3, $\boldsymbol{u}_i$ is continuously distributed conditional on $\boldsymbol{X}_i = \boldsymbol{x}$ for any value $\boldsymbol{x}$, and thus $\boldsymbol{Y}_i$ is also continuously distributed conditional on $\boldsymbol{X}_i = \boldsymbol{x}$ (regardless of the law of α_i).

We now turn to the second assertion. By assumption 2.1, $\boldsymbol{m}(\boldsymbol{X}_i, \alpha_i)$ and $\boldsymbol{u}_i$ are independent conditional on $\boldsymbol{X}_i$. By a standard argument, we obtain that

$$f_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) = \int f_{\boldsymbol{u}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y} - \boldsymbol{m}(\boldsymbol{x}, a)) f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) \pi_\alpha(da), \tag{48}$$

where $f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}$ is the conditional density of α_i given $\boldsymbol{X}_i = \boldsymbol{x}$ with respect to some dominating measure π_α (assumption D.4). We assert that under assumptions 2.1-2.3, D.1-D.4, $f_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}|\boldsymbol{x})$ is τ times differentiable in $\boldsymbol{x}$ and for $k \leq \tau$

$$\nabla^k_{\boldsymbol{x}} f_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}) = \int \nabla^k_{\boldsymbol{x}} \left[f_{\boldsymbol{u}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y} - \boldsymbol{m}(\boldsymbol{x}, a)) f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) \right] \pi_\alpha(da), \tag{49}$$

$$\sup_{\boldsymbol{x},\boldsymbol{y}} \left\| \nabla^k_{\boldsymbol{x}} f_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{x}) \right\| < \infty, \tag{50}$$

where, throughout, $\nabla^k f(\boldsymbol{x})$ stands for a suitable vector of all kth order partial derivatives and $\left\|\nabla^k f(\boldsymbol{x})\right\|$ is the 2-norm of this vector. The ordering of partial derivatives inside the vector is irrelevant.

Consider $k = 1$. Differentiate the expression under the integral in eq. (48) with respect to x_1:

$$\left[\partial_{x_1}\left(f_{\boldsymbol{u}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{m}(\boldsymbol{x},a))\right)\right] f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) + f_{\boldsymbol{u}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{m}(\boldsymbol{x},a))\left[\partial_{x_1} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)\right]. \tag{51}$$

By assumption 2.1, $f_{\boldsymbol{u}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{u}) = f_{u_1|X_1=x_1}(u_1) f_{u_2|X_2=x_2}(u_2)$. Let the operator ∂_{cond} be defined as in assumption D.3. Analogously, let $\partial_{arg} f_{u_t|X_t=x}(v)$ be the partial derivative of $f_{u_t|X_t=x}(v)$ with respect to the main argument v. Then

$$\begin{aligned}\partial_{x_1}\left(f_{\boldsymbol{u}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{m}(\boldsymbol{x},a))\right) = f_{u_2|X_2=x_2}(y_2 - m(x_2,a))\Big[&\partial_{cond} f_{u_1|X_1=x_1}(y_1 - m(x_1,a)) \\ &- \partial_{arg} f_{u_1|X_1=x_1}(y_1 - m(x_1,a))\partial_x m(x_1,a)\Big].\end{aligned}$$

Under assumptions D.1 and D.3, $\partial_{x_1} f_{\boldsymbol{u}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{m}(\boldsymbol{x},a))$ and $f_{\boldsymbol{u}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{u})$ are both bounded uniformly over $(\boldsymbol{x},\boldsymbol{y},a,\boldsymbol{u})$ by some finite constant C_u. The expression in eq. (51) can be bounded by $C_u(\sup_{\boldsymbol{x}} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) + \sup_{\boldsymbol{x}} \left\|\nabla_{\boldsymbol{x}} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)\right\|)$. By assumption D.4

$$C_u \int \left(\sup_{\boldsymbol{x}} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) + \sup_{\boldsymbol{x}} \left\|\nabla_{\boldsymbol{x}} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)\right\|\right) \pi_\alpha(da) < \infty.$$

The same logic applies to the derivative with respect to x_2. This argument establishes differentiability of $f_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}$ with respect to $\boldsymbol{x}$ and justifies interchanging the integral and the derivative in eq. (49) for $k = 1$. Further, again using an upper bound for the expression in eq. (51), we obtain eq. (50) for $k = 1$ as

$$\begin{aligned}&\sup_{\boldsymbol{x},\boldsymbol{y}} \left\| \int \nabla_{\boldsymbol{x}} \left[f_{\boldsymbol{u}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{m}(\boldsymbol{x},a)) f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)\right] \pi_\alpha(da)\right\| \\ &\leq \int \sup_{\boldsymbol{x},\boldsymbol{y}} \left\|\nabla_{\boldsymbol{x}} \left[f_{\boldsymbol{u}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{m}(\boldsymbol{x},a)) f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)\right]\right\| \pi_\alpha(da) \\ &\leq 2C_u \int \left(\sup_{\boldsymbol{x}} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) + \sup_{\boldsymbol{x}} \left\|\nabla_{\boldsymbol{x}} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)\right\|\right) \pi_\alpha(da) < \infty.\end{aligned}$$

To extend the result to the kth derivatives ($k \leq \tau$), we note that $\partial_{x_1}^j \partial_{x_2}^{k-j} f_{\boldsymbol{u}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{y} - \boldsymbol{m}(\boldsymbol{x},a)) f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)$ can be written as a sum of terms of the form $[\partial_{x_1}^p \partial_{x_2}^l f_{\boldsymbol{u}|\boldsymbol{X}}(\boldsymbol{y} - \boldsymbol{m}(\boldsymbol{x},a))] \times [\partial_{x_1}^{j-p} \partial_{x_2}^{k-j-l} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)]$. Proceeding as above and using assumptions 2.1, D.1, D.3, and D.4, we obtain eqs. (49) and (50) for all $k \leq \tau$. Finally, since $f_{\boldsymbol{Y},\boldsymbol{X}} = f_{\boldsymbol{Y}|\boldsymbol{X}} f_{\boldsymbol{X}}$ and $f_{\boldsymbol{X}}$ is τ times continuously differentiable with bounded derivatives by assumption D.2, the second conclusion of the lemma follows. □

Let k be fixed. Define the finite class $\mathcal{G}_k$ of functions of (y_1, y_2) as

$$\mathcal{G}_k = \left\{(y_2 - y_1)^j, y_1^{j-1}(y_1 - y_2), y_2^{j-1}(y_2 - y_1), y_1^{j-1} y_2, j \in 1, 2, \ldots, k\right\}. \tag{52}$$

$\mathcal{G}_k$ is the class of functions g used in estimation of $\mu_k(x)$.

Let J be as in assumption 2.3 and $\boldsymbol{x} = (x_1, x_2)$. For $g_k \in \mathcal{G}$ and $\boldsymbol{x} \in J$, let $r_g(\boldsymbol{x})$ be defined as in eq. (6).

Lemma D.2. *Let $k < \infty$ be fixed and let $\mathcal{G}_k$ be defined as in eq. (52). Let $g \in \mathcal{G}_k$ and let $r_g(\boldsymbol{x})$ be defined as in eq. (6). Let assumptions 2.1-2.3, C.1, D.1-D.4 hold. Then $r_g(\boldsymbol{x})$ is τ times differentiable in $\boldsymbol{x}$ with all partial derivatives up to order τ bounded.*

Proof. We prove the assertion for $g(y_1, y_2) = (y_2 - y_1)^p, p \leq k$, the proof for the other $g \in \mathcal{G}_k$ is analogous. Under assumption 2.1, α_i, u_{i1} and u_{i2} are independent, and so

$$r_g(\boldsymbol{x}) = \mathbb{E}\left[(Y_{i2} - Y_{i1})^p | \boldsymbol{X}_i = \boldsymbol{x}\right] = \sum_{j=0}^{p} \binom{p}{j} \mathbb{E}\left[(m(x_2, \alpha_i) - m(x_1, \alpha_i))^j | \boldsymbol{X}_i = \boldsymbol{x}\right] \tag{53}$$
$$\times \left[\sum_{l=0}^{p-j} \binom{p-j}{l} \mathbb{E}[u_{i1}^l | X_{i1} = x_1] \, \mathbb{E}[u_{i2}^{p-j-l} | X_{i2} = x_2]\right].$$

We proceed as in the proof of lemma D.1. We show that both kinds of moments that appear in eq. (53) are differentiable τ times with bounded derivatives. First consider the differentiability of $\mathbb{E}\left[(m(x_2, \alpha_i) - m(x_1, \alpha_i))^j | \boldsymbol{X}_i = \boldsymbol{x}\right] = \int (m(x_2, a) - m(x_1, a))^j f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) \pi_\alpha(da)$. The first derivative of the expression under the integral with respect to x_1 is given by

$$-j(m(x_2, a) - m(x_1, a))^{j-1} (\partial_{x_1} m(x_1, a)) f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) + (m(x_2, a) - m(x_1, a))^j (\partial_{x_1} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)).$$

By assumption D.1, there exists some constant $C_m < \infty$ such that the above display can be bounded by $C_m \left[\sup_{\boldsymbol{x}} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) + \sup_{\boldsymbol{x}} \left\| f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) \right\|\right]$. By D.4, $\int [\sup_{\boldsymbol{x}} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) + \sup_{\boldsymbol{x}} \left\| f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) \right\|] \pi_\alpha(da) < \infty$. We conclude that the integral and the derivative may be interchanged as

$$\partial_{x_1} \mathbb{E}\left[(m(x_2, \alpha_i) - m(x_1, \alpha_i))^j | \boldsymbol{X}_i = \boldsymbol{x}\right] = \int \partial_{x_1} \left[(m(x_2, a) - m(x_1, a))^j f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)\right] \pi_\alpha(da).$$

Furthermore, the above partial derivative is bounded over $\boldsymbol{x}$ as

$$\begin{aligned}
&\sup_{\boldsymbol{x}} \partial_{x_1} \left| \mathbb{E}\left[(m(x_2, \alpha_i) - m(x_1, \alpha_i))^j | \boldsymbol{X}_i = \boldsymbol{x}\right] \right| \\
&\leq \int \sup_{\boldsymbol{x}} \left| \partial_{x_1} \left[(m(x_2, a) - m(x_1, a))^j f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a)\right] \right| \pi_\alpha(da) \\
&\leq C_m \int [\sup_{\boldsymbol{x}} f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) + \sup_{\boldsymbol{x}} \left\| f_{\alpha|\boldsymbol{X}=\boldsymbol{x}}(a) \right\|] \pi_\alpha(da) < \infty.
\end{aligned}$$

The same argument applies to the derivative with respect to x_2, showing that $\mathbb{E}[(m(x_2, \alpha_i) - m(x_1, \alpha_i))^j | \boldsymbol{X}_i = \boldsymbol{x}]$ is differentiable with bounded first derivatives. Logic similar to that of

the proof of lemma D.1 extends this to higher-order derivatives up to order τ. The same argument applies for all $j = 0, \ldots, p$.

Second, we turn to $\mathbb{E}\left[u_{i1}^l | X_{i1} = x_1\right] = \int u^l f_{u_1|X_1=x_1}(u)du$ and proceed similarly. Derivative of the expression under the integral is given by $u^l \partial_{cond} f_{u_1|X_1=x_1}(u)$ for ∂_{cond} as in assumption D.3. By assumption D.3, $\int |u|^l \sup_{x_1} \left|\partial_{cond} f_{u_1|X_1=x_1}(u)\right| du < \infty$, as $l \leq \tau$. As above, we obtain that $\partial_{x_1} \mathbb{E}\left[u_1^l | X_1 = x_1\right] = \int u^l l(\partial_{cond} f_{u_1|X_1=x_1}(u))du$ and $\sup_{\boldsymbol{x}} \left|\partial_{x_1} \mathbb{E}\left[u_1^l | X_1 = x_1\right]\right| < \infty$ The same logic applies to higher-order derivatives and to the moments of u_2.

Combining the above arguments together, we conclude that $r_{\boldsymbol{g}}(\boldsymbol{x})$ is τ times differentiable with all partial derivatives up to order τ are bounded over $\boldsymbol{x}$. □

Lemma D.3. *Let the assumptions of theorem 3.1 hold. Then $\nu_{u_t^p}(x \pm h)$, $\nu_{m^k}(x \pm h)$, $\nu_{(u_2-u_1)^p}(x, h)$, and $D_p(x, h)$ are τ times differentiable in h for all $h \in (-\epsilon, \epsilon)$ for all $x \in I$ for $p = 0, 1, \ldots, k$ with all derivatives uniformly bounded over $x \in I$ and $h \in (-\epsilon, \epsilon)$.*

Proof of lemma D.3. Consider first $\nu_{u_1^p}(x - h)$ and $\nu_{m^p}(x - h)$. We establish the result by finite induction on p. The result is immediate for $p = 0$. Consider $p = 1$. Then $\nu_{u_1}(x-h) = 0$, which is k times differentiable with respect to h. By lemma 2.1 $\nu_{m^1}(x-h) = r_{y_2}(x-h, x-h)$. Then differentiability and boundedness of derivatives follow directly from lemma D.2. Now suppose that the conclusion of the lemma holds for moments of order up to $p-1$ and consider the pth moments. Eq. (8) of lemma 2.1 expresses $\nu_{u_1^p}$ as a smooth function of $r_{y_1^{p-1}(y_1-y_2)}$ and lower-order moments of u_{i1} and m. Using the inductive assumption, we conclude that $\nu_{u_1^p}$ is τ times differentiable. Let $l \in \{0, 1, \ldots, \tau\}$. The lth derivative of $\nu_{u_1^p}(x - h)$ with respect to h is given by

$$\begin{aligned}\partial_h^l \nu_{u_1^p}(x-h) = \partial_h^l r_{y_1^{p-1}(y_1-y_2)}(x-h, x-h) \\ - \sum_{j=1}^{p-1} \binom{p-1}{j} \left[\sum_{i=0}^{l} \binom{l}{i} \left(\partial_h^i \nu_{m^j}(x-h)\right) \left(\partial_h^{l-i} \nu_{u_1^{p-j}}(x-h)\right)\right].\end{aligned}$$

From lemma D.2 and the inductive assumption it immediately follows that

$$\sup_{x \in I, h \in (-\epsilon, \epsilon)} \left|\partial_h^l \nu_{u_1^p}(x-h)\right| < \infty.$$

A similar argument establishes the conclusions for moments of u_{i1} and m.

The result for $\nu_{(u_2-u_1)^p}(x, h)$ follows from eq. (7) and from the corresponding results for the moments of u_{i1} and u_{i2}.

Finally, consider $D_p(x, h)$. We proceed by finite induction on p, as above. First, consider $p = 1$. Then

$$D_1(x, h) = r_{(y_2-y_1)}(x - h, x + h).$$

By lemma D.2 the conclusion of the lemma then holds for D_1. Now suppose that the result holds for D_j for all $j \leq p-1$. Eq. (14) expresses $D_p(x,h)$ as a smooth function of $r_{(y_2-y_1)^p}(x-h,x+h)$, $D_j(x,h)$ and $\nu_{(u_2-u_1)^{p-j}}(x,h)$ for $j=0,\ldots,p-1$. Thus, D_p is differentiable τ times. Now let $l \in \{0,1,\ldots,\tau\}$. The lth derivative of $D_p(x,h)$ with respect to h is given by

$$\begin{aligned}\partial_h^l D_p(x,h) = \partial_h^l r_{(y_2-y_1)^p}(x-h,x+h) \\ - \sum_{j=0}^{p-1} \binom{p}{j} \left[\sum_{i=0}^{l} \binom{l}{i} \left(\partial_h^i D_j(x,h)\right) \left(\partial_h^{l-i} \nu_{(u_2-u_1)^{(p-j)}}(x,h)\right) \right]\end{aligned}$$

We conclude that by the inductive assumption, the corresponding result for derivatives of $\nu^{(p-j)}_{(u_2-u_1)}(x,h)$, and lemma D.2 that $\sup_{x\in I, h\in(-\epsilon,\epsilon)} \left|\partial_h^l D_p(x,h)\right| < \infty$. □

D.3 Proof of Theorem 3.1

Proof of theorem 3.1. First, identification of $D_k(x,h)$ follows immediately from eq. (14) and from lemma 2.1. Second, lemma D.3 establishes differentiability of $D_k(x,h)$.

We now turn to the third assertion of the theorem. Recall that $B_{x-h,2h} = \{X_{i1} = x-h, X_{i2} = x+h\}$. By the mean value theorem and lemma C.1, there exists some measurable $\tilde{x} = \tilde{x}(h,\alpha_i) \in [x-h, x+h]$ such that

$$\begin{aligned}D_k(x,h) &\equiv \mathbb{E}\left[(m(x+h,\alpha_i) - m(x-h,\alpha_i))^k | B_{x-h,2h}\right] \\ &= (2h)^k \,\mathbb{E}\left[\left(m\left(\tilde{x}(h,\alpha_i),\alpha\right)\right)^k | B_{x-h,2h}\right].\end{aligned}$$

We can then represent D_k as

$$D_k(x,h) = (2h)^k \mu_k(x) + \underbrace{(2h)^k \left[\mathbb{E}\left[(m(\tilde{x}(h,\alpha_i),\alpha_i))^k | B_{x-h,2h}\right] - \mu_k(x)\right]}_{=:\theta(h)}. \tag{54}$$

By lemma D.3 $D_k(x,h)$ is differentiable k times in h for h around 0. $\theta(h)$ is the difference of two k times differentiable functions. We conclude that $\theta(h)$ is also k times differentiable with a continuous kth derivative for h around 0.

We now show that the $\theta^{(k)}(0) = 0$. Consider the kth central difference of $\theta(h)$ around $h=0$:

$$\begin{aligned}&\frac{\sum_{j=0}^{k}(-1)^j \binom{k}{j} \theta\left(\left(\frac{k}{2}-j\right)h\right)}{h^k} \\ &= 2^k \sum_{j=0}^{k} \binom{k}{j} (-1)^j \left(\mathbb{E}\left[\left(m\left(\tilde{x}\left(\left(\frac{k}{2}-j\right)h, \alpha_i\right), \alpha_i\right)\right)^k \middle| B_{x-\left(\frac{k}{2}-j\right)h, 2\left(\frac{k}{2}-j\right)h} \right] - \mu_k(x) \right)\end{aligned}$$

Consider each term under the sum separately. As in the proof of theorem 2.2, it holds for each $j \in \{0, 1, \ldots, k\}$ as $h \to 0$ that

$$\mathbb{E}\left[\left(m\left(\tilde{x}\left(\left(\frac{k}{2}-j\right)h, \alpha_i\right), \alpha_i\right)\right)^k \middle| B_{x-\left(\frac{k}{2}-j\right)h, 2\left(\frac{k}{2}-j\right)h}\right] - \mu_k(x) \to 0.$$

Further, as $h \to 0$, the kth central difference approximates $\theta^{(k)}(0)$. We conclude that $\theta^{(k)}(0) = 0$.

Now we differentiate eq. (54) k times with respect to h and evaluate at $h = 0$ to obtain

$$\partial_h^k D_k(x, h) = 2^k (k!) \mu_k(x).$$

Eq. (13) follows immediately. □

E Proof of Proposition 1

Proof of proposition 1. Define the vector $\boldsymbol{b}(x)$ of Bernstein polynomials of order p_x:

$$\boldsymbol{b}(x) = (b_{0,p_x}(x), \ldots, b_{p_x,p_x}(x))'.$$

The objective function $\hat{Q}(\boldsymbol{\gamma})$ may be represented as

$$\begin{aligned}\hat{Q}(\boldsymbol{\gamma}) &= \int \left[\tilde{\boldsymbol{\mu}}_K(x) - \boldsymbol{M}_{\Psi,K}\boldsymbol{\gamma}\boldsymbol{b}(x)\right]' \boldsymbol{\Omega} \left[\tilde{\boldsymbol{\mu}}_K(x) - \boldsymbol{M}_{\Psi,K}\boldsymbol{\gamma}\boldsymbol{b}(x)\right] \pi(dx) \\ &= \int \left[\boldsymbol{b}(x)'\boldsymbol{\gamma}'\boldsymbol{M}'_{\Psi,K}\boldsymbol{\Omega}\boldsymbol{M}_{\Psi,K}\boldsymbol{\gamma}\boldsymbol{b}(x) - 2\tilde{\boldsymbol{\mu}}_K(x)\boldsymbol{\Omega}\boldsymbol{M}_{\Psi,K}\boldsymbol{\gamma}\boldsymbol{b}(x)\right] \pi(dx) + c, \end{aligned} \tag{55}$$

where the vector $\tilde{\boldsymbol{\mu}}_K(x)$ is defined in eq. (37) and the c term does not depend on $\boldsymbol{\Gamma}$.

We first show that

$$\hat{Q}(\boldsymbol{\gamma}) = \boldsymbol{g}'\boldsymbol{H}\boldsymbol{g} - 2\boldsymbol{h}'\boldsymbol{\gamma} + c. \tag{56}$$

We consider the quadratic and the linear terms in eq. (55) separately. Define the $p_v(p_x+1)$-vector $\boldsymbol{g}$ as

$$\boldsymbol{g} = \mathrm{vec}(C^{1/2}\boldsymbol{\gamma}') \tag{57}$$

and note that $\boldsymbol{\gamma}$ can be obtained from $\boldsymbol{g}$ as

$$\boldsymbol{\gamma} = \left(\boldsymbol{I}_{p_v} \otimes \boldsymbol{g}'\right)\left(\mathrm{vec}(\boldsymbol{I}_{p_v}) \otimes \boldsymbol{I}_{p_x+1}\right) C^{-1/2}. \tag{58}$$

First we tackle the quadratic term. Let $\boldsymbol{A} = \boldsymbol{M}'_{\Psi,K}\boldsymbol{\Omega}\boldsymbol{M}_{\Psi,K}$, $\boldsymbol{D} = \boldsymbol{\gamma}'\boldsymbol{A}\boldsymbol{\gamma}$. Let d_{ij} be the

(i,j)th element of $\boldsymbol{D}$; c_{ij} be the (i,j)th element of $\boldsymbol{C}$. Then

$$\int \boldsymbol{b}(x)\boldsymbol{D}\boldsymbol{b}(x)\pi(dx) = \sum_{i=1}^{p_x+1}\sum_{i=1}^{p_x+1} d_{ij}\int b_{i-1,p_x}(x)b_{j-1,p_x}(x)\pi(dx) = \sum_{i=1}^{p_x+1}\sum_{i=1}^{p_x+1} d_{ij}c_{ij}. \quad (59)$$

We further expand d_{ij} as follows. Let $\boldsymbol{\gamma}_j$ be the jth column of $\boldsymbol{\gamma}'$, that is, $\boldsymbol{\gamma}' = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{p_v})$. Also, let a_{ij} be the (i,j)th element of $\boldsymbol{A}$. Then $\boldsymbol{\gamma}'\boldsymbol{A}\boldsymbol{\gamma} = \sum_{k=1}^{p_v}\sum_{l=1}^{p_v} a_{kl}\boldsymbol{\gamma}_l\boldsymbol{\gamma}_k'$, and so

$$d_{ij} = \sum_{k=1}^{p_v}\sum_{l=1}^{p_v} a_{kl}\gamma_{l,i-1}\gamma_{k,j-1}, \quad i,j = 1,2\ldots,p_x+1.$$

Then

$$\begin{aligned}\sum_{i=1}^{p_x+1}\sum_{i=1}^{p_x+1} d_{ij}c_{ij} &= \sum_{i=1}^{p_x+1}\sum_{i=1}^{p_x+1}\left(\sum_{k=1}^{p_v}\sum_{l=1}^{p_v} a_{kl}\gamma_{l,i-1}\gamma_{k,j-1}\right)c_{ij} = \sum_{k=1}^{p_v}\sum_{l=1}^{p_v} a_{kl}\sum_{i=1}^{p_x+1}\sum_{i=1}^{p_x+1}\gamma_{l,i-1}\gamma_{k,j-1}c_{ij} \\ &= \sum_{k=1}^{p_v}\sum_{l=1}^{p_v} a_{kl}\boldsymbol{\gamma}_l'\boldsymbol{C}\boldsymbol{\gamma}_k = \sum_{k=1}^{p_v}\sum_{l=1}^{p_v} a_{kl}\tilde{\boldsymbol{\gamma}}_l'\tilde{\boldsymbol{\gamma}}_k = \mathrm{vec}(\tilde{\boldsymbol{\gamma}}')'(\boldsymbol{A}\otimes\boldsymbol{I}_{p_x+1})\,\mathrm{vec}(\tilde{\boldsymbol{\gamma}}') \quad (60)\end{aligned}$$

where $\tilde{\boldsymbol{\gamma}}_k = C^{1/2}\boldsymbol{\gamma}_k$ and $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma}C^{1/2}$.

From eqs. (39), (57), (59), and (60) it now follows that

$$\int \left[\boldsymbol{b}(x)'\boldsymbol{\gamma}'\boldsymbol{M}_{\Psi,K}'\boldsymbol{\Omega}\boldsymbol{M}_{\Psi,K}\boldsymbol{\gamma}\boldsymbol{b}(x)\right]\pi(dx) = \boldsymbol{g}'\boldsymbol{H}\boldsymbol{g}. \quad (61)$$

Similar analysis applies to the linear term in eq. (55). It holds that

$$\begin{aligned}\int \boldsymbol{b}'(x)\boldsymbol{\gamma}'\boldsymbol{M}_{\Psi,K}'\boldsymbol{\Omega}\tilde{\boldsymbol{\mu}}_K(x)\pi(dx) &= \int \mathrm{vec}\left(\boldsymbol{b}'(x)\boldsymbol{\Gamma}'\boldsymbol{M}_{\Psi,K}'\boldsymbol{\Omega}\tilde{\boldsymbol{\mu}}_K(x)\right)\pi(dx) \\ &= \int \mathrm{vec}\left(\boldsymbol{b}'(x)\boldsymbol{C}^{-1/2}\tilde{\boldsymbol{\gamma}}'\boldsymbol{M}_{\Psi,K}'\boldsymbol{\Omega}\tilde{\boldsymbol{\mu}}_K(x)\right)\pi(dx) \\ &= \left[\int\left((\tilde{\boldsymbol{\mu}}_K(x)'\boldsymbol{\Omega}\boldsymbol{M}_{\Psi,K})\otimes(\boldsymbol{b}'(x)C^{-1/2})\right)\pi(dx)\right]\mathrm{vec}(\tilde{\boldsymbol{\gamma}}') \quad (62)\end{aligned}$$

The $(p_x+1)(j-1)+i$th element ($j = 1,\ldots,p_v$, $i = 1,\ldots,p_x+1$) of the leading matrix takes form

$$\sum_{k=0}^{K-1}\sum_{l=0}^{p_x}(\boldsymbol{C}^{-1/2})_{l+1,i}\left(\boldsymbol{\Omega}\boldsymbol{M}_{\Psi,K}\right)_{k+1,j}\int \tilde{\mu}_k(x)b_{l,p_x}(x)\pi(dx).$$

It follows from eqs. (40), (57), (62) that

$$\int \boldsymbol{b}'(x)\boldsymbol{\gamma}'\boldsymbol{M}_{\Psi,K}'\boldsymbol{\Omega}\tilde{\boldsymbol{\mu}}_K(x)\pi(dx) = \boldsymbol{h}'\boldsymbol{g} \quad (63)$$

Eq. (56) now follows from eqs. (55), (61), and (63).

Now consider the constraints of the problem (38). We restate the constraints on $\boldsymbol{\gamma}$ in terms of equivalent constraints on $\boldsymbol{g}$.

First we focus on equality constraints. Each column of $\boldsymbol{\gamma}$ need to sum to 1. To represent this constraint in matrix form, let $\imath_k$ be a k-dimensional vector of all ones. Then the original constraint is equivalent to

$$\imath'_{p_v}\boldsymbol{\gamma} = \imath'_{p_x+1}$$

Postmultiply by $C^{1/2}$ and transpose to obtain $\tilde{\boldsymbol{\gamma}}'\imath_J = C^{1/2}\imath_{p_x+1}$. Vectorizing this, we obtain the new constraint in terms of $\boldsymbol{g}$

$$\left(\imath'_{p_v} \otimes \boldsymbol{I}_{p_x+1}\right) \operatorname{vec}(\tilde{\boldsymbol{\Gamma}}') = \operatorname{vec}(C^{1/2}\imath_{p_x+1}) = \left(\mathbb{I}_1 \otimes C^{1/2}\right) \operatorname{vec}(\imath_{p_x+1}) = C^{1/2}\imath_{p_x+1}$$

Note that $\operatorname{vec}(C^{1/2}\imath_{p_x+1}) = \left(\mathbb{I}_1 \otimes C^{1/2}\right) \operatorname{vec}(\imath_{p_x+1}) = C^{1/2}\imath_{p_x+1}$. Then by eq. (57) the equality requirement on $\boldsymbol{\gamma}$ imply that

$$\left(\imath'_{p_v} \otimes \boldsymbol{I}_{p_x+1}\right) \boldsymbol{g} = \operatorname{vec}(C^{1/2}\imath_{p_x+1}) = C^{1/2}\imath_{p_x+1} \tag{64}$$

Proceeding in reverse order, we obtain that the above constraint on $\boldsymbol{g}$ implies the corresponding equality constraints on $\boldsymbol{\gamma}$.

Now consider the inequality constraints on $\boldsymbol{\gamma}$. Each $\gamma_{l,j}$ is required to be non-negative. As above, let $\boldsymbol{\gamma}_k$ be the kth column of $\boldsymbol{\gamma}'$. The non-negativity requirement can be written as the vector inequality $0 \leq \boldsymbol{\gamma}_k$, where the inequality applies pointwise. Observe that $\boldsymbol{\gamma}_k$ may be represented as $\boldsymbol{\gamma}_k = C^{-1/2}C^{1/2}\boldsymbol{\gamma}_k = C^{-1/2}\tilde{\boldsymbol{\gamma}}_k$, yielding equivalent inequalities

$$0 \leq C^{-1/2}\tilde{\boldsymbol{\gamma}}_k, \quad k = 1, \ldots, p_v.$$

Stacking these inequalities on top of each other across k, we obtain the following non-negativity constraints

$$0 \leq (\boldsymbol{I}_k \otimes C^{-1/2}) \operatorname{vec}(\tilde{\boldsymbol{\Gamma}}') \tag{65}$$

By eq. (57), the above is identical to

$$0 \leq (\boldsymbol{I}_k \otimes C^{-1/2})\boldsymbol{g}'.$$

Finally, to handle the penalty, note that

$$\sum_{j=1}^{p_v}\sum_{l=0}^{p_x} \gamma_{j,l}^2 = \operatorname{tr}(\boldsymbol{\gamma}\boldsymbol{\gamma}'). \tag{66}$$

Using eq. (58), we obtain that

$$\begin{aligned}\operatorname{tr}(\boldsymbol{\gamma}\boldsymbol{\gamma}') &= \operatorname{tr}\left((\boldsymbol{I}_{p_v}\otimes\boldsymbol{g}')\left(\operatorname{vec}(\boldsymbol{I}_{p_v})\otimes\boldsymbol{I}_{p_x+1}\right)C^{-1}\left(\operatorname{vec}(\boldsymbol{I}_{p_v})'\otimes\boldsymbol{I}_{p_x+1}\right)(\boldsymbol{I}_{p_v}\otimes\boldsymbol{g})\right) \\ &= \operatorname{tr}\left(\boldsymbol{W}(\boldsymbol{I}_{p_v}\otimes\boldsymbol{g})(\boldsymbol{I}_{p_v}\otimes\boldsymbol{g}')\right) = \operatorname{tr}\left(\boldsymbol{W}(\boldsymbol{I}_{p_v}\otimes\boldsymbol{g}\boldsymbol{g}')\right)\end{aligned} \tag{67}$$

where $\boldsymbol{W} = \left(\operatorname{vec}(\boldsymbol{I}_{p_v})\otimes\boldsymbol{I}_{p_x+1}\right)C^{-1}\left(\operatorname{vec}(\boldsymbol{I}_{p_v})'\otimes\boldsymbol{I}_{p_x+1}\right)$. To evaluate the above trace, let $\boldsymbol{Z} = \mathbb{I}_{p_v}\otimes(\boldsymbol{g}\boldsymbol{g}')$. Fix $l\in\{1,2,\ldots,p_v^2(p_x+1)\}$. Let k and i be such that $l = p_v(p_x+1)(k-1)+i$ for $1\le i\le p_v(p_x+1)$, $k=1,\ldots,p_v$. The (l,l)th element of $\boldsymbol{W}\boldsymbol{Z}$ is given by

$$\begin{aligned}\sum_{j=1}^{p_v^2(p_x+1)} w_{lj}z_{jl} &= \sum_{j=p_v(p_x+1)(k-1)+1}^{p_v(p_x+1)k} w_{lj}z_{jl} = \sum_{j=1}^{p_v(p_x+1)} w_{l,p_v(p_x+1)(k-1)+j}g_jg_i \\ &= \sum_{j=1}^{p_v(p_x+1)} w_{p_v(p_x+1)(k-1)+i,p_v(p_x+1)(k-1)+j}g_jg_i\end{aligned}$$

Thus,

$$\begin{aligned}\operatorname{tr}(\boldsymbol{W}\boldsymbol{Z}) &= \sum_{i=1}^{p_v(p_x+1)}\sum_{j=1}^{p_v(p_x+1)} w_{p_v(p_x+1)(k-1)+i,p_v(p_x+1)(k-1)+j}g_jg_i \\ &= \sum_{i=1}^{p_v(p_x+1)}\sum_{j=1}^{p_v(p_x+1)} g_jg_i\left[\sum_{k=1}^{p_v} w_{p_v(p_x+1)(k-1)+i,p_v(p_x+1)(k-1)+j}\right]\end{aligned} \tag{68}$$

Define $\boldsymbol{V}$ be the $p_v(p_x+1)\times p_v(p_x+1)$ matrix with the (i,j) element given by $(\boldsymbol{V})_{ij} = \left[\sum_{k=1}^{p_v} w_{p_v(p_x+1)(k-1)+i,p_v(p_x+1)(k-1)+j}\right]$. We conclude by eqs. (66), (67), and (68) that

$$\sum_{j=1}^{p_v}\sum_{l=0}^{p_x}\gamma_{j,l}^2 = \boldsymbol{g}'\boldsymbol{V}\boldsymbol{g}. \tag{69}$$

The assertion of the proposition now follows from eqs. (57), (58), (56), and (69), and the discussion related to eqs. (64) and (65). □

F Proof of Theorem 4.1

Proof of theorem 4.1. For clarity, we break the proof down into 5 steps:

(I) We first introduce a convenient parametrization for conditional moments of (Y_{i1},Y_{i2}) given (X_{i1},X_{i2}) that matches the approach to implementation proposed in the Supplementary Appendix.

(II) We obtain convergence rates for the estimators of conditional moments of (Y_{i1},Y_{i2}) given (X_{i1},X_{i2}) and of their derivatives.

(III) We transfer these first-step convergence rates to the estimators of $\nu_{u_t^k}$, ν_{m^k} and of their derivatives.

(IV) The rates for $\nu_{u_t^k}$ are used to obtain the convergence rates of estimators of $\nu_{(u_2-u_1)^p}$ and of its derivatives.

(V) Finally, the second and the fourth steps are leveraged to obtain convergence rates for the estimators of D_p and their derivatives. From these rates we deduce the convergence rate of $\hat{\mu}_k$.

I We begin by introducing a convenient parametrization for the conditional expectations of function of Y given X, mirroring the approach to implementation described in in the Supplementary Appendix.

(1) Define the variables $W_{i1}^{(\Delta)} = (X_{i2} + X_{i1})/2$ and $W_{i2}^{(\Delta)} = (X_{i2} - X_{i1})/2$. Let the class $\mathcal{G}_k$ be defined as in eq. (52). For any function $g(y_1, y_2) \in \mathcal{G}_k$ let

$$R_g^{(\Delta)}(w_1, w_2) = \mathbb{E}\left[g(Y_{i1}, Y_{i2}) \middle| W_{i1}^{(\Delta)} = w_1, W_{i2}^{(\Delta)} = w_2\right].$$

Observe that $B_{x-h,2h} \equiv \{X_{i1} = x - h, X_{i2} = x + h\} = \{W_{i1}^{(\Delta)} = x, W_{i2}^{(\Delta)} = h\}$. Correspondingly, it holds that $r_g(x - h, x + h) = R_g^{(\Delta)}(x, h)$, where r_g is defined in eq. (6). It follows that $\partial_h^l r_g(x - h, x + h)$ with respect to h at $h = 0$ can be obtained as $\partial_{w_2}^l R_g^{(\Delta)}(w_1, w_2)$ at $(w_1, w_2) = (x, 0)$.

(2) Define $W_{i1}^{(-)} = (X_{i2} - X_{i1})/2$ and $W_{i,2}^{(-)} = -(X_{i1} + X_{i2})/2$. For a function $g(y_1, y_2)$ define

$$R_g^{(-)}(w_1, w_2) = \mathbb{E}\left[g(Y_1, Y_2) \middle| W_{i1}^{(-)} = w_1, W_{i2}^{(-)} = w_2\right].$$

Then $r_g(x - h, x - h) = R_g^{(-)}(0, -x + h)$. As above, it follows that $\partial_h^l r_g(x - h, x - h)$ at $h = 0$ is equal to $\partial_{w_2}^l R_g^{(-)}(w_1, w_2)$ at $(w_1, w_2) = (0, -x)$.

(3) Define $W_{i1}^{(+)} = (X_{i2} - X_{i1})/2$ and $W_{i2}^{(+)} = (X_{i1} + X_{i2})/2$. For a function $g(y_1, y_2)$ set

$$R_g^{(+)}(w_1, w_2) = \mathbb{E}\left[g(Y_1, Y_2) \middle| W_{i1}^{(+)} = w_1, W_{i2}^{(+)} = w_2\right].$$

Then $r_g(x + h, x + h) = R_g^{(+)}(0, x + h)$, and so $\partial_h^l r_g(x + h, x + h)$ at $h = 0$ is equal to $\partial_{w_2}^l R_g^{(+)}(w_1, w_2)$ at $(w_1, w_2) = (0, x)$.

II As the first building block for the rates of the moment estimators, we now consider the estimators of the different R functions. Let $\widehat{\partial_{w_2}^l R_g^{(\Delta)}(w_1, w_2)}$ is the local polynomial LP(q) estimator of $R_g^{(\Delta)}(w_1, w_2)$; analogously for $\widehat{\partial_{w_2}^l R_g^{(-)}(w_1, w_2)}$ and $\widehat{\partial_{w_2}^l R_g^{(+)}(w_1, w_2)}$. To obtain the convergence rates for these estimators, we verify the conditions of theorem 6 of Masry (1996a):

- Consider condition (1a). First, the kernel automatically has $2q$ finite moments, as Ψ_{LP} has bounded support under assumption 4.1. Second, $f_{\boldsymbol{X}}(\boldsymbol{x})$ is bounded uniformly in $\boldsymbol{x}$ under assumption D.2. Third, under assumption D.2 $f_{\boldsymbol{X}}$ has bounded first derivatives, and thus $f_{\boldsymbol{X}}$ is automatically uniformly continuous.
- Condition (2b) follows immediately from (1a).
- Condition (3) holds by assumption 4.1 on the kernel Ψ_{LP}.
- Let $g \in \mathcal{G}_k$. Condition (5d) holds, as by lemma D.2 all $(q+1)$th derivatives of $r_g(\boldsymbol{x})$ are bounded uniformly in $\boldsymbol{x}$.
- Fix $g \in \mathcal{G}_k$. By lemma D.2 all $(q+2)$th derivatives of $r_g(\boldsymbol{x})$ exist and are uniformly bounded over $\boldsymbol{x}$. It then readily follows each $(q+1)$th derivative of $r_g(\boldsymbol{x})$ is Lipschitz in $\boldsymbol{x}$, establishing condition (6).
- Condition (4b) is automatically implied by conditions (5d) and (6).
- We turn to condition (7). First consider condition (7a). Let $\delta' = \delta/k > 0$ for δ of the assumptions of the theorem. For any $g \in \mathcal{G}_k$ it holds that $\sup_{\boldsymbol{x}\in J} \mathbb{E}\left[|g(Y_{i1}, Y_{i2})|^{2+\delta'} | \boldsymbol{X}_i = \boldsymbol{x}\right] < \infty$. We show this for $g(y_1, y_2) = (y_2 - y_1)^j, j \leq k$, the argument for other functions is analogous:

$$
\begin{aligned}
&\sup_{\boldsymbol{x}} \mathbb{E}\left[|(Y_{i2} - Y_{i1})|^{2+\delta'} | \boldsymbol{X}_i = \boldsymbol{x}\right] \\
&\leq \sup_{\boldsymbol{x}} 2^{2j+\delta j/k}\, \mathbb{E}\left[\left||Y_{i1}|^j + |Y_{i2}|^j\right|^{2+\delta'} | \boldsymbol{X}_i = \boldsymbol{x}\right] \\
&\leq 2^{2j+\delta j/k}\left[2 \sup_{x,a} |m(x,a)|^{2j+\delta j/k} + \sum_{t=1}^{2} \sup_{\boldsymbol{x}} \mathbb{E}\left[|u_{it}|^{2j+\delta j/k} | \boldsymbol{X}_i = \boldsymbol{x}\right]\right] \\
&< \infty,
\end{aligned}
\tag{70}
$$

  where the last inequality follows from the assumption D.1 and the assumption that $\sup_{\boldsymbol{x}} \mathbb{E}\left[|u_{it}|^{2k+\delta} | \boldsymbol{X}_i = \boldsymbol{x}\right] < \infty$ combined with the observation that $2j + \delta j/k \leq 2k + \delta$. Conditions (7b)-(7c) follow from lemma D.1. Finally, (7d) holds as $\{\boldsymbol{Y}_i, \boldsymbol{X}_i\}$ is iid across i.
- In condition (8), take D equal to J for J of assumption 2.3. By assumption 2.3, $f_{\boldsymbol{X}}(\boldsymbol{x})$ is bounded away from 0 on J.
- Condition (4.5) on s holds under the assumptions of the theorem with $\sigma = \nu = \infty$, as $\{\boldsymbol{Y}_i, \boldsymbol{X}_i\}$ is iid across i.

It follows from theorem 6 of Masry (1996a) that for any $l \in \{0, \ldots, k\}$ and any $g \in \mathcal{G}_k$

$$
\begin{aligned}
\sup_{x\in I} \left| \widehat{\partial^l_{w_2} R_g^{(\Delta)}(x,0)} - \partial^l_{w_2} R_g^{(\Delta)}(x,0) \right| &= O_{a.s.}(\delta_{l,N}), \\
\sup_{x\in I} \left| \widehat{\partial^l_{w_2} R_g^{(-)}(0,-x)} - \partial^l_{w_2} R_g^{(-)}(0,-x) \right| &= O_{a.s.}(\delta_{l,N}),
\end{aligned}
$$

$$\sup_{x\in I}\left|\widehat{\partial^l_{w_2} R^{(+)}_g(0,x)} - \partial^l_{w_2} R^{(+)}_g(0,x)\right| = O_{a.s.}(\delta_{l,N}) \tag{71}$$

where

$$\delta_{l,N} = \sqrt{\frac{\log(N)}{N s^{2+2l}}} + s^{q-l+1}.$$

III We now establish convergence rates for $\nu_{u_t^k}$, ν_{m^k} and their derivatives. Note that $\nu_{u_t^k}$ and ν_{m^k} are differentiable at least k times by lemma D.3. First, we rewrite the expressions of lemma 2.1 in terms of above notation for regression functions. If $k = 1$

$$\begin{aligned}
&\partial^l_h \nu_{u_t}(x-h)|_{h=0} = 0, \quad t = 1,2,\\
&\partial^l_h \nu_m(x \pm h)|_{h=0} = \partial^l_{w_2} R^{(\pm)}_{y_1 y_2}(0, \pm x).
\end{aligned} \tag{72}$$

If $k = 2$, we instead have

$$\begin{aligned}
&\partial^l_h \nu_{u_1^k}(x-h)|_{h=0}\\
&= \partial^l_{w_2} R^{(-)}_{y_1^{k-1}(y_1-y_2)}(0,-x)\\
&\quad - \sum_{j=1}^{k-1}\binom{k-1}{j}\left[\sum_{i=0}^{l}\binom{l}{i}\left(\partial^i_h \nu_{m^j}(x-h)|_{h=0}\right)\left(\partial^{l-i}_h \nu_{u_1^{k-j}}(x-h)|_{h=0}\right)\right],\\
&\partial^l_h \nu_{u_2^k}(x+h)|_{h=0}\\
&= \partial^l_{w_2} R^{(+)}_{y_2^{k-1}(y_2-y_1)}(0,x)\\
&\quad - \sum_{j=1}^{k-1}\binom{k-1}{j}\left[\sum_{i=0}^{l}\binom{l}{i}\left(\partial^i_h \nu_{m^j}(x+h)|_{h=0}\right)\left(\partial^{l-i}_h \nu_{u_2^{k-j}}(x+h)|_{h=0}\right)\right],\\
&\partial^l_h \nu_{m^k}(x \pm h)|_{h=0}\\
&= \partial^l_{w_2} R^{(\pm)}_{y_1^{k-1} y_2}(0,\pm x)\\
&- \sum_{j=1}^{k-1}\binom{k-1}{j-1}\left[\sum_{i=0}^{l}\binom{l}{i}\left(\partial^i_h \nu_{m^j}(x\pm h)|_{h=0}\right)\left(\partial^{l-i}_h \nu_{u_t^{k-j}}(x\pm h)|_{h=0}\right)\right],
\end{aligned}$$

where in the last line we take t equal to 1 if the argument of $\nu_{u_t^{(k-j)}}$ is $x-h$, and $t = 2$ otherwise.

Let $\widehat{\partial^l_h \nu_{u_1^k}(x-h)|_{h=0}}$ be the estimator for $\partial^l_h \nu_{u_1^k}(x-h)|_{h=0}$, formed as the sample analog of the above equations (see algorithm 1). Likewise, let $\widehat{\partial^l_h \nu_{m^k}(x\pm h)|_{h=0}}$ be the estimator for $\partial^l_h \nu_{m^k}(x\pm h)|_{h=0}$.

We assert that for each $p = \{0, \dots, k\}$ it holds for all $l \in \{0, \dots, k\}$ that

$$\begin{aligned}
\sup_{x\in I}\left|\widehat{\partial_h^l \nu_{u_1^p}(x-h)|_{h=0}} - \partial_h^l \nu_{u_1^p}(x-h)|_{h=0}\right| &= O_{a.s.}\left(\delta_{l,N}\right) \\
\sup_{x\in I}\left|\widehat{\partial_h^l \nu_{u_2^p}(x+h)|_{h=0}} - \partial_h^l \nu_{u_2^p}(x+h)|_{h=0}\right| &= O_{a.s.}\left(\delta_{l,N}\right) \\
\sup_{x\in I}\left|\widehat{\partial_h^l \nu_{m^p}(x\pm h)|_{h=0}} - \partial_h^l \nu_{m^p}(x\pm h)|_{h=0}\right| &= O_{a.s.}\left(\delta_{l,N}\right).
\end{aligned} \tag{73}$$

We prove the above assertion by finite induction on p. First, the results for $p = 0, 1$ follows immediately from eqs. (71) and (72). Suppose that (73) holds for moments of order up to $p-1$. Then consider the pth moments. We only consider $\widehat{\partial_h^l \nu_{u_1^p}(x-h)|_{h=0}}$ explicitly, the argument is completely analogous for the other estimators.

The estimation error can be bounded as

$$\begin{aligned}
&\sup_{x\in I}\left|\widehat{\partial_h^l \nu_{u_1^p}(x-h)|_{h=0}} - \partial_h^l \nu_{u_1^p}(x-h)|_{h=0}\right| \\
&\leq \sup_{x\in I}\left|\widehat{\partial_{w_2}^l R^{(-)}_{y_1^{p-1}(y_1-y_2)}(0,-x)} - \partial_{w_2}^l R^{(-)}_{y_1^{p-1}(y_1-y_2)}(0,-x)\right| \\
&\quad + \sum_{j=1}^{p-1}\binom{p-1}{j}\sum_{i=0}^{l}\binom{l}{i}\sup_{x\in I}\Bigg|\left(\widehat{\partial_h^i \nu_{m^j}(x-h)|_{h=0}}\right)\left(\widehat{\partial_h^{l-i} \nu_{u_1^{p-j}}(x-h)|_{h=0}}\right) \\
&\qquad\qquad - \left(\partial_h^i \nu_{m^j}(x-h)|_{h=0}\right)\left(\partial_h^{l-i} \nu_{u_1^{p-j}}(x-h)|_{h=0}\right)\Bigg|.
\end{aligned} \tag{74}$$

The term under the sum can further be upper bounded as

$$\begin{aligned}
&\sup_{x\in I}\Bigg|\left(\widehat{\partial_h^i \nu_{m^j}(x-h)|_{h=0}}\right)\left(\widehat{\partial_h^{l-i} \nu_{u_1^{p-j}}(x-h)|_{h=0}}\right) \\
&\qquad - \left(\partial_h^i \nu_{m^j}(x-h)|_{h=0}\right)\left(\partial_h^{l-i} \nu_{u_1^{p-j}}(x-h)|_{h=0}\right)\Bigg| \\
&\leq \sup_{x\in I}\left|\partial_h^{l-i} \nu_{u_1^{p-j}}(x-h)|_{h=0}\right| \sup_{x\in I}\left|\widehat{\partial_h^i \nu_{m^j}(x-h)|_{h=0}} - \partial_h^i \nu_{m^j}(x-h)|_{h=0}\right| \\
&\quad + \sup_{x\in I}\left|\widehat{\partial_h^i \nu_{m^j}(x-h)|_{h=0}}\right| \sup_{x\in I}\left|\widehat{\partial_h^{l-i} \nu_{u_1^{p-j}}(x-h)|_{h=0}} - \partial_h^{l-i} \nu_{u_1^{p-j}}(x-h)|_{h=0}\right| \\
&= O(1) O_{a.s.}\left(\delta_{i,N}\right) + O_{a.s.}(1) O_{a.s.}\left(\delta_{l-i,N}\right),
\end{aligned} \tag{75}$$

where we use the inductive assumption and lemma D.3. Observe that $\delta_{j,N} = o(\delta_{j+1,N})$ under the assumptions of the theorem. We conclude that (73) holds for $\widehat{\partial_h^l \nu_{u_1^p}(x-h)|_{h=0}}$ by combining together eqs. (71), (74), and (75). The argument is analogous for moments of u_2 and m.

IV Now we establish convergence rates of the estimator $\nu_{(u_2-u_1)^p}(x,h)$ for $p \in \{0,1,\dots,k\}$. First, note that the lth derivative of $\nu_{(u_2-u_1)^p}(x,h)$ with respect to h is given by

$$\partial_h^l \nu_{(u_2-u_1)^p}(x,h) = \sum_{j=0}^{p} \binom{p}{j} \left[\sum_{i=0}^{l} \binom{l}{i} \left(\partial_h^i \nu_{u_1^j}(x-h) \right) \left(\partial_h^{l-i} \nu_{u_2^{p-j}}(x+h) \right) \right].$$

The estimator $\widehat{\partial_h^l \nu_{(u_2-u_1)^p}(x,0)}$ replaces all population objects on the right hand side above with their sample versions, in line with the algorithm 1. Correspondingly, the estimation error error at $h=0$ can be bounded as

$$\begin{aligned}
&\sup_{x\in I} \left| \widehat{\partial_h^l \nu_{(u_2-u_1)^p}(x,0)} - \partial_h^l \nu_{(u_2-u_1)^p}(x,0) \right| \\
&\leq \sum_{j=0}^{p} \binom{p}{j} \sum_{i=0}^{l} \binom{l}{i} \sup_{x\in I} \Bigg| \left(\widehat{\partial_h^i \nu_{u_1^j}(x-h)|_{h=0}} \right) \left(\widehat{\partial_h^{l-i} \nu_{u_2^{p-j}}(x+h)|_{h=0}} \right) \\
&\qquad\qquad - \left(\partial_h^i \nu_{u_1^j}(x-h)|_{h=0} \right) \left(\partial_h^{l-i} \nu_{u_2^{p-j}}(x+h)|_{h=0} \right) \Bigg|. \qquad (76)
\end{aligned}$$

Each term under the sums can be further bounded as

$$\begin{aligned}
&\sup_{x\in I} \Bigg| \left(\widehat{\partial_h^i \nu_{u_1^j}(x-h)|_{h=0}} \right) \left(\widehat{\partial_h^{l-i} \nu_{u_2^{p-j}}(x+h)|_{h=0}} \right) \\
&\qquad\qquad - \left(\partial_h^i \nu_{u_1^j}(x-h)|_{h=0} \right) \left(\partial_h^{l-i} \nu_{u_2^{p-j}}(x+h)|_{h=0} \right) \Bigg| \\
&\leq \sup_{x\in I} \left| \partial_h^i \nu_{u_1^j}(x-h)|_{h=0} \right| \sup_{x\in I} \left| \widehat{\partial_h^{l-i} \nu_{u_2^{p-j}}(x+h)|_{h=0}} - \partial_h^{l-i} \nu_{u_2^{p-j}}(x+h)|_{h=0} \right| \\
&\quad + \sup_{x\in I} \left| \widehat{\partial_h^{l-i} \nu_{u_2^{p-j}}(x+h)|_{h=0}} \right| \sup_{x\in I} \left| \widehat{\partial_h^i \nu_{u_1^j}(x-h)|_{h=0}} - \partial_h^i \nu_{u_1^j}(x-h)|_{h=0} \right| \\
&= O(1) O_{a.s.}(\delta_{l-i,N}) + O_{a.s.}(1) O_{a.s.}(\delta_{i,N}). \qquad (77)
\end{aligned}$$

It now follows from $\delta_{j,N} = o(\delta_{j+1,N})$ and eqs. (73), (76), and (77) that

$$\sup_{x\in I} \left| \widehat{\partial_h^l \nu_{(u_2-u_1)^p}(x,0)} - \partial_h^l \nu_{(u_2-u_1)^p}(x,0) \right| = O_{a.s.}(\delta_{l,N}). \qquad (78)$$

V Recall that $D_p(x,h) = \mathbb{E}\left[(m(x_0+h,\alpha_i) - m(x_0-h,\alpha_i))^p | B_{x_0-h,2h} \right]$ (see eq. (47)). Using eq. (14) and the above notation, we can write the lth derivative of D_p as

$$\partial_h^l D_p(x,h) = \partial_{w_2}^l R^{(\Delta)}_{(y_2-y_1)^p}(x,h)$$

$$- \sum_{j=0}^{p-1} \binom{p}{j} \left[\sum_{i=0}^{l} \binom{l}{i} \left(\partial_h^i D_j(x,h) \right) \left(\partial_h^{l-i} \nu_{(u_2-u_1)^{p-j}}(x,h) \right) \right],$$

The estimator $\widehat{\partial_h^l D_p(x,0)}$ replaces all population objects on the right hand side with their sample versions (see algorithm 1).

We assert that for each $p \in \{0,1,\ldots,k\}$ it holds for all $l \in \{0,1,\ldots,k\}$ that

$$\sup_{x \in I} \left| \widehat{\partial_h^l D_p(x,0)} - \partial_h^l D_p(x,0) \right| = O_{a.s.}(\delta_{l,N}) \tag{79}$$

The estimation error can be bounded as

$$\begin{aligned}
&\sup_{x \in I} \left| \widehat{\partial_h^l D_p(x,0)} - \partial_h^l D_p(x,0) \right| \\
&\leq \sup_{x \in I} \left| \widehat{\partial_{w_2}^l R^{(\Delta)}_{(y_2-y_1)^p}(x,0)} - \partial_{w_2}^l R^{(\Delta)}_{(y_2-y_1)^p}(x,0) \right| \\
&\quad + \sum_{j=0}^{p-1} \binom{p}{j} \sum_{i=0}^{l} \binom{l}{i} \sup_{x \in I} \Bigg| \left(\widehat{\partial_h^i D_j(x,0)} \right) \left(\widehat{\partial_h^{l-i} \nu_{(u_2-u_1)^{p-j}}(x,0)} \right) \\
&\qquad\qquad - \left(\partial_h^i D_j(x,0) \right) \left(\partial_h^{l-i} \nu_{(u_2-u_1)^{p-j}}(x,0) \right) \Bigg|
\end{aligned} \tag{80}$$

The term under the sums can be bounded from above as

$$\begin{aligned}
&\sup_{x \in I} \Bigg| \left(\widehat{\partial_h^i D_j(x,0)} \right) \left(\widehat{\partial_h^{l-i} \nu_{(u_2-u_1)^{p-j}}(x,0)} \right) \\
&\qquad\qquad - \left(\partial_h^i D_j(x,0) \right) \left(\partial_h^{l-i} \nu_{(u_2-u_1)^{p-j}}(x,0) \right) \Bigg| \\
&\leq \sup_{x \in I} \left| \widehat{\partial_h^{l-i} \nu_{(u_2-u_1)^{p-j}}(x,0)} \right| \sup_{x \in I} \left| \widehat{\partial_h^i D_j(x,0)} - \partial_h^i D_j(x,0) \right| \\
&\quad + \sup_{x \in I} \left| \partial_h^i D_j(x,0) \right| \sup_{x \in I} \left| \widehat{\partial_h^{l-i} \nu_{(u_2-u_1)^{p-j}}(x,0)} - \partial_h^{l-i} \nu_{(u_2-u_1)^{p-j}}(x,0) \right| \\
&= O_{a.s.}(1) O_{a.s.}(\delta_{i,N}) + O(1) O_{a.s.}(\delta_{l-i,N}).
\end{aligned} \tag{81}$$

(79) now follows from eqs. (71), (80), and (81).

Finally, by theorem 3.1 and by definition of $\hat{\mu}_k(x)$ we have that

$$\mu_k(x) = \frac{1}{2^k k!} \partial_h^k D_k(x,0), \quad \hat{\mu}_k(x) = \frac{1}{2^k k!} \widehat{\partial_h^k D_k(x,0)}.$$

Then

$$\sup_{x\in I}|\hat{\mu}_k(x)-\mu_k(x)| = \frac{1}{2^k k!}\sup_{x\in I}\left|\widehat{\partial_h^k D_k(x,0)} - \partial_h^k D_k(x,0)\right| = O_{a.s.}(\delta_{k,N}),$$

establishing the result of the theorem. □

G Proof of Theorem 4.2

G.1 Supporting Lemmas

In order to establish asymptotic normality of the moment estimator $\hat{\mu}_k(x)$, we first establish that the underlying local polynomial estimators of conditional moments of (Y_{i1}, Y_{i2}) and their derivatives are also asymptotically normally distributed. It is convenient to establish joint asymptotic normality of all the LP(q) estimators involved.

We first introduce some notation. Let $\mathcal{G}_k$ be defined as in eq. (52). Let k and x be as in theorem 4.2. Let the vector $\hat{\boldsymbol{V}}_{\mathcal{G}_k}(x)$ stack all the $\widehat{\partial_{w_2}^k R_g^{(\Delta)}(x,0)}$, $\widehat{\partial_{w_2}^k R_g^{(-)}(0,-x)}$, and $\widehat{\partial_{w_2}^k R_g^{(+)}(0,x)}$ that appear in the proof of theorem 4.1. Let $\boldsymbol{V}_{\mathcal{G}_k}(x)$ be its population counterpart. Formally, number the elements of $\mathcal{G}_k$ as $\{g_1,\ldots,g_{|\mathcal{G}_k|}\}$. Let $\hat{\boldsymbol{V}}_{\mathcal{G}_k}(x)$ be a $3|\mathcal{G}_k|$-vector with the jth element given by $\widehat{\partial_{w_2}^k R_{g_j}^{(\Delta)}(x,0)}$; the $(|\mathcal{G}_k|+j)$th element given by $\widehat{\partial_{w_2}^k R_{g_j}^{(-)}(0,-x)}$; and the $(2|\mathcal{G}_k|+j)$th element given by $\widehat{\partial_{w_2}^k R_g^{(-)}(0,x)}$. Similarly, let $\boldsymbol{V}_{\mathcal{G}_k}(x)$ be a $3|\mathcal{G}_k|$-vector with the jth element given by $\partial_{w_2}^k R_{g_j}^{(\Delta)}(x,0)$; the $(|\mathcal{G}_k|+j)$th element given by $\partial_{w_2}^k R_{g_j}^{(-)}(0,-x)$; and the $(2|\mathcal{G}_k|+j)$th element given by $\partial_{w_2}^k R_g^{(-)}(0,x)$

Lemma G.1. *Let the assumption of theorem 4.2 hold. Let $L<\infty$ be a fixed integer. For $l=1,\ldots,L$ let $x_l\in I$. Then the vector $\sqrt{Ns^{2+2k}}[(\hat{\boldsymbol{V}}_{\mathcal{G}_k}(x_1)-\boldsymbol{V}_{\mathcal{G}_k}(x_1))',\ldots,(\hat{\boldsymbol{V}}_{\mathcal{G}_k}(x_L)-\boldsymbol{V}_{\mathcal{G}_k}(x_L))']'$ converges weakly to a mean-zero normally distributed random vector. Further, if $l\neq j$, then $\sqrt{Ns^{2+2k}}(\hat{\boldsymbol{V}}_{\mathcal{G}_k}(x_l)-\boldsymbol{V}_{\mathcal{G}_k,N}(x_l))$ and $\sqrt{Ns^{2+2k}}(\hat{\boldsymbol{V}}_{\mathcal{G}_k}(\boldsymbol{x}_j)-\boldsymbol{V}_{\mathcal{G}_k}(x_j))$ are asymptotically independent.*

Proof. We establish convergence by theorem 5 of Masry (1996b), whose conditions we now briefly verify. The conditions overlap significantly with the conditions of theorem 6 of Masry (1996a) used in the proof of theorem 4.1, and we refer to that proof for some additional details.

- Conditions (1a) and (2a) follow from assumption 4.1 on the kernel Ψ_{LP}. Condition (1b) follow from assumption D.2. Condition (1c) holds as $\{(\boldsymbol{Y}_i,\boldsymbol{X}_i)\}$ is iid over i.
- Condition (2b) follows from lemma D.1. The moment condition of (2c) holds as in the proof of theorem 4.1: for any $g\in\mathcal{G}_k$ it holds that $\sup_{\boldsymbol{x}\in J}\mathbb{E}\left[|g(Y_{i1},Y_{i2})|^{2+\delta'}|\boldsymbol{X}_i=\boldsymbol{x}\right]<\infty$, where $\delta'=\delta/k>0$ for the δ of the theorem statement (see eq. (70) and the preceding discussion).

- Condition (3) holds by the assumptions of the theorem on s.
- Condition (4) holds by lemma D.1.
- The differentiability requirement on p. 83 for r_g holds as each r_g is differentiable at least $(q+2)$ times by lemma D.2.

Asymptotic normality now follows by theorem theorem 5 of Masry (1996b).[7] The requirement that $Ns^{2k+4} \to 0$ eliminates the bias and ensures that the limit distribution is centered at zero. Independence holds by a standard argument (see e.g. p. 113 in Bierens (1987)). □

We now establish asymptotic normality of the components of the moment estimator. The results are intuitive, as the $\hat{\nu}_{\cdot}$ estimators are built up as differentiable transformations of moments of Y.

Define $\hat{\boldsymbol{V}}_{u_1,p}(x)$ as a $(p+1)$-vector with jth element given by $\widehat{\partial_h^k \nu_{u_1^{j-1}}(x-h)|_{h=0}}$. Let $\boldsymbol{V}_{u_1,p}(x)$ be a $(p+1)$-vector with the jth element given by $\partial_h^k \nu_{u_1^{j-1}}(x-h)|_{h=0}$. Similarly, define $\hat{\boldsymbol{V}}_{u_2,p}(x)$ as a $(p+1)$-vector with the jth element $\widehat{\partial_h^k \nu_{u_2^{j-1}}(x+h)|_{h=0}}$ and $\boldsymbol{V}_{u_2,p}(x)$ as the $(p+1)$-vector with the jth element $\partial_h^k \nu_{u_2^{j-1}}(x+h)|_{h=0}$.

Lemma G.2. *Let the assumptions of theorem 4.2 holds. Then* $\sqrt{Ns^{2+2k}}\big((\hat{\boldsymbol{V}}_{u_1,k}(x)-\boldsymbol{V}_{u_1,k}(x))', (\hat{\boldsymbol{V}}_{u_2,k}(x)-\boldsymbol{V}_{u_2,k}(x))'\big)$ *weakly converges to a mean zero normally distributed random vector. Convergence is joint with the vector* $\sqrt{Ns^{2+2k}}(\hat{\boldsymbol{V}}_{\mathcal{G}_k}(x)-\boldsymbol{V}_{\mathcal{G}_k}(x))$ *of lemma G.1.*

Proof. It will be convenient to prove a slightly stronger assertion. First, let $\hat{\boldsymbol{V}}_{m,p}^{(+)}(x)$ be a $(p+1)$-vector with the jth element given by $\widehat{\partial_h^k \nu_{m^{j-1}}(x+h)|_{h=0}}$, and $\boldsymbol{V}_{m,p}^{(+)}(x)$ be its population equivalent. Similarly, define $\hat{\boldsymbol{V}}_{m,p}^{(-)}(x)$ as the $(p+1)$-vector with the jth element $\widehat{\partial_h^k \nu_{m^{j-1}}(x-h)|_{h=0}}$ and $\boldsymbol{V}_{m,p}^{(-)}(x)$ as its population equivalent.[8] Then we assert that for every $p=0,1,\ldots,k$ the vector $\sqrt{Ns^{2+2k}}\big((\hat{\boldsymbol{V}}_{u_1,p}(x)-\boldsymbol{V}_{u_1,p}(x))', (\hat{\boldsymbol{V}}_{u_2,p}(x)-\boldsymbol{V}_{u_2,p}(x))', (\hat{\boldsymbol{V}}_{m,p}^{(\pm)}(x)-\boldsymbol{V}_{m,p}^{(\pm)}(x))'\big)$ weakly converges to a mean zero normally distributed random vector, jointly with $\sqrt{Ns^{2+2k}}(\hat{\boldsymbol{V}}_{\mathcal{G}_k}(x)-\boldsymbol{V}_{\mathcal{G}_k}(x))$.

We show the assertion by finite induction on p. The assertion is trivial is for $p=0$. Consider $p=1$. Then

$$\sqrt{Ns^{2+2k}}(\hat{\boldsymbol{V}}_{u_1,1}(x)-\boldsymbol{V}_{u_1,1}(x)) = \sqrt{Ns^{2+2k}}(\hat{\boldsymbol{V}}_{u_2,1}(x)-\boldsymbol{V}_{u_2,1}(x)) = 0,$$

[7] Theorem 5 of Masry (1996b) is stated for the case of estimating one conditional moment. However, it extends immediately to establishing joint normality of a fixed finite collection of conditional moments by an application of the Cramér-Wold device.

[8] Note that $\partial_h^k \nu_{m^p}(x-h)|_{h=0} = (-1)^k \partial_h^k \nu_{m^p}(x+h)|_{h=0}$. We can then consider only $\nu_{m^p}(x+h)$ if Ψ_{LP} is a product kernel of the form $\Psi_{LP} = \Psi_{1d}^2$ where Ψ_{1d} is a symmetric kernel, as the local polynomial estimators of the derivatives will satisfy the same property as the population derivatives.

and

$$\sqrt{Ns^{2+2k}}(\hat{\boldsymbol{V}}_{m,1}(x) - \boldsymbol{V}_{m,1}(x)) = \begin{pmatrix} 0 \\ \sqrt{Ns^{2+2k}}\left(\widehat{\partial^l_{w_2} R^{(+)}_{y_1y_2}(0,x)} - \partial^l_{w_2} R^{(+)}_{y_1y_2}(0,x)\right) \end{pmatrix}.$$

The assertion then follows immediately from lemma G.1.

Now suppose that the assertion is true up to $p-1$. Consider $p \leq k$. Then consider the last element of $\hat{\boldsymbol{V}}_{u_1,p}$, corresponding to the kth derivative of $\nu_{u_1^p}$. As in the proof of theorem 4.1, the following representation holds:

$$\begin{aligned}
&\sqrt{Ns^{2+2k}}\left(\widehat{\partial^k_h \nu_{u_1^p}(x-h)|_{h=0}} - \partial^k_h \nu_{u_1^p}(x-h)|_{h=0}\right) \\
&= \sqrt{Ns^{2+2k}}\left(\widehat{\partial^k_{w_2} R^{(-)}_{y_1^{p-1}(y_1-y_2)}(0,-x)} - \partial^k_{w_2} R^{(-)}_{y_1^{p-1}(y_1-y_2)}(0,-x)\right) \\
&\quad + \sum_{j=1}^{p-1}\binom{p-1}{j}\sum_{i=0}^{k}\binom{k}{i}\sqrt{Ns^{2+2k}}\Bigg(\left(\widehat{\partial^i_h \nu_{m^j}(x-h)|_{h=0}}\right)\left(\widehat{\partial^{k-i}_h \nu_{u_1^{p-j}}(x-h)|_{h=0}}\right) \\
&\qquad\qquad - \left(\partial^i_h \nu_{m^j}(x-h)|_{h=0}\right)\left(\partial^{k-i}_h \nu_{u_1^{p-j}}(x-h)|_{h=0}\right)\Bigg) \\
&= \sqrt{Ns^{2+2k}}\left(\widehat{\partial^k_{w_2} R^{(-)}_{y_1^{p-1}(y_1-y_2)}(0,-x)} - \partial^k_{w_2} R^{(-)}_{y_1^{p-1}(y_1-y_2)}(0,-x)\right) \\
&\quad + \sum_{j=1}^{p-1}\binom{p-1}{j}\sum_{i=0}^{k}\binom{k}{i}\sqrt{Ns^{2+2k}}\left(\widehat{\partial^{k-i}_h \nu_{u_1^{p-j}}(x-h)|_{h=0}}\right) \\
&\qquad\qquad \times\left(\widehat{\partial^i_h \nu_{m^j}(x-h)|_{h=0}} - \partial^i_h \nu_{m^j}(x-h)|_{h=0}\right) \\
&\quad + \sum_{j=1}^{p-1}\binom{p-1}{j}\sum_{i=0}^{k}\binom{k}{i}\sqrt{Ns^{2+2k}}\left(\partial^i_h \nu_{m^j}(x-h)|_{h=0}\right) \\
&\qquad\qquad \times\left(\widehat{\partial^{k-i}_h \nu_{u_1^{p-j}}(x-h)|_{h=0}} - \partial^{k-i}_h \nu_{u_1^{p-j}}(x-h)|_{h=0}\right). \qquad (82)
\end{aligned}$$

Consider the (j,i)th term in the first term with $i < k$. By eq. (73) in the proof theorem 4.2 it satisfies

$$\begin{aligned}
&\sqrt{Ns^{2+2k}}\left(\widehat{\partial^{k-i}_h \nu_{u_1^{p-j}}(x-h)|_{h=0}}\right)\left(\widehat{\partial^i_h \nu_{m^j}(x-h)|_{h=0}} - \partial^i_h \nu_{m^j}(x-h)|_{h=0}\right) \\
&= \sqrt{Ns^{2+2k}}\left(\partial^{k-i}_h \nu_{u_1^{p-j}}(x-h)|_{h=0}\right)\left(\widehat{\partial^i_h \nu_{m^j}(x-h)|_{h=0}} - \partial^i_h \nu_{m^j}(x-h)|_{h=0}\right)
\end{aligned}$$

$$
\begin{aligned}
&+\sqrt{Ns^{2+2k}}O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2(k-i)}}}+s^{q-(k-i)+1}\right)O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2i}}}+s^{q-i+1}\right)\\
&=\sqrt{Ns^{2+2k}}O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2i}}}+s^{q-i+1}\right)\\
&\quad+\sqrt{Ns^{2+2k}}O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2(k-i)}}}+s^{q-(k-i)+1}\right)O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2i}}}+s^{q-i+1}\right)\\
&=o_{a.s.}(1),
\end{aligned}
$$

where the last equality holds by the assumption that $Ns^{2q+4}\to 0$ and $s^2\log(N)\to 0$. We conclude that all the terms in the first sum of eq. (82) with $i<k$ are $o_{a.s.}(1)$. Analogously, all the terms in the second sum of eq. (82) with $i>0$ are $o_{a.s.}(1)$.

In light of this, we can write

$$
\begin{aligned}
&\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^k\nu_{u_1^p}(x-h)|_{h=0}}-\partial_h^k\nu_{u_1^p}(x-h)|_{h=0}\right)+o_{a.s.}(1) \qquad (83)\\
&=\sqrt{Ns^{2+2k}}\left(\widehat{\partial_{w_2}^k R^{(-)}_{y_1^{p-1}(y_1-y_2)}(0,-x)}-\partial_{w_2}^k R^{(-)}_{y_1^{p-1}(y_1-y_2)}(0,-x)\right)\\
&\quad+\sum_{j=1}^{p-1}\binom{p-1}{j}\sqrt{Ns^{2+2k}}\left(\nu_{u_1^{p-j}}(x-h)|_{h=0}\right)\left(\widehat{\partial_h^k\nu_{m^j}(x-h)|_{h=0}}-\partial_h^k\nu_{m^j}(x-h)|_{h=0}\right)\\
&\quad+\sum_{j=1}^{p-1}\binom{p-1}{j}\sqrt{Ns^{2+2k}}\left(\nu_{m^j}(x-h)|_{h=0}\right)\left(\widehat{\partial_h^k\nu_{u_1^{p-j}}(x-h)|_{h=0}}-\partial_h^k\nu_{u_1^{p-j}}(x-h)|_{h=0}\right).
\end{aligned}
$$

Similar representations hold for $\widehat{\partial_h^k\nu_{u_2^p}(x+h)|_{h=0}}$ and $\widehat{\partial_h^k\nu_{m^p}(x\pm h)|_{h=0}}$.

We conclude that that $\sqrt{Ns^{2+2k}}\big((\hat{\boldsymbol{V}}_{u_1,p}(x)-\boldsymbol{V}_{u_1,p}(x))',(\hat{\boldsymbol{V}}_{u_2,p}(x)-\boldsymbol{V}_{u_2,p}(x))',(\hat{\boldsymbol{V}}^{(\pm)}_{m,p}(x)-\boldsymbol{V}^{(\pm)}_{m,p}(x))'\big)$ can be written as a smooth transformation of $\sqrt{Ns^{2+2k}}\big((\hat{\boldsymbol{V}}_{u_1,p-1}(x)-\boldsymbol{V}_{u_1,p-1}(x))',(\hat{\boldsymbol{V}}_{u_2,p-1}(x)-\boldsymbol{V}_{u_2,p-1}(x))',(\hat{\boldsymbol{V}}^{(\pm)}_{m,p-1}(x)-\boldsymbol{V}^{(\pm)}_{m,p-1}(x))'\big)$ and $\sqrt{Ns^{2+2k}}(\hat{\boldsymbol{V}}_{\mathcal{G}_k}(x)-\boldsymbol{V}_{\mathcal{G}_k}(x))$, plus an $o_{a.s.}(1)$ component. The assertion then immediately follows by the delta method from the inductive assumption. □

We now transfer the convergence results of lemma G.2 to estimates of the kth derivatives of $\nu_{(u_2-u_1)^j}(x,h)$.

Define $\hat{\boldsymbol{V}}_{u_2-u_1,k}(x)$ as a $(k+1)$-vector with jth element given by $\widehat{\partial_h^k\nu_{(u_2-u_1)^{j-1}}(x,h)|_{h=0}}$. Let $\boldsymbol{V}_{u_2-u_1,k}(x)$ be a $(p+1)$-vector with the jth element given by $\partial_h^k\nu_{(u_2-u_1)^{j-1}}(x,h)|_{h=0}$.

Lemma G.3. *Let assumptions of theorem 4.2 hold. Then the vector* $\sqrt{Ns^{2+2k}}\big(\hat{\boldsymbol{V}}_{u_2-u_1,k}(x)-\boldsymbol{V}_{u_2-u_1,k}(x,h)\big)$ *weakly converges to a mean-zero normally distributed random vector. Convergence is joint with the vector* $\sqrt{Ns^{2+2k}}(\hat{\boldsymbol{V}}_{\mathcal{G}_k}(x)-\boldsymbol{V}_{\mathcal{G}_k}(x))$ *of lemma G.1.*

Proof. Consider the $(p+1)$st element of $\sqrt{Ns^{2+2k}}\left(\hat{\boldsymbol{V}}_{u_2-u_1,k}(x)-\boldsymbol{V}_{u_2-u_1,k}(x,h)\right)$:

$$
\begin{aligned}
&\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^k\nu_{(u_2-u_1)^p}(x,0)}-\partial_h^k\nu_{(u_2-u_1)^p}(x,0)\right)\\
&=\sum_{j=0}^{p}\binom{p}{j}\sum_{i=0}^{k}\binom{k}{i}\sqrt{Ns^{2+2k}}\Bigg[\left(\widehat{\partial_h^i\nu_{u_1^j}(x-h)|_{h=0}}\right)\left(\widehat{\partial_h^{k-i}\nu_{u_2^{p-j}}(x+h)|_{h=0}}\right)\\
&\qquad\qquad-\left(\partial_h^i\nu_{u_1^j}(x-h)|_{h=0}\right)\left(\partial_h^{k-i}\nu_{u_2^{p-j}}(x+h)|_{h=0}\right)\Bigg]\\
&=\sum_{j=0}^{p}\binom{p}{j}\sum_{i=0}^{k}\binom{k}{i}\sqrt{Ns^{2+2k}}\left(\partial_h^i\nu_{u_1^j}(x-h)|_{h=0}\right)\\
&\qquad\times\left(\widehat{\partial_h^{k-i}\nu_{u_2^{p-j}}(x+h)|_{h=0}}-\partial_h^{k-i}\nu_{u_2^{p-j}}(x+h)|_{h=0}\right)\\
&\quad+\sum_{j=0}^{p}\binom{p}{j}\sum_{i=0}^{k}\binom{k}{i}\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^{k-i}\nu_{u_2^{p-j}}(x+h)|_{h=0}}\right)\\
&\qquad\times\left(\widehat{\partial_h^{i}\nu_{u_1^{j}}(x-h)|_{h=0}}-\partial_h^{i}\nu_{u_1^{j}}(x-h)|_{h=0}\right)
\end{aligned}
$$

Consider the (j,i)th term in the second sum with $i<k$. Exactly as in the proof of lemma G.3 it holds that

$$
\begin{aligned}
&\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^{k-i}\nu_{u_2^{p-j}}(x+h)|_{h=0}}\right)\left(\widehat{\partial_h^i\nu_{u_1^j}(x-h)|_{h=0}}-\partial_h^i\nu_{u_1^j}(x-h)|_{h=0}\right)\\
&=\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^{k-i}\nu_{u_2^{p-j}}(x+h)|_{h=0}}\right)\left(\widehat{\partial_h^i\nu_{u_1^j}(x-h)|_{h=0}}-\partial_h^i\nu_{u_1^j}(x-h)|_{h=0}\right)\\
&\quad+\sqrt{Ns^{2+2k}}O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2(k-i)}}}+s^{q-(k-i)+1}\right)O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2i}}}+s^{q-i+1}\right)\\
&=\sqrt{Ns^{2+2k}}O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2i}}}+s^{q-i+1}\right)\\
&\quad+\sqrt{Ns^{2+2k}}O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2(k-i)}}}+s^{q-(k-i)+1}\right)O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2i}}}+s^{q-i+1}\right)\\
&=o_{a.s.}(1).
\end{aligned}
$$

A similar result applies to the all the terms in the first sum with $i>0$. It follows that

$$
\begin{aligned}
&\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^k\nu_{(u_2-u_1)^p}(x,0)}-\partial_h^k\nu_{(u_2-u_1)^p}(x,0)\right)+o_{a.s.}(1) \qquad (84)\\
&=\sum_{j=0}^{p}\binom{p}{j}\sqrt{Ns^{2+2k}}\left(\partial_h^i\nu_{u_1^j}(x-h)|_{h=0}\right)\left(\widehat{\partial_h^{k-i}\nu_{u_2^{p-j}}(x+h)|_{h=0}}-\partial_h^{k-i}\nu_{u_2^{p-j}}(x+h)|_{h=0}\right)
\end{aligned}
$$

$$+\sum_{j=0}^{p}\binom{p}{j}\sqrt{Ns^{2+2k}}\left(\partial_h^{k-i}\nu_{u_2^{p-j}}(x+h)|_{h=0}\right)\left(\widehat{\partial_h^i\nu_{u_1^j}(x-h)|_{h=0}}-\partial_h^i\nu_{u_1^j}(x-h)|_{h=0}\right)$$

The above expression shows that $\sqrt{Ns^{2+2k}}\big(\hat{\boldsymbol{V}}_{u_2-u_1,k}(x)-\boldsymbol{V}_{u_2-u_1,k}(x,h)\big)$ is a smooth transformation of the vector $\sqrt{Ns^{2+2k}}\big((\hat{\boldsymbol{V}}_{u_1,k}(x)-\boldsymbol{V}_{u_1,k}(x))',(\hat{\boldsymbol{V}}_{u_2,k}(x)-\boldsymbol{V}_{u_2,k}(x))'\big)$ of lemma G.2, plus an $o_{a.s.}(1)$ component. The assertion of the lemma follows by the delta method and lemma G.2. □

As the final building block, we now establish asymptotic normality of the estimator $\widehat{\partial_h^k D_k(x,0)}$ of $\partial_h^k D_k(x,0)$, leveraging lemmas G.1 and G.3.

Lemma G.4. *Let the assumptions of theorem 4.2 hold. Then* $\sqrt{Ns^{2+2k}}\big(\widehat{\partial_h^k D_k(x,0)}-\partial_h^k D_k(x,0)\big)$ *converges weakly to a mean-zero normally distributed random variable.*

Proof. We prove a slightly stronger result. Let $\hat{\boldsymbol{V}}_{D,p}(x)$ be a p-vector with jth element given by $\widehat{\partial_h^k D_j(x,0)}$. Similarly, let $\boldsymbol{V}_{D,p}(x)$ be a p-vector with jth element $\partial_h^k D_j(x,0)$. We claim that $\sqrt{Ns^{2+2k}}\left(\hat{\boldsymbol{V}}_{D,p}(x)-\boldsymbol{V}_{D,p}(x)\right)$ weakly converges to a mean zero normally distributed random vector for all $p=1,\dots,k$, jointly with the vectors $\sqrt{Ns^{2+2k}}(\hat{\boldsymbol{V}}_{\mathcal{G}_k}(x)-\boldsymbol{V}_{\mathcal{G}_k}(x))$ of lemma G.1 and $\sqrt{Ns^{2+2k}}\big(\hat{\boldsymbol{V}}_{u_2-u_1,k}(x)-\boldsymbol{V}_{u_2-u_1,k}(x,h)\big)$ of lemma G.3.

The proof is by (finite) induction on p. The result is immediate for $p=1$, as

$$\sqrt{Ns^{2+2k}}\left(\hat{V}_{D,1,k}(x)-V_{D,1,k}(x)\right)=\sqrt{Ns^{2+2k}}\left(\widehat{R^{(\Delta)}_{(y_2-y_1)}(x,0)}-R^{(\Delta)}_{(y_2-y_1)}(x,0)\right).$$

The result then immediately follows from lemma G.1.

Now suppose that the result holds up to $p-1$, $p\le k$. The kth derivatives of $D_p(x,h)$ can be represented as

$$\begin{aligned}
&\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^k D_p(x,0)}-\partial_h^k D_p(x,0)\right)\\
&=\sqrt{Ns^{2+2k}}\left(\widehat{\partial^k_{w_2}R^{(\Delta)}_{(y_2-y_1)^p}(x,0)}-\partial^k_{w_2}R^{(\Delta)}_{(y_2-y_1)^p}(x,0)\right)\\
&\quad+\sum_{j=0}^{p-1}\binom{p}{j}\sum_{i=0}^{k}\binom{k}{i}\sqrt{Ns^{2+2k}}\Big[\left(\widehat{\partial_h^i D_j(x,0)}\right)\left(\widehat{\partial_h^{k-i}\nu_{(u_2-u_1)^{p-j}}(x,0)}\right)\\
&\qquad\qquad-\left(\partial_h^i D_j(x,0)\right)\left(\partial_h^{k-i}\nu_{(u_2-u_1)^{p-j}}(x,0)\right)\Big]\\
&=\sqrt{Ns^{2+2k}}\left(\widehat{\partial^k_{w_2}R^{(\Delta)}_{(y_2-y_1)^p}(x,0)}-\partial^k_{w_2}R^{(\Delta)}_{(y_2-y_1)^p}(x,0)\right)\\
&\quad+\sum_{j=0}^{p-1}\binom{p}{j}\sum_{i=0}^{k}\binom{k}{i}\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^i D_j(x,0)}\right)\left(\widehat{\partial_h^{k-i}\nu_{(u_2-u_1)^{p-j}}(x,0)}-\partial_h^{k-i}\nu_{(u_2-u_1)^{p-j}}(x,0)\right)
\end{aligned}$$

$$+\sum_{j=0}^{p-1}\binom{p}{j}\sum_{i=0}^{k}\binom{k}{i}\sqrt{Ns^{2+2k}}\left(\partial_h^{k-i}\nu_{(u_2-u_1)^{p-j}}(x,0)\right)\left(\widehat{\partial_h^i D_j(x,0)}-\partial_h^i D_j(x,0)\right).$$

Consider the (j,i)th term in first sum for $i>0$. Analogously to the proof of lemma G.2, we obtain by eq. (78) that

$$\begin{aligned}
&\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^i D_j(x,0)}\right)\left(\widehat{\partial_h^{k-i}\nu_{(u_2-u_1)^{p-j}}(x,0)}-\partial_h^{k-i}\nu_{(u_2-u_1)^{p-j}}(x,0)\right)\\
&=\sqrt{Ns^{2+2k}}\left(\partial_h^i D_j(x,0)\right)\left(\widehat{\partial_h^{k-i}\nu_{(u_2-u_1)^{p-j}}(x,0)}-\partial_h^{k-i}\nu_{(u_2-u_1)^{p-j}}(x,0)\right)\\
&\quad+\sqrt{Ns^{2+2k}}O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2(k-i)}}}+s^{q-(k-i)+1}\right)O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2i}}}+s^{q-i+1}\right)\\
&=\sqrt{Ns^{2+2k}}O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2(k-i)}}}+s^{q-(k-i)+1}\right)\\
&\quad+\sqrt{Ns^{2+2k}}O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2(k-i)}}}+s^{q-(k-i)+1}\right)O_{a.s.}\left(\sqrt{\frac{\log(N)}{Ns^{2+2i}}}+s^{q-i+1}\right)\\
&=o_{a.s.}(1).
\end{aligned}$$

It then holds that

$$\begin{aligned}
&\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^k D_p(x,0)}-\partial_h^k D_p(x,0)\right)+o_{a.s.}(1) \qquad (85)\\
&=\sqrt{Ns^{2+2k}}\left(\widehat{\partial_{w_2}^k R^{(\Delta)}_{(y_2-y_1)^p}(x,0)}-\partial_{w_2}^k R^{(\Delta)}_{(y_2-y_1)^p}(x,0)\right)\\
&\quad+\sum_{j=0}^{p-1}\binom{p}{j}\left(D_j(x,0)\right)\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^k\nu_{(u_2-u_1)^{p-j}}(x,0)}-\partial_h^k\nu_{(u_2-u_1)^{p-j}}(x,0)\right)\\
&\quad+\sum_{j=0}^{p-1}\binom{p}{j}\left(\nu_{(u_2-u_1)^{p-j}}(x,0)\right)\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^k D_j(x,0)}-\partial_h^k D_j(x,0)\right).
\end{aligned}$$

We conclude that the vector $\sqrt{Ns^{2+2k}}\left(\hat{\boldsymbol{V}}_{D,p}(x)-\boldsymbol{V}_{D,p,}(x)\right)$ is a smooth transformation of $\sqrt{Ns^{2+2k}}\big(\hat{\boldsymbol{V}}_{D,p-1}(x)-\boldsymbol{V}_{D,p-1}(x)\big)$, $\sqrt{Ns^{2+2k}}(\hat{\boldsymbol{V}}_{\mathcal{G}_k}(x)-\boldsymbol{V}_{\mathcal{G}_k}(x))$, and $\sqrt{Ns^{2+2k}}\big(\hat{\boldsymbol{V}}_{u_2-u_1,k}(x)-\boldsymbol{V}_{u_2-u_1,k}(x,h)\big)$, plus an $o_{a.s.}(1)$ term. Asymptotic normality and jointness of convergence now follows from the inductive assumption, lemmas G.1 and G.3, and the delta method. □

G.2 Proof of Theorem 4.2

Proof of theorem 4.2. Lemma G.4 shows that $\sqrt{Ns^{2+2k}}\left(\widehat{\partial_h^k D_k(x,0)} - \partial_h^k D_k(x,0)\right)$ is asymptotically normally distributed. The result of the theorem follows as

$$\sqrt{Ns^{2+2k}}\left(\hat{\mu}_k(x) - \mu_k(x)\right) = \frac{1}{2^k k!}\sqrt{Ns^2}\left(\widehat{\partial_h^k D_k(x,0)} - \partial_h^k D_k(x,0)\right). \tag{86}$$

Asymptotic independence of $\sqrt{Ns^{2+2k}}\left(\hat{\mu}_k(x_1) - \mu_k(x_1)\right)$ and $\sqrt{Ns^{2+2k}}\left(\hat{\mu}_k(x_2) - \mu_k(x_2)\right)$ for $x_1 \neq x_2$ follows from lemma G.1. □

Remark 10 (On $V_k(x)$)**.** An expression for the asymptotic variance $V_k(x)$ of $\hat{\mu}_k(x)$ may be obtained from eqs. (85) and (86). Note that all components in eq. (85) are jointly asymptotically normal. Accordingly, $V_k(x)$ is a sum of variances and covariances of the estimator for kth derivatives of expectation of $(Y_{i2} - Y_{i1})^p$, the estimators for kth derivatives of $\nu_{(u_2-u_1)^p}$, $p = 2, \ldots, k-1$, and the estimators for kth derivatives of D_p, $p = 1, \ldots, k-1$. There are $(2k-3)(2k-2)$ such terms in $V_k(x)$. Note that the variances and covariances for intermediate estimators may be obtained using representations (83), (84), and (85).

Remark 11. A plug-in estimator of $V_k(x)$ may be constructed using its characterization in the proof of theorem 4.2, and inference may be done using the asymptotic normality result of eq. (24). However, this approach has two disadvantages relative to the bootstrap. First, estimators based on such an analytical expression may perform poorly in finite samples. To see the issue, note that derivatives of all orders up to k enter the estimator $\hat{\mu}_k(x)$, as algorithm 1 makes clear. The asymptotic variance of the derivatives of order $j < k$ is negligible relative to that of the kth derivatives. Correspondingly, lower-order derivative make no contribution to the asymptotic variance of the moment estimator. However, these derivatives may still contribute significantly to finite-sample variability. Not accounting for this contribution may then lead to poor performance of resulting confidence intervals and tests. Second, the resulting expression for $V_k(x)$ will be complex for $k \geq 2$. To see see, note that the expression for $\hat{\mu}_k(x)$ involves $(k+1)$ kth derivatives, yielding a total of $(k+1)(k+2)/2$ distinct variance and covariance terms. Of these, k terms correspond to variances of $\widehat{\partial_h^k \nu_{(u_2-u_1)^p}(x,0)}$. Each of these terms in turn relies on k kth derivatives of moments of u_{it}, which further rely on additional k kth derivatives.[9]

[9]For $k = 1$ the estimator satisfies $\hat{\mu}_1(x) = \frac{1}{2}\widehat{\partial_h^l r_{(y_2-y_1)}(x-h, x+h)|_{h=0}}$. The asymptotic variance V_1 can then be obtained fairly straightforwardly from variance expressions for local polynomial estimators.

H Proof of Theorem 5.1

Define the $\bar{p}_v \times \bar{p}_v$ matrix $\boldsymbol{M}_{\Psi,\bar{p}_v}$ as having the (k,j)th element $\int v^{k-1}\Psi(d(v - v_{j,p}))$; that is, the kth row of $\boldsymbol{M}_{\Psi,\bar{p}_v}$ are the $(k-1)$th moments of the individual mixture components.

Lemma H.1. *Let assumption 5.1 hold. Then $\boldsymbol{M}_{\psi,\bar{p}_v}$ has rank $\bar{p}_v$*

Proof. Let $\psi(v) = \Psi'(v)$, which exists under assumption 5.1. Define $\mu_{j,\psi} = \int v^j \psi(v) dv$. The kth moment of the jth mixture component $\Psi(v - v_{j,\bar{p}_v})$ can be represented as

$$\begin{aligned}
\int v^k \psi(v - v_{j,\bar{p}_v}) dv &= \int ((v - v_{j,\bar{p}_v}) + v_{j,\bar{p}_v})^k \psi(v - v_{j,\bar{p}_v}) dv \\
&= \sum_{i=0}^{k} v_{j,\bar{p}_v}^i \int (v - v_{j,\bar{p}_v})^{k-i} \psi(v - v_{j,\bar{p}_v}) dv \\
&= \sum_{i=0}^{k} \binom{k}{i} v_{j,\bar{p}_v}^i \mu_{k-i,\psi}.
\end{aligned}$$

$\int v^k \psi(v - v_{j,\bar{p}_v}) dv$ is the $(k+1,j)$th element of $\boldsymbol{M}_{\Psi,\bar{p}_v}$ Accordingly, $\boldsymbol{M}_{\psi,\bar{p}_v}$ may be written out as

$$\boldsymbol{M}_{\psi,\bar{p}_v} = \begin{pmatrix}
\binom{0}{0}\mu_{0,\psi} & 0 & \cdots & 0 & 0 & \cdots & 0 \\
\binom{1}{0}\mu_{1,\psi} & \binom{1}{1}\mu_{0,\psi} & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\binom{k}{0}\mu_{k,\psi} & \binom{k}{1}\mu_{k-1,\psi} & \cdots & \binom{k}{k}\mu_{0,\psi} & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\binom{\bar{p}_v-1}{0}\mu_{\bar{p}_v-1,\psi} & \binom{\bar{p}_v-1}{1}\mu_{\bar{p}_v-2,\psi} & \cdots & \binom{\bar{p}_v-1}{k}\mu_{\bar{p}_v-1-k,\psi} & \binom{\bar{p}_v-1}{k+1}\mu_{\bar{p}_v-2-k,\psi} & \cdots & \binom{\bar{p}_v-1}{\bar{p}_v-1}\mu_{0,\psi}
\end{pmatrix} \boldsymbol{V}$$

$$\boldsymbol{V} \coloneqq \begin{pmatrix}
1 & 1 & \dots & 1 \\
v_{1,\bar{p}_v} & v_{2,\bar{p}_v} & \dots & v_{\bar{p}_v,\bar{p}_v} \\
\vdots & \vdots & \ddots & \vdots \\
v_{1,\bar{p}_v}^{\bar{p}_v-1} & v_{2,\bar{p}_v}^{\bar{p}_v-1} & \dots & v_{\bar{p}_v,\bar{p}_v}^{\bar{p}_v-1}
\end{pmatrix} \tag{87}$$

As $v_{i,\bar{p}_v} \neq v_{j,\bar{p}_v}$ for $i \neq j$, the Vandermonde matrix $\boldsymbol{V}$ is non-singular. Then $\boldsymbol{M}_{\Psi,\bar{p}_v}$ is written as a product of a lower triangular matrix with ones on the diagonal and a full-rank matrix. We conclude that $\boldsymbol{M}_{\Psi,\bar{p}_v}$ is full rank as well. □

Proof of theorem 5.1. Let $\boldsymbol{\rho}(x) = (\rho_1(x), \dots, \rho_{\bar{p}_v}(x))$ be a vector of candidate mixing probabilities, that is, $\sum_{j=1}^{\bar{p}_v} \rho_(x) = 1$, $\rho(x) \geq 0$ for all $x \in I$, and each ρ is twice differentiable.

Let π be a finite measure on I such that the Lebesgue measure is absolutely continuous

with respect to π. Consider the following criterion function

$$\bar{Q}(\boldsymbol{\rho}(x)) = \int \sum_{k=0}^{\bar{p}_v - 1} \frac{1}{k!} \left[\mu_k(x) - \int v^k \sum_{j=1}^{\bar{p}_v} \rho_j(x) \psi(v - v_{j,\bar{p}_v}) dv \right]^2 \pi(dx). \tag{88}$$

Further, let

$$\boldsymbol{\Omega} = \text{diag}\{1/0!, 1/1!, \ldots, 1/(\bar{p}_v - 1)!\} \tag{89}$$

Last, note that

$$\mu_k(x) = \int v^k \sum_{j=1}^{\bar{p}_v} \rho_{0,j}(x) \psi(v - v_{j,\bar{p}_v}) dv.$$

Observe that the objective function can then be written as

$$\bar{Q}(\boldsymbol{\rho}) = \int \left[\boldsymbol{\rho}_0(x) - \boldsymbol{\rho}(x)\right]' \boldsymbol{M}'_{\Psi,\bar{p}_v} \boldsymbol{\Omega} \boldsymbol{M}_{\Psi,\bar{p}_v} \left[\boldsymbol{\rho}_0(x) - \boldsymbol{\rho}(x)\right] \pi(dx), \tag{90}$$

where $\boldsymbol{\rho}_0(x) = (\rho_{0,\bar{p}_v}(x), \ldots, \rho_{0,\bar{p}_v}(x))$ is the vector of true mixing probabilities and the matrix $\boldsymbol{M}_{\Psi,\bar{p}_v}$ is defined before lemma H.1.

We show that $\boldsymbol{\rho}_0$ is identified as the unique vector of mixing probabilities such that $Q(\boldsymbol{\rho}) = 0$. Let $\boldsymbol{\rho}$ be such that $Q(\boldsymbol{\rho}) = 0$. By assumption, the integrand in (90) is continuous and $\pi(I)$ is absolutely continuous with respect to the Lebesgue measure. We conclude that for all x it then holds that

$$\left[\boldsymbol{\rho}_0(x) - \boldsymbol{\rho}(x)\right]' \boldsymbol{M}'_{\Psi,\bar{p}_v} \boldsymbol{\Omega} \boldsymbol{M}_{\Psi,\bar{p}_v} \left[\boldsymbol{\rho}_0(x) - \boldsymbol{\rho}(x)\right] = 0$$

By lemma H.1, the matrix $\boldsymbol{M}_{\Psi,\bar{p}_v}$ has maximal rank $\bar{p}_v$. Correspondingly, $\boldsymbol{M}'_{\Psi,\bar{p}_v} \boldsymbol{\Omega} \boldsymbol{M}_{\Psi,\bar{p}_v}$ is positive definite. We conclude that $\boldsymbol{\rho}_0(x) = \boldsymbol{\rho}(0)$ for all $x \in I$. □

I Proof of Theorem 5.2

Proof of theorem 5.2. We only establish the second assertion. The proof of the first one is analogous, but simpler. Let $\tilde{\boldsymbol{\mu}}_{\bar{p}_v}(x)$ be the vector of the first $\bar{p}_v$ estimated moments of marginal effects, starting for the zeroth moment, and let $\boldsymbol{\mu}_{\bar{p}_v}(x)$ be its population counterpart:

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_{\bar{p}_v}(x) &= (\tilde{\mu}_0(x), \tilde{\mu}_1(x), \ldots, \tilde{\mu}_{\bar{p}_v - 1}(x))', \\ \boldsymbol{\mu}_{\bar{p}_v} &= (\mu_0(x), \mu_1(x), \ldots, \mu_{\bar{p}_v - 1}(x))'. \end{aligned}$$

We begin by extending the definition of the objective function (21) to the space of distributions of assumption 5.1. Let $(\rho_1(x), \ldots, \rho_{\bar{p}_v}(x))$ be a vector of mixing probabilities for $x \in I$, that is, $\rho_j(x) \geq 0$, $\sum_{j=1}^{\bar{p}_v} \rho_j(x) = 1$ for all $x \in I$. The sample objective function

evaluate at a mixture distribution with mixing probabilities $\rho_j(x)$ is given by

$$\begin{aligned}\hat{Q}(\boldsymbol{\rho}) &= \int \sum_{k=0}^{\bar{p}_v-1} \frac{1}{k!}\left[\hat{\mu}_k(x) - \int v^k \sum_{j=1}^{\bar{p}_v} \rho_j(x)\psi(v-v_j)dv\right]^2 \pi(dx) \\ &= \int \left[\hat{\boldsymbol{\mu}}_{\bar{p}_v} - \boldsymbol{M}_{\psi,\bar{p}_v}\boldsymbol{\rho}(x)\right]' \boldsymbol{\Omega}\left[\hat{\boldsymbol{\mu}}_{\bar{p}_v} - \boldsymbol{M}_{\psi,\bar{p}_v}\boldsymbol{\rho}(x)\right]\pi(dx)\end{aligned}$$

where the matrix $\boldsymbol{M}_{\Psi,\bar{p}_v}$ is defined before lemma H.1 and $\boldsymbol{\Omega}$ is defined in eq. (89). Note that this agrees with the definition of eq. (21) if ρ_j can be written as sums of Bernstein polynomials.

We now bound the difference between $\hat{Q}(\boldsymbol{\rho})$ and the function $\bar{Q}(\boldsymbol{\rho})$ of eq. (88) uniformly in $\boldsymbol{\rho}$ permitted by assumption 5.1:

$$\begin{aligned}&\left|\hat{Q}(\boldsymbol{\rho}) - \bar{Q}(\boldsymbol{\rho})\right| \\ &= \left| \int \left[\left[\tilde{\boldsymbol{\mu}}_{\bar{p}_v} - \boldsymbol{M}_{\Psi,\bar{p}_v}\boldsymbol{\rho}(x)\right]' \boldsymbol{\Omega}\left[\tilde{\boldsymbol{\mu}}_{\bar{p}_v} - \boldsymbol{M}_{\Psi,\bar{p}_v}\boldsymbol{\rho}(x)\right]\right.\right. \\ &\quad \left.\left. - \left[\boldsymbol{\mu}_{\bar{p}_v} - \boldsymbol{M}_{\Psi,\bar{p}_v}\boldsymbol{\rho}(x)\right]' \boldsymbol{\Omega}\left[\boldsymbol{\mu}_{\bar{p}_v} - \boldsymbol{M}_{\Psi,\bar{p}_v}\boldsymbol{\rho}(x)\right]\right]\pi(dx)\right| \\ &= \left| \int \left[2\left[\tilde{\boldsymbol{\mu}}_{\bar{p}_v}(x) - \boldsymbol{\mu}_{\bar{p}_v}(x)\right]' \boldsymbol{\Omega}\boldsymbol{M}_{\Psi,\bar{p}_v}\boldsymbol{\rho}(x) + \left[\tilde{\boldsymbol{\mu}}_{\bar{p}_v}(x)'\tilde{\boldsymbol{\mu}}_{\bar{p}_v}(x) - \boldsymbol{\mu}_{\bar{p}_v}(x)'\boldsymbol{\mu}_{\bar{p}_v}(x)\right]\right]\pi(dx)\right| \\ &\leq \pi(I)\left[\sup_{x\in I}\left\|\tilde{\boldsymbol{\mu}}_{\bar{p}_v}(x) - \boldsymbol{\mu}_{\bar{p}_v}(x)\right\|_\infty \|\boldsymbol{\Omega}\|_1 \left\|\boldsymbol{M}_{\Psi,\bar{p}_v}\right\|_1 \sup_{x\in I}\|\boldsymbol{\rho}(x)\|_1\right. \\ &\qquad \left. + \sup_{x\in I}\left|\left\|\tilde{\boldsymbol{\mu}}_{\bar{p}_v}(x)\right\|_2^2 - \left\|\boldsymbol{\mu}_{\bar{p}_v}(x)\right\|_2^2\right|\right] \\ &\leq \pi(I)\left[\sup_{x\in I}\left\|\tilde{\boldsymbol{\mu}}_{\bar{p}_v}(x) - \boldsymbol{\mu}_{\bar{p}_v}(x)\right\|_\infty \|\boldsymbol{\Omega}\|_1 \left\|\boldsymbol{M}_{\Psi,\bar{p}_v}\right\|_1 + \max_{x\in I}\left|\left\|\tilde{\boldsymbol{\mu}}_{\bar{p}_v}(x)\right\|_2^2 - \left\|\boldsymbol{\mu}_{\bar{p}_v}(x)\right\|_2^2\right|\right] \\ &\leq \pi(I)\left[\sup_{x\in I}\left\|\tilde{\boldsymbol{\mu}}_{\bar{p}_v}(x) - \boldsymbol{\mu}_{\bar{p}_v}(x)\right\|_\infty \left\|\boldsymbol{M}_{\Psi,\bar{p}_v}\right\|_1 + \max_{x\in I}\left|\left\|\tilde{\boldsymbol{\mu}}_{\bar{p}_v}(x)\right\|_2^2 - \left\|\boldsymbol{\mu}_{\bar{p}_v}(x)\right\|_2^2\right|\right] \\ &=: \eta_N \end{aligned} \tag{91}$$

where we use the fact that

$$\|\boldsymbol{\rho}(x)\|_1 = \sum_{j=1}^{\bar{p}_v}|\rho_j(x)| = 1.$$

η_N is controlled by the errors in the first $\bar{p}_v$ moments, uniformly over the space of mixing probabilities:

$$\eta_N = O_{a.s.}\left(\max_{k=1,\ldots,\bar{p}_v-1}\{\delta_{k,N}\}\right). \tag{92}$$

Approximation (20) models mixing probabilities $\boldsymbol{\rho}(x)$ using Bernstein polynomials of order p_x. Correspondingly, define the $\bar{p}_v$-vector $\boldsymbol{B}(\boldsymbol{\gamma}, x)$ as

$$\boldsymbol{B}(\boldsymbol{\gamma}, x) = \left(\sum_{l=0}^{p_x} \gamma_{1,l} b_{l,p_x}(x), \ldots, \sum_{l=0}^{p_x} \gamma_{\bar{p}_v,l} b_{l,p_x}(x) \right)'$$

where $\boldsymbol{\gamma}$ is as $\bar{p}_v \times (p_x + 1)$ matrix that satisfies $\sum_{j=1}^{p_v} \gamma_{j,l} = 1$ for $l = 0, 1, \ldots p_x$, $\gamma_{j,l} \geq 0$ for all j, l. Note that $\boldsymbol{B}(\boldsymbol{\gamma}, x)$ is a vector of mixing probabilities for every x and every valid $\boldsymbol{\gamma}$ as $b_{l,p_x}(x) \geq 0$ for all x and

$$\sum_{j=1}^{p_v} \left(\sum_{l=0}^{p_x} \gamma_{j,l} b_{l,p_x}(x) \right) = \sum_{l=0}^{p_x} b_{l,p_x}(x) \sum_{j=1}^{p_v} \gamma_{j,l} = \sum_{l=0}^{p_x} b_{l,p_x}(x) = 1.$$

For future reference, define the Bersntein weights $\boldsymbol{\gamma}^B$ as

$$\gamma^B_{j,l} = \rho_{0,j} \left(x_{lb} + \frac{l(x_{ub} - x_{lb})}{p_x} \right).$$

where $I = [x_{lb}, x_{ub}]$. Observe that for any l it holds that $\sum_{j=1}^{\bar{p}_v} \gamma^B_{j,l} = 1$, as those are mixing probabilities at point $x = x_{lb} + l(x_{ub} - x_{lb})/p_x$. Further, all are non-negative. Correspondingly, $\boldsymbol{\gamma}^B$ is a feasible choice in the optimization problem (23). Further, $\boldsymbol{\gamma}^B$ satisfies

$$\sup_{x \in I} \left| \rho_{0,j}(x) - \sum_{l=0}^{p_x} \gamma_{j,l} b_{l,p_x}(x) \right| = O(p_x^{-1}), \quad j = 0, 1, \ldots, \bar{p}_v. \tag{93}$$

by theorem 4.29 of Bustamante (2017).

We now turn to the estimator $\hat{F}_N(v|x)$ of (22). The total variation distance between $\hat{F}_N(v|x)$ and $F_0(v|x)$ can be expressed in terms of L^1 distance between the corresponding densities (which exist under assumption 5.1). Let $\hat{f}_N(v|x) = \partial_v \hat{F}_N(v|x)$ and $f_0(v|x) = \partial_v F_0(v|x)$. These densities may be written as

$$\hat{f}_N(v|x) = \sum_{j=1}^{\bar{p}_v} \sum_{l=0}^{p_x} \hat{\gamma}_{j,l} b_{l,p_x}(x) \psi(v - v_{j,\bar{p}_v}), \quad f_0(v|x) = \sum_{j=1}^{\bar{p}_v} \rho_{0,j}(x) \psi(v - v_{j,\bar{p}_v}).$$

where $\hat{\boldsymbol{\gamma}}$ is defined in eq. (23) and $\psi = \Psi'$. Then for any $x \in I$

$$d_{TV}(\hat{F}_N(\cdot|x), F_0(\cdot|x)) = \frac{1}{2} \int |\hat{f}_N(v|x) - f_0(v|x)| dv.$$

Correspondingly, the distance of interest may be bounded as

$$2 \int d_{TV}(\hat{F}_N(\cdot|x), F_0(\cdot|x)) \pi(dx)$$

$$
\begin{aligned}
&= \iint \left| \sum_{j=1}^{\bar{p}_v} \left[\sum_{l=0}^{p_x} \hat{\gamma}_{j,l} b_{l,p_x}(x) - \rho_{0,j}(x) \right] \psi(v - v_{j,\bar{p}_v}) \right| dv\pi(dx) \\
&\leq \iint \sum_{j=1}^{\bar{p}_v} \left| \sum_{l=0}^{p_x} \hat{\gamma}_{j,l} b_{l,p_x}(x) - \rho_{0,j}(x) \right| \psi(v - v_{j,\bar{p}_v}) dv\pi(dx) \\
&= \int \sum_{j=1}^{\bar{p}_v} \left| \sum_{l=0}^{p_x} \hat{\gamma}_{j,l} b_{l,p_x}(x) - \rho_{0,j}(x) \right| \pi(dx) \\
&= \int \left\| \boldsymbol{B}(\hat{\boldsymbol{\gamma}}, x) - \boldsymbol{\rho}_0(x) \right\|_1 \pi(dx). \qquad (94)
\end{aligned}
$$

To bound $\left\| \boldsymbol{B}(\hat{\boldsymbol{\gamma}}, x) - \boldsymbol{\rho}_0(x) \right\|_1$, consider the quadratic form $\boldsymbol{z}' \left(\boldsymbol{M}'_{\Psi,\bar{p}_v} \boldsymbol{\Omega} \boldsymbol{M}_{\Psi,\bar{p}_v} \right) \boldsymbol{z}$. Let $\lambda_{\min}(\boldsymbol{A})$ be the minimal eigenvalue of a square matrix $\boldsymbol{A}$. The following bound obtains:

$$
\begin{aligned}
&\lambda_{\min}(\boldsymbol{M}'_{\Psi,\bar{p}_v} \boldsymbol{\Omega} \boldsymbol{M}_{\Psi,\bar{p}_v}) \int \left\| \boldsymbol{B}(\hat{\boldsymbol{\gamma}}, x) - \boldsymbol{\rho}_0(x) \right\|_2^2 \pi(dx) \\
&\leq \int \left(\boldsymbol{B}(\hat{\boldsymbol{\gamma}}, x) - \boldsymbol{\rho}_0(x) \right)' \boldsymbol{M}'_{\Psi,\bar{p}_v} \boldsymbol{\Omega} \boldsymbol{M}_{\Psi,\bar{p}_v} \left(\boldsymbol{B}(\hat{\boldsymbol{\gamma}}, x) - \boldsymbol{\rho}_0(x) \right) \pi(dx) \\
&\equiv \bar{Q}(\boldsymbol{B}(\hat{\boldsymbol{\gamma}}, \cdot)) \leq \hat{Q}(\boldsymbol{B}(\hat{\boldsymbol{\gamma}}, \cdot)) + \eta_N \leq \hat{Q}(\boldsymbol{B}(\boldsymbol{\gamma}^B, \cdot)) + \eta_N \leq \bar{Q}(\boldsymbol{B}(\boldsymbol{\gamma}^B, \cdot)) + 2\eta_N \\
&= O_{a.s.} \left(p_x^{-2}, \max_{k=1,\ldots,\bar{p}_v - 1} \delta_{k,N} \right), \qquad (95)
\end{aligned}
$$

where in the first equality we use eq. (90); in the inequalities we use the uniform convergence of $\hat{Q}$ to $\bar{Q}$ of eq. (91); in the last line we use (92) and

$$
\begin{aligned}
\bar{Q}(\boldsymbol{B}(\boldsymbol{\gamma}^B, \cdot)) &= \int \left(\boldsymbol{B}(\boldsymbol{\gamma}^B, x) - \boldsymbol{\rho}_0(x) \right)' \boldsymbol{M}'_{\Psi,\bar{p}_v} \boldsymbol{\Omega} \boldsymbol{M}_{\Psi,\bar{p}_v} \left(\boldsymbol{B}(\boldsymbol{\gamma}^B, x) - \boldsymbol{\rho}_0(x) \right) \pi(dx) \\
&\leq \lambda_{\max}(\boldsymbol{M}'_{\psi,\bar{p}_v} \boldsymbol{\Omega} \boldsymbol{M}_{\Psi,\bar{p}_v}) \int \left\| \boldsymbol{B}(\boldsymbol{\gamma}^B, x) - \boldsymbol{\rho}_0(x) \right\|_2^2 \pi(dx) \\
&\leq O(p_x^{-2}),
\end{aligned}
$$

where the last line follows from (93).

The result of the theorem follows from observing that (1) $\lambda_{\min}(\boldsymbol{M}'_{\Psi,\bar{p}_v} \boldsymbol{\Omega} \boldsymbol{M}_{\Psi,\bar{p}_v}) > 0$, as $\boldsymbol{M}_{\Psi,\bar{p}_v}$ is full-rank by lemma H.1, (2) $\left\| \boldsymbol{B}(\hat{\boldsymbol{\gamma}}, x) - \boldsymbol{\rho}_0(x) \right\|_1 \leq \bar{p}_v \left\| \boldsymbol{B}(\hat{\boldsymbol{\gamma}}, x) - \boldsymbol{\rho}_0(x) \right\|_2$ and (3) combining eqs. (94) and (95). □

J Moment Metrics

We first show that $d_{2\mu}$ and $d^{\pi}_{2,\mu}$ are metrics.

Lemma J.1. *Let $\mathcal{Q}$ be the class of cumulative distribution functions with bounded support.*

Define

$$d_{2,\mu}(F,G) = \left(\sum_{k=1}^{\infty} \frac{1}{k!} \left[\int v^k F(dv) - \int v^k G(dv) \right]^2 \right)^{1/2}.$$

Then (1) $d_{2,\mu}(F,G) < \infty$ for all $F, G \in \mathcal{Q}$; (2) $d_{2,\mu}(\cdot,\cdot)$ is a metric on $\mathcal{Q}$.

Proof. For (1), let $F, G \in \mathcal{Q}$. By definition of $\mathcal{Q}$, there exists some $M > 0$ such that $\text{supp}(F) \subset [-M, M]$ and $\text{supp}(G) \subset [-M, M]$. Then $\int |v|^k F(dv) \leq M^k$ and $\int |v|^k G(dv) \leq M^k$. Thus,

$$d_{2,\mu}(F,G) = \sum_{k=1}^{\infty} \frac{1}{k!} \left[\int v^k F(dv) - \int v^k G(dt) \right]^2 \leq \sum_{k=1}^{\infty} \frac{2M^{2k}}{k!} < \infty.$$

For (2), we check the defining properties of a metric:

1. $d_{2,\mu}^2(F,G) = 0$ implies that $\int v^k F(dv) = \int v^k G(dv)$ for all $k = 1, 2, \ldots$. Both F and G have bounded support, hence the moments uniquely determine the corresponding distributions. We conclude that $F = G$.

2. Symmetry is immediate.

3. As $d_{2,\mu}$ has the form of an L^2 metric, the triangle inequality is established by a standard application of the Cauchy-Schwarz inequality: if $F, G, H \in \mathcal{Q}$

$$\begin{aligned}
d_{2,\mu}^2(F,G) &= \sum_{k=1}^{\infty} \frac{1}{k!} \left[\int v^k F(dv) - \int v^k G(dv) \right]^2 \\
&= \sum_{k=1}^{\infty} \frac{1}{k!} \left[\int v^k F(dv) - \int v^k H(dv) \right]^2 + \sum_{k=1}^{\infty} \frac{1}{k!} \left[\int v^k H(dv) - \int v^k G(dv) \right]^2 \\
&\quad + \sum_{k=1}^{\infty} \frac{1}{k!} \left[\int v^k F(dv) - \int v^k H(dv) \right] \left[\int v^k H(dv) - \int v^k G(dv) \right] \\
&\leq d_{2,\mu}^2(F,H) + d_{2,\mu}^2(H,G) + 2 d_{2,\mu}(F,H) d_{2,\mu}(H,G) \\
&= \left[d_{2,\mu}(F,H) + d_{2,\mu}(H,G) \right]^2.
\end{aligned}$$

□

Remark 12. $d_{2,\mu}$ is in fact a metric of weak convergence on the class of distributions with support lying in $[-M, M]$, where M is a fixed constant. On the larger class $\mathcal{Q}$ the metric $d_{2,\mu}$ metrizes a notion of convergence that it stronger than weak convergence. To see this, let the distribution F_n place mass $(1/n)$ on $n!$ and mass $(1 - 1/n)$ on 0. Then F_n weakly converges to a point mass at 0, but does not converge in the moment metric.

We now show that $d^{\pi}_{2,\mu}$ is a metric.

Lemma J.2. . *Let $I \subset \mathbb{R}$ be a closed interval. Let $\mathcal{Q}^I$ be the space of bivariate functions $F(v|x) : \mathbb{R} \times I \to [0,1]$ such that:*
(1) for each $t \in \mathbb{R}$ the function $F(v|x)$ is continuous in x;
(2) for each $x \in I$ the function $F(v|x)$ is a cdf
(3) the conditional support of $F(v|x)$ is bounded uniformly in x, that is $\text{supp}(F(v|x)) \subset [-M_F, M_F]$ for some $M_F \geq 0$ that does not depend on x (but may depend on F).
Let $d_{2,\mu}$ be defined as in lemma J.1. Let π be a finite measure on I such that the Lebesgue measure on I is absolutely continuous with respect to π. Define

$$d^{\pi}_{2,\mu}(F,G) = \int_I d_{2,\mu}\left(F(\cdot|x), G(\cdot|x)\right) \pi(dx)$$

Then $d^{\pi}_{2,\mu}$ is a metric on $\mathcal{Q}^I_{M_Q}$.

Proof. Let $d^{\pi}_{2,\mu}(F,G) = 0$. Then $d_{2,\mu}\left(F(\cdot|x), G(\cdot|x)\right) = 0$ for π-almost all (a.a.) x. By assumption, then also $d_{2,\mu}\left(F(\cdot|x), G(\cdot|x)\right) = 0$ for Lebesgue-a.a. x . Since $d_{2,\mu}\left(F(\cdot|x), G(\cdot|x)\right)$ is continuous in x, we conclude that equality holds for all $x \in I$. By lemma J.1 then for each x it holds that $F(v|x) = G(v|x)$ for all v. We conclude that $F = G$ as functions in $\mathcal{Q}^I$. By lemma J.1 $d^{\pi}_{2,\mu}$ satisfies all other properties of a metric. □

The convergence result of theorem 5.4 is stated of terms the Kolmogorov metric, and not the moment metric used in estimation. Accordingly, the following lemma establishes an upper bound on the former metric in terms of the latter one for sequences that are convergent or divergent in the moment metrics. More generally, the proof provides a generic upper bound on the Kolmogorov metric in terms of the moment metrics for any pair of distributions (eq. (99)). A related bound using the Zolotarev λ-metric is obtained by Tardella (2001) for the case where a given number of moments are matched exactly.

Lemma J.3. *Let $\mathcal{F}, \mathcal{F}^I$ be as in lemma 5.3.*
(1) Let $F, F_n \in \mathcal{F}$ Then

$$\sup_{v \in \mathbb{R}} |F_n(v) - F(v)| = \begin{cases} O\left(d_{2,\mu}(F_N, F) \log(d_{2,\mu}(F_N, F))\right), & d_{2,\mu}(F_n, F) \to \infty, \\ O\left([-\log(d_{2,\mu}(F_n, F))]^{-1/2}\right), & d_{2,\mu}(F_n, F) \to 0. \end{cases} \tag{96}$$

(2) Let $F_n, F \in \mathcal{F}^I$. Then

$$\int_I \sup_{v \in \mathbb{R}} |F_n(v|x) - F(v|x)| \pi(dx) = \begin{cases} O\left(d^{\pi}_{2,\mu}(F_N, F) \log(d^{\pi}_{2,\mu}(F_N, F))\right), & d^{\pi}_{2,\mu}(F_n, F) \to \infty, \\ O\left([-\log(d^{\Pi}_{2,\mu}(F_n, F))]^{-1/2}\right), & d^{\pi}_{2,\mu}(F_n, F) \to 0. \end{cases} \tag{97}$$

Proof. By theorem 1.5.2 in Ibragimov and Linnik (1971), the uniform distance between two cdfs can be bounded from above as

$$\sup_{v\in\mathbb{R}}|F(v|x)-G(v|x)| \leq \frac{2}{\pi}\int_0^L \left|\frac{\varphi_F(w|x)-\varphi_G(w|x)}{w}\right| dw + \frac{24}{\pi}\frac{\|g\|_\infty}{L} \tag{98}$$

where φ_F and φ_G are the characteristic functions of F and G, respectively, and $g = G'$, and $L > 0$ is a scalar. The difference of characteristic functions can be expressed as

$$\varphi_F(w|x)-\varphi_G(w|x) = \sum_{k=0}^{\infty}\frac{(iw)^k}{k!}\left[\int z^k F(dz) - \int z^k G(gz)\right].$$

By Hölder's inequality for positive w it holds that

$$\begin{aligned}
|\varphi_F(w|x)-\varphi_G(w|x)| &\leq \sum_{k=1}^{\infty}\frac{w^k}{\sqrt{k!}}\frac{|\int z^k F(dz) - \int z^k G(dz)|}{\sqrt{k!}} \\
&\leq d_{2,\mu}(F(\cdot|x),G(\cdot|x))\sqrt{\sum_{k=1}^{\infty}\frac{(w^2)^k}{k!}} \\
&= d_{2,\mu}(F(\cdot|x),G(\cdot|x))(e^{w^2}-1)^{1/2}.
\end{aligned}$$

Then

$$\begin{aligned}
&\frac{2}{\pi}\int_0^L \left|\frac{\varphi_F(w|x)-\varphi_G(w|x)}{w}\right| dv + \frac{24}{\pi}\frac{\|g\|_\infty}{L} \\
&\leq \frac{2d_{2,\mu}(F(\cdot|x),G(\cdot|x))}{\pi}\left[\int_0^1 \sqrt{e^{w^2}-1}dv + \int_1^L e^{w^2/2}dv\right] + \frac{24}{\pi}\frac{\|g\|_\infty}{L} \\
&\leq c_1 d_{2,\mu}(F(\cdot|x),G(\cdot|x)) + c_2 L e^{L^2/2} d_{2,\mu}(F(\cdot|x),G(\cdot|x)) + c_3 L^{-1},
\end{aligned} \tag{99}$$

where c_1, c_2, c_3 are suitable finite constants and we note that

$$\int e^{w^2}dv = \sum_{n=0}^{\infty}\frac{w^{2n+1}}{n!(2n+1)} \leq w\sum_{n=0}^{\infty}\frac{w^{2n}}{n!} = we^{w^2}.$$

Taking $L = \sqrt{-\log(d_{2,\mu}(F(\cdot|x),G(\cdot|x)))}$ if $d_{2,\mu} \leq 1$ and $L = \sqrt{\log d_{2,\mu}(F(\cdot|x),G(\cdot|x))}$ otherwise in eqs. (98) and (99) yields eq. (96).

To prove eq. (97), integrate with respect to $x \in I$ in eq. (99) to obtain

$$\int \sup_{v\in\mathbb{R}}|F(v|x)-G(v|x)|\pi(dx) \leq c_1 d^{\pi}_{2,\mu}(F,G) + c_2 L e^{L^2/2} d^{\pi}_{2,\mu}(F,G) + c_3\pi(I)L^{-1}.$$

Proceeding as above, we obtain eq. (97). □

Proof of lemma 5.3. The results follow immediately from lemmas J.1, J.2, and J.3. □

K Sieve Properties

In lemmas K.1 and K.2 we establish the approximation properties of the sieves introduced in section 3.2 in terms of moment metrics.

Lemma K.1. *Let Ψ satisfy assumption 5.3 and $F_0(\cdot|x_0)$ satisfy assumption 5.2. Let p be a positive integer and $\delta > 0$. Let the interval $[-\delta p/2, \delta p/2]$ be partitioned into p equal-length closed intervals of length δ, labeled $B_{j,p}$, $j = 1, \ldots, p$ from left to right. Let $v_{j,p}$ be the center of $B_{j,p}$. Let $\sigma_p > 0$. Define the class $\mathcal{L}_{p,\sigma_p,\delta}$ as*

$$\mathcal{L}_{p,\sigma_p,\delta} = \left\{ \Lambda^{(x_0)}_{p,\sigma_p,\delta}(v|\boldsymbol{\gamma}) = \sum_{j=1}^{p} \gamma_j \Psi\left(\frac{v - v_{j,p}}{\sigma_p}\right), \quad \sum_{j=1}^{p} \gamma_j = 1, \quad \gamma_j \geq 0 \right\}.$$

Then if $\delta p \to \infty$, there exists a function $\Pi^{(x_0)}_{p,\sigma_p,\delta} F_0 \in \mathcal{L}_{p,\sigma_p,\delta}$ such that

$$d_{2,\mu}\left(\Pi^{(x_0)}_{p,\sigma_p,\delta} F_0, F_0(\cdot|x_0)\right) = O\left(\max\left\{\delta e^{(\delta p/2)^2/2}, \sigma_p(\delta p/2 + \sigma_p) e^{(\delta p/2 + \sigma_p)^2/2}\right\}\right). \tag{100}$$

In particular, if

$$\delta = \frac{\sqrt{\log(\log(p))}}{p}, \quad \sigma_p = \frac{1}{p\sqrt{\log(\log(p))} e^{\sqrt{\log(\log(p))}}},$$

then

$$d_{2,\mu}\left(\Pi^{(x_0)}_{p,\sigma_p,\delta} F_0, F_0(\cdot|x_0)\right) = O\left(\frac{\log(p)}{p}\right)$$

Proof. Define the function $\Pi^{(x_0)}_{p,\sigma_p,\delta} F_0 : \mathbb{R} \to [0, 1]$ as

$$\left(\Pi^{(x_0)}_{p,\sigma_p,\delta} F_0\right)(v) = \sum_{j=1}^{p} F_0(B_{j,p}|x_0) \Psi\left(\frac{v - v_{j,p}}{\sigma_p}\right). \tag{101}$$

Observe that by definition of the $B_{j,p}$, the mixture weights in eq. (101) are non-negative. Further, observe that $B_{j,p}$ may only intersect at their endpoints. As $F_0(v|x_0)$ is continuous in v under assumption 5.2, the weights sum to 1. Thus, $\Pi^{(x_0)}_{p,\sigma_p,\delta} F_0 \in \mathcal{L}_{p,\sigma_p,\delta}$ for any $p, \sigma_p > 0, \delta > 0$.

We first derive a representation for moments of $\Pi^{(x_0)}_{p,\sigma_p,\delta} F_0$ and two useful related inequalities.Let $\psi \coloneqq \Psi'$. The kth moment of $\Pi^{(x_0)}_{p,\sigma_p,\delta} F_0$ is given by

$$\sum_{j=1}^{p} F_0(B_{j,p}|x_0) \int v^k \frac{1}{\sigma_p} \psi\left(\frac{v - v_{j,p}}{\sigma_p}\right) dv$$

$$
\begin{aligned}
&= \sum_{j=1}^{p} F_0(B_{j,p}|x_0) \int ((v - v_{j,p}) + v_{j,p})^k \frac{1}{\sigma_p} \psi\left(\frac{t - v_{j,p}^p}{\sigma_p}\right) dv \\
&= \sum_{j=1}^{p} F_0(B_{j,p}|x_0) \sum_{i=0}^{k} \binom{k}{i} v_{j,p}^i \int (v - v_{j,p})^{k-i} \frac{1}{\sigma_p} \psi\left(\frac{v - v_{j,p}}{\sigma_p}\right) dv \\
&= \sum_{j=1}^{p} F_0(B_{j,p}|x_0) \sum_{i=0}^{k} \binom{k}{i} v_{j,p}^i \sigma_p^{k-i} \int v^{k-i} \psi(v) dv \\
&= \sum_{j=1}^{p} F_0(B_{j,p}|x_0) \sum_{i=0}^{k} \binom{k}{i} v_{j,p}^i \sigma_p^{k-i} \mu_{k-i,\psi}, \qquad (102)
\end{aligned}
$$

where $\mu_{j,\psi} = \int v^j \psi(v) dv$ is the j moment of Ψ.

Consider the terms with $i = k$ in eq. (102). By theorem 6 in Dragomir (2000):

$$
\left| \sum_{j=1}^{p} F_0(B_{j,p}|x_0) v_{j,p}^k - \int_{-\delta p/2}^{\delta p/2} v^k F_0(dt|x_0) \right| \leq C_{f_0} \delta (\delta p)^k \bigvee_{-\delta p/2}^{\delta p/2} v^k = 2 C_{f_0} \delta \left(\frac{\delta p}{2}\right)^k, \qquad (103)
$$

where $C_{f_0} = \sup_{x,v} |F_0'(v|x)| < \infty$ by assumption 5.2 and $\bigvee_a^b g$ is the total variation of the function g over $[a, b]$. Further, observe that eventually $\int_{-\delta p/2}^{\delta p/2} v^k F_0(dv|x_0) = \mu_k(x_0)$ as $\delta p \to \infty$ by the assumption of the lemma.

Now consider all the terms with $i \neq k$ in eq. (102). First we note that

$$
\begin{aligned}
&\left| \sum_{i=0}^{k-1} \binom{k}{i} v_{j,p}^i \sigma_p^{k-i} \mu_{k-i,\psi} \right| \leq \sum_{i=0}^{k-1} \binom{k}{i} |v_{j,p}|^i \sigma_p^{k-i} = \sigma_p \sum_{i=0}^{k-1} \frac{k}{k-i} \binom{k-1}{i} |v_{j,p}|^i \sigma_p^{k-1-i} \\
&\leq k \sigma_p (|v_{j,p}| + \sigma_p)^{k-1} \leq k \sigma_p \left(\frac{\delta p}{2} + \sigma_p\right)^{k-1} \qquad (104)
\end{aligned}
$$

where in the first inequality we use the fact that $\text{supp}(\Psi) \subset [-1, 1]$ and the last one that $|v_{j,p}| \leq \delta p/2$ by definition.

Now we turn to bounding the distance between $\Pi^{(x_0)}_{p,\sigma_p,\delta} F_0$ and $F_0(\cdot|x_0)$. By eq. (102):

$$
\begin{aligned}
&d_{2,\mu}^2 \left(\Pi^{(x_0)}_{p,\sigma_p,\delta} F_0, F_0(\cdot|x_0) \right) \\
&= \sum_{k=1}^{\infty} \frac{1}{k!} \left[\sum_{j=1}^{p} F_0(B_{j,p}|x_0) \int v^k \frac{1}{\sigma_p} \psi\left(\frac{v - v_{j,p}}{\sigma_p}\right) dv - \mu_k(x_0) \right]^2 \\
&= \sum_{k=1}^{\infty} \frac{1}{k!} \left[\sum_{j=1}^{p} F_0(B_{j,p}|x_0) \sum_{i=0}^{k} \binom{k}{i} v_{j,p}^i \sigma_p^{k-i} \mu_{k-i,\psi} - \mu_k(x_0) \right]^2
\end{aligned}
$$

$$= \sum_{k=1}^{\infty} \frac{1}{k!} \left[\sum_{j=1}^{p} F_0(B_{j,p}|x_0) v_{j,p}^k - \mu_k(x_0) + \sum_{j=1}^{p} F_0(B_{j,p}|x_0) \sum_{i=0}^{k-1} \binom{k}{i} v_{j,p}^i \sigma_p^{k-i} \mu_{k-i,\psi} \right]^2$$

$$\leq \sum_{k=1}^{\infty} \frac{2}{k!} \left[F_0(B_{j,p}|x_0) v_{j,p}^k - \mu_k(x_0) \right]^2 + \sum_{k=1}^{\infty} \frac{2}{k!} \left[\sum_{j=1}^{p} F_0(B_{j,p}|x_0) \sum_{i=0}^{k-1} \binom{k}{i} v_{j,p}^i \sigma_p^{k-i} \mu_{k-i,\psi} \right]^2$$

By eq. (103) it holds that

$$\sum_{k=1}^{\infty} \frac{2}{k!} \left[F_0(B_{j,p}|x_0) v_{j,p}^k - \mu_k(x_0) \right]^2 \leq \sum_{k=1}^{\infty} \frac{2C_{f_0}^2}{k!} \delta^2 \left(\frac{\delta p}{2} \right)^{2k} = 2C_{f_0}^2 \delta^2 e^{(\delta p/2)^2}. \tag{105}$$

By eq. (104) we obtain

$$\sum_{k=1}^{\infty} \frac{1}{k!} \left[\sum_{j=1}^{p} F_0(B_{j,p}|x_0) \sum_{i=0}^{k-1} \binom{k}{i} v_{j,p}^i \sigma_p^{k-i} \mu_{k-i,\psi} \right]^2 \leq \sum_{k=1}^{\infty} \frac{1}{k!} k^2 \sigma_p^2 \left(\frac{\delta p}{2} + \sigma_p \right)^{2k-2}$$

$$= \sum_{k=1}^{\infty} \frac{k}{(k-1)!} \sigma_p^2 \left(\frac{\delta p}{2} + \sigma_p \right)^{2k-2}$$

$$= \sum_{k=1}^{\infty} \frac{(k-1)}{(k-1)!} \sigma_p^2 \left(\frac{\delta p}{2} + \sigma_p \right)^{2k-2} + \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \sigma_p^2 \left(\frac{\delta p}{2} + \sigma_p \right)^{2k-2}$$

$$= \sum_{k=2}^{\infty} \frac{1}{(k-2)!} \sigma_p^2 \left(\frac{\delta p}{2} + \sigma_p \right)^{2k-2} + \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \sigma_p^2 \left(\frac{\delta p}{2} + \sigma_p \right)^{2k-2}$$

$$= \sigma_p^2 (\delta p + \sigma_p)^2 \sum_{k=0}^{\infty} \frac{(\delta p/2 + \sigma_p)^{2k}}{k!} + \sigma_p^2 \sum_{k=0}^{\infty} \frac{(\delta p/2 + \sigma_p)^{2k}}{k!}$$

$$\leq 2\sigma_p^2 (\delta p/2 + \sigma_p)^2 e^{(\delta p/2 + \sigma_p)^2}. \tag{106}$$

Eq. (100) follows from eqs. (105) and (106). □

Lemma K.2. *Let Ψ satisfy assumption 5.3 and let F_0 satisfy assumption 5.2. Let p_v, p_x be positive integers; let δ and σ_p be positive numbers. Let the interval $[-\delta p_v/2, \delta p_v/2]$ be partitioned into p_v equal-length closed intervals of length δ, labeled B_{j,p_v}, $j = 1, \ldots, p_v$ from left to right. Let v_{j,p_v} be the center of B_{j,p_v}. Define the class $\mathcal{L}^I_{(p_v,\delta,\sigma_p),p_x}$ as*

$$\mathcal{L}^I_{(p_v,\delta,\sigma_p),p_x} = \left\{ \Lambda_{(p_v,\delta,\sigma_p),p_x}(v|x,\boldsymbol{\gamma}) = \sum_{j=1}^{p_v} \sum_{l=0}^{p_x} \gamma_{j,l} b_{l,p_x}(x) \Psi \left(\frac{v - v_{j,p_v}}{\sigma_p} \right), \gamma_{jl} \geq 0, \sum_{j=1}^{p_v} \gamma_{j,l} = 1 \ \forall l \right\},$$

$$b_{l,p_x}(x) = \binom{p_x}{l} \left(\frac{x - x_{lb}}{x_{ub} - x_{lb}} \right)^l \left(\frac{x_{ub} - x}{x_{ub} - x_{lb}} \right)^{p_x - l}.$$

Then:

(1) For each valid vector $\boldsymbol{\gamma}$ and each $x \in I$ the function $\Lambda_{(p_v,\delta,\sigma_p),p_x}(v|x,\boldsymbol{\gamma})$ is a cdf. Support

of $\Lambda_{(p_v,\delta,\sigma_p),p_x}(v|x,\boldsymbol{\gamma})$ *is bounded uniformly in* x

(2) *Let* π *be a finite measure on* I*. If* $\delta p_v \to \infty$*, there exists* $\Pi_{(p_v,\delta,\sigma_p),p_x}F_0(v|x) \in \mathcal{L}^I_{(p_v,\delta,\sigma_p),p_x}$ *such that*

$$\begin{aligned}
&d^{\pi}_{2,\mu}\left(\Pi_{(p_v,\delta,\sigma_p),p_x}F_0(v|x), F_0(v|x)\right) \\
&\leq \frac{1}{\pi(I)}\int d^2_{2,\mu}\left(\Pi_{(p_v,\delta,\sigma_p),p_x}F_0(v|x), F_0(v|x)\right)\pi(dx) \\
&= O\left(\max\left\{\delta e^{(\delta p/2)^2/2}, p_x^{-1}e^{(\delta p/2+2)^2/2}, \sigma_p(\delta p/2+\sigma_p)e^{(\delta p/2+\sigma_p)^2/2}\right\}\right)
\end{aligned}$$

(3) *In particular, if*

$$p_v = p_x, \quad \delta = \frac{\sqrt{\log(\log(p_v))}}{p_v}, \quad \sigma_p = \frac{1}{p\sqrt{\log(\log(p_v))}e^{\sqrt{\log(\log(p_v))}}},$$

then

$$\int d^2_{2,\mu}\left(\Pi_{(p_v,\delta,\sigma_p),p_x}F_0(v|x), F_0(v|x)\right)\pi(dx) = O\left(\frac{\log(p_v)}{p_v}\right).$$

Proof. We begin with the first assertion. First, $\sum_{l=0}^{p_x}\gamma_{j,l}b_{l,p_x}(x) \geq 0$ for all $j = 1,\ldots,p_v$ as $\gamma_{j,l} \geq 0$ and for any $x \in I$, $b_{l,p_x}(x) \geq 0$. Second, the mixture weights sum to unity: since $\sum_{l=0}^{p_x} b_{l,p_x}(x) = 1$ for each $x \in I$, it holds that

$$\sum_{j=1}^{p_v}\left(\sum_{l=0}^{p_x}\gamma_{j,l}b_{l,p_x}(x)\right) = \sum_{l=0}^{p_x}b_{l,p_x}(x)\sum_{j=1}^{p_v}\gamma_{j,l} = \sum_{l=0}^{p_x}b_{l,p_x}(x) = 1.$$

Define

$$\Pi_{(p_v,\delta,\sigma_p),p_x}F_0(v|x) = \sum_{j=1}^{p_v}\sum_{l=0}^{p_x}F_0\left(B_{j,p_v}\middle|x_{lb} + \frac{l(x_{ub}-x_{lb})}{p_x}\right)b_{l,p_x}(x)\Psi\left(\frac{t-v_{j,p_v}}{\sigma_p}\right),$$

where x_{lb} and x_{ub} are the endpoints of I. Observe that by construction $\Pi_{(p_v,\delta,\sigma_p),p_x}F_0(v|x) \in \mathcal{L}^I_{(p_v,\delta,\sigma_p),p_x}$, as the intervals B_{j,p_v} may only interest at their endpoints and $F_0(v|x)$ is continuous in t for each x.

To prove approximation properties, we proceed similarly to lemma K.1. The kth moment of the approximation at x is given by

$$\begin{aligned}
&\sum_{j=1}^{p_v}\sum_{l=0}^{p_x}F_0\left(B_{j,p_v}\middle|x_{lb} + \frac{l(x_{ub}-x_{lb})}{p_x}\right)b_{l,p_x}(x)\int v^k\frac{1}{\sigma_p}\psi\left(\frac{v-v_{j,p_v}}{\sigma_p}\right)dv \\
&= \sum_{j=1}^{p_v}\sum_{l=0}^{p_x}F_0\left(B_{j,p_v}\middle|x_{lb} + \frac{l(x_{ub}-x_{lb})}{p_x}\right)b_{l,p_x}(x)\sum_{i=0}^{k}\binom{k}{i}v^i_{j,p_v}\sigma_p^{k-i}\mu_{k-i,\psi}, \qquad (107)
\end{aligned}$$

We being by considering the term corresponding to $i = k$ in the third sum. This term targets

$\mu_k(x)$. Further, the difference may be decomposed as

$$\sum_{j=1}^{p_v} v_{j,p_v}^k \left[\sum_{l=0}^{p_x} F_0\left(B_{j,p_v} \middle| x_{lb} + \frac{l(x_{ub} - x_{lb})}{p_x} \right) b_{l,p_x}(x) - F_0\left(B_{j,p_v} \middle| x \right) \right] + \sum_{j=1}^{p_v} F_0\left(B_{j,p_v} \middle| x \right) v_{j,p_v}^k - \mu_k(x). \tag{108}$$

First, as in lemma K.1, by theorem 6 of Dragomir (2000) we obtain that

$$\sup_{x \in I} \left| \sum_{j=1}^{p_v} F_0\left(B_{j,p_v} \middle| x \right) v_{j,p_v}^k - \mu_k(x) \right| \leq C_{f_0} \delta \left(\frac{\delta p}{2} \right)^k,$$

for $C_{f_0} = \sup_{v \in \mathbb{R}, x \in I} F_0(v|x)$.

Second, we consider the leading sum in eq. (108). By assumption 5.2, the function (of x) $F_0(B_{j,p_v}|x) = \int_{B_{j,p_v}} F_0(v|x)dv$ is twice-differentiable with a bounded continuous second derivative on I. Label the endpoints of B_{j,p_v} as $B_{j,p_v} = [w_{j-1,p_v}, w_{j,p_v}]$ and recall that $w_{j,p_v} - w_{j-1,p_v} = \delta$. We can estimate the second derivative of $F_0(B_{j,p_v}|x)$ as

$$\left| \partial_x^2 F_0(B_{j,p_v}|x) \right| = \left| \delta \frac{1}{w_{j,p_v} - w_{j-1,p_v}} \int_{w_{j-1,p_v}}^{w_{j,p_v}} \partial_x F_0(v|x) dv \right| \leq C_{\partial_x f_0} \delta.$$

where $C_{\partial_x F_0} = \sup_{v,x} |\partial_x F_0(v|x)|$. Then from theorem 4.29 of Bustamante (2017) it follows that for some $C_\Pi < \infty$

$$\max_{x \in I} \left| \sum_{l=0}^{p_x} F_0\left(B_{j,p_v} \middle| x_{lb} + \frac{l(x_{ub} - x_{lb})}{p_x} \right) b_{l,p_x}(x) - F_0(B_{j,p_v}|x) \right| \leq \frac{C_\Pi \delta}{p_x}.$$

It follows that

$$\left| \sum_{j=1}^{p_v} v_{j,p_v}^k \sum_{l=0}^{p_x} \left[F_0\left(B_{j,p_v} \middle| x_{lb} + \frac{l(x_{ub} - x_{lb})}{p_x} \right) b_{l,p_x}(x) - F_0(x) \right] \right| \leq \frac{C_\Pi \delta}{p_x} \sum_{j=1}^{p_v} |v_{j,p_v}|^k \leq \frac{C_\pi \delta^{k+1}}{p_x} \frac{(\lceil \frac{p_v}{2} \rceil + 1)^{k+1}}{k+1},$$

where we observe that

$$\sum_{j=1}^{p_v} |v_{j,p_v}|^k \leq 2 \sum_{j=1}^{\lceil p_v/2 \rceil} ((j+1)\delta)^k = 2\delta^k \sum_{j=1}^{\lceil p_v/2 \rceil} (j+1)^k \leq 2\delta^k \int_1^{\lceil p_v/2 \rceil + 1} (x+1)^k \leq 2\delta^k \frac{(\lceil \frac{p_v}{2} \rceil + 1)^{k+1}}{k+1}.$$

Now consider the sum over i from 0 to $k-1$ in eq. (107). By eq. (104)

$$\left|\sum_{i=0}^{k-1}\binom{k}{i}v_{j,p}^{i}\sigma_p^{k-i}\mu_{k-i,\psi}\right| \leq k\sigma_p\left(\frac{\delta p_v}{2}+\sigma_p\right)^{k-1}.$$

Correspondingly,

$$\begin{aligned}
&\left|\sum_{j=1}^{p_v}\sum_{l=0}^{p_x}F_0\left(B_{j,p_v}\middle|x_{lb}+\frac{l(x_{ub}-x_{lb})}{p_x}\right)b_{l,p_x}(x)\sum_{i=0}^{k-1}\binom{k}{i}v_{j,p_v}^{i}\sigma_p^{k-i}\mu_{k-i,\psi}\right| \\
&\leq k\sigma_p\left(\frac{\delta p_v}{2}+\sigma_p\right)^{k-1}\sum_{j=1}^{p_v}\sum_{l=0}^{p_x}F_0\left(B_{j,p_v}\middle|x_{lb}+\frac{l(x_{ub}-x_{lb})}{p_x}\right)b_{l,p_x}(x) \\
&= k\sigma_p\left(\frac{\delta p_v}{2}+\sigma_p\right)^{k-1}.
\end{aligned}$$

Then it holds that

$$\begin{aligned}
&\int d_{2,\mu}^2\left(\Pi_{(p_v,\delta,\sigma_p),p_x}F_0(v|x),F_0(v|x)\right)\pi(dx) \\
&\leq 8\int_I\sum_{k=0}^{\infty}\frac{1}{k!}\left[\left[C_{f_0}\delta\left(\frac{\delta p_v}{2}\right)^k\right]^2+\left[\frac{C_\pi\delta^{k+1}}{p_x}\frac{(\lceil\frac{p_v}{2}\rceil+1)^{k+1}}{k+1}\right]^2+\left[k\sigma_p\left(\frac{\delta p_v}{2}+\sigma_p\right)^{k-1}\right]^2\right]\pi(dx) \\
&\leq 8\pi(I)\left[2C_{f_0}^2\delta^2e^{(\delta p_v/2)^2}+e^{(\delta p_v/2+2)^2}\frac{C_\pi^2}{p_x^2}+2\sigma_p^2(\delta p_v/2+\sigma_p)^2e^{(\delta p_v/2+\sigma_p)^2}\right].
\end{aligned}$$

The equality in (2) follows by the assumption that $\pi(I)<\infty$. To see the inequality in (2), observe that by Jensen's inequality for any $F,G\in\mathcal{F}^I$ it holds that

$$\left[d_{2,\mu}^{\pi}(F,G)\right]^2=\left[\int d_{2,\mu}(F(\cdot|x),G(\cdot|x))\pi(dx)\right]^2\leq\frac{1}{\pi(I)}\int_I d_{2,\mu}^2(G(\cdot|x),G(\cdot|x))\pi(dx). \quad (109)$$

(3) can be obtained by substituting the proposed values for p_v,p_x,δ,σ_p. □

L Proof of Theorem 5.4

We will use the following high-level lemma to establish the consistency (theorem 5.4) of the distribution estimators. The lemma be viewed as a modification and amalgamation of lemmas A.2 and B.1 of Chen and Pouzo (2012).

Lemma L.1. *Let $\tilde{Q}_N(\cdot|x_0)$ and $Q(\cdot|x_0)$ be defined as in eqs. (28) and (29), respectively. Let $\mathcal{L}_p$ be the pth sieve space defined in eq. (25) and $\mathcal{F}$ be as in assumption 5.2, both equipped with the metric topology induced by $d_{2,\mu}$ and the corresponding Borel σ-algebra. Let*

(1) $p = p(N)$ *be a non-decreasing sequence such that* $p \to \infty$ *as* $N \to \infty$.
(2) $\sup_{F \in \mathcal{L}_p} \left| \tilde{Q}_N(F|x_0) - Q(F|x_0) \right| \leq \eta_{Q,N}$ *where* $\eta_{Q,N} = O_{a.s.}(\delta_{Q,N})$ *for a deterministic sequence* $\delta_{Q,N} = o(1)$.
(3) $\tilde{F}_N(v|x_0) = \arg\min_{F \in \mathcal{L}_p} \left[\tilde{Q}_N(F|x_0) + \lambda_N P_p(F) \right]$ *for some non-negative* λ_N *and a non-negative penalty function* $P_p : \mathcal{L}_p \to \mathbb{R}_+$.
(4) There exists an approximating function $\Pi_p^{(x_0)} F_0 \in \mathcal{L}_p$ *such that* $d_{2,\mu}(\Pi_p^{(x_0)} F_0, F_0) \to 0$.
(5) $\lambda_N P_p \left(\Pi_p^{(x_0)} F_0 \right) \to 0$.
(6) $\tilde{F}_N(v|x_0)$ *is a measurable function of the data.*
Then

$$
\begin{aligned}
&d_{2,\mu}(\tilde{F}_N(\cdot|x_0), F_0(\cdot|x_0)) \\
&= O_{a.s.}\left(\max\left\{ d_{2,\mu}\left(\Pi_p^{(x_0)} F_0, F_0 \right), \sqrt{\delta_{Q,N}}, \sqrt{\lambda_N P_p\left(\Pi_p^{(x_0)} F_0 \right)} \right\} \right) = o_{a.s.}(1).
\end{aligned}
$$

Proof. Note that by definition $\mathcal{L}_p \subset \mathcal{F}$ for any p, and so both $Q(\hat{F}_N^{(x_0)}|x_0)$ and $Q(\Pi_p^{(x_0)} F_0|x_0)$ are defined.

Consider the following chain of inequalities:

$$
\begin{aligned}
Q(\hat{F}_N^{(x_0)}|x_0) &\leq \tilde{Q}_N(\tilde{F}_N(\cdot|x_0)|x_0) + \eta_{Q,N} + \lambda_N P_p(\tilde{F}_N(\cdot|x_0)) \\
&\leq \tilde{Q}_N(\Pi_p^{(x_0)} F_0) + \eta_{Q,N} + \lambda_N P_p(\Pi_p^{(x_0)} F_0|x_0) \\
&\leq Q(\Pi_p^{(x_0)} F_0|x_0) + 2\eta_{Q,N} + \lambda_N P_p(\Pi_p^{(x_0)} F_0).
\end{aligned}
$$

where the first inequality holds by assumption of uniform convergence of $\tilde{Q}_N(\cdot|x_0)$ to $Q(\cdot|x_0)$ over $\mathcal{L}_p$ and non-negativity of the penalty term; second inequality by definition of $\tilde{F}_N(v|x_0)$ and the fact that $\Pi_p^{(x_0)} F_0 \in \mathcal{L}_p$; and the last inequality by applying uniform convergence again. Finally, recall that $Q(F|x_0) = d_{2,\mu}^2(F, F_0(\cdot|x_0))$ by eq. (29). Further, $Q(\cdot|x_0)$ is continuous on $\mathcal{F}$ (as the square of the metric generating the topology of $\mathcal{F}$), and thus measurable. The result follows. □

An analogous result holds in the interval case.

Lemma L.2. *Let* $\hat{Q}_N$ *and* Q *be defined as in eqs. (30) and (31), respectively. Let* $\mathcal{L}^I_{p_v,p_x}$ *be the* (p_v, p_x)*th sieve space defined in eq. (27) and* $\mathcal{F}^I$ *be as in assumption 5.2, both equipped with the metric topology induced by* $d^\pi_{2,\mu}$ *and the corresponding Borel* σ*-algebra for* $d^\pi_{2,\mu}$. *Let*
(1) $p_v = p_v(N), p_x = p_x(N)$ *be non-decreasing sequences such that* $p_v, p_x \to \infty$ *as* $N \to \infty$.
(2) $\sup_{F \in \mathcal{L}^I_{p_v,p_x}} \left| \hat{Q}_N(F) - Q(F) \right| \leq \eta^I_{Q,N}$ *where* $\eta^I_{Q,N} = O_{a.s.}(\delta^I_{Q,N})$ *for a deterministic sequence that satisfies* $\delta_{Q,N} = o(1)$.

(3) $\hat{F}_N = \arg\min_{F \in \mathcal{L}^I_{p_v,p_x}} \left[\hat{Q}_N(F) + \lambda^I_N P^I_p(F)\right]$ *for some non-negative* λ^I_N *and a non-negative penalty function* $P^I_{p_v,p_x} : \mathcal{L}^I_{p_v,p_x} \to \mathbb{R}_+$.

(4) There exists some $\Pi_{p_v,p_x} F_0 \in \mathcal{L}_{p_v,p_x}$ *such that* $\int d^2_{2,\mu}(\Pi_{p_v,p_x} F_0(\cdot|x), F_0(\cdot|x)) \times \pi(dx) = O(\delta_{\Pi,N})$ *for* $\delta_{\Pi,N} = o(1)$.

(5) $\lambda^I_N P^I_{p_v,p_x}(\Pi_{p_v,p_x} F_0) \to 0$.

(6) $\hat{F}_N$ *is a measurable function of the data.*

Then

$$d^\pi_{2,\mu}(\hat{F}_N, F_0) = O_{a.s.}\left(\max\left\{\sqrt{\delta_{\Pi,N}}, \sqrt{\delta^I_{Q,N}}, \sqrt{\lambda^I_N P^I_{p_v,p_x}(\Pi_{p_v,p_x} F_0)}\right\}\right) = o_{a.s.}(1).$$

Proof. Proceeding as in lemma L.1, we obtain that

$$Q(\hat{F}_N) \leq Q(\Pi_{p_v,p_x} F_0) + 2\eta^I_{Q,N} + \lambda^I_N P^I_{p_v,p_x}(\Pi_{p_v,p_x} F_0).$$

Observe that $Q(F) = \int_I d^2_{2,\mu}(F(\cdot|x), G(\cdot,x))\pi(dx)$ for any $F \in \mathcal{F}^I$. It then holds that $d^\pi_{2,\mu}(F) \leq \sqrt{Q(F)}$ by eq. (109). The conclusion follows. □

The following lemma shows that the sample objective functions converge to the population objective functions, verifying conditions (2) of lemmas L.1 and L.2.

Lemma L.3. *Let assumptions of theorem 5.4 hold. Let* $\tilde{Q}_N(\cdot|\cdot), \hat{Q}_N(\cdot)$, $Q(\cdot|\cdot)$, *ad* $Q(\cdot)$ *be defined as in eqs.* (28)-(31). *Let the sieve spaces* $\mathcal{L}_p, \mathcal{L}^I_{p_v,p_x}$ *be as in the eqs.* (25) *and* (27), *respectively. Then*

$$\sup_{F \in \mathcal{L}_p} \left|\tilde{Q}_N(F|x) - Q(F|x)\right| \xrightarrow{a.s.} 0, \quad \text{for any } x \in I \tag{110}$$

$$\sup_{F \in \mathcal{L}^I_{p_v,p_x}} \left|\hat{Q}_N(F) - Q(F)\right| \xrightarrow{a.s.} 0. \tag{111}$$

Proof. We establish eq. (111). The proof of eq. (110) is analogous, but easier. Define the functions $\hat{H}_N(F,x), H(F,x) : \mathcal{F}^I \times I \to \mathbb{R}$ as

$$\hat{H}_N(F,x) = \sum_{k=1}^{K-1} \frac{1}{k!}\left[\tilde{\mu}_k(x) - \int v^k F(dv|x)\right]^2,$$

$$H(F,x) = \sum_{k=1}^{\infty} \frac{1}{k!}\left[\mu_k(x) - \int v^k F(dv|x)\right]^2.$$

As in the proof of lemma J.1, observe that $H(F,x) < \infty$ for any $x \in I$ and $F \in \mathcal{F}^I$. As $\mathcal{L}^I_{p_v,p_x} \subset \mathcal{F}^I$ for all (p_v,p_x), H and $\hat{H}$ are automatically defined on $\mathcal{L}^I_{p_v,p_x}$. It holds $\hat{Q}_N(F) = \int_I \hat{H}_N(F,x)\pi(dx)$ and $Q(F) = \int H(F,x)\pi(dx)$.

Bound the difference of $\hat{H}_N(F,x)$ and $H(F,x)$ for given $F \in \mathcal{L}^I_{p_v,p_x}, x \in I$ as

$$\begin{aligned}
&\left|\hat{H}_N(F,x) - H(F,x)\right| \\
&\leq \left|\sum_{k=1}^{K-1} \frac{1}{k!}\left(\tilde{\mu}_k^2(x) - \mu_k^2(x)\right)\right| + 2\left|\sum_{k=1}^{K-1} \frac{1}{k!}\left[(\tilde{\mu}_k(x) - \mu_k(x))\int v^k F(dv|x)\right]\right| \\
&\quad + \left|\sum_{k=K+}^{\infty} \frac{1}{k!}\mu_k^2(x)\right| + 2\left|\sum_{k=K}^{\infty} \frac{1}{k!}\mu_k(x)\int v^k F(dv|x)\right| + \left|\sum_{k=K}^{\infty} \frac{1}{k!}\left[\int v^k F(dv|x)\right]^2\right|.
\end{aligned}$$

We will now show each sum tends to 0 a.s. uniformly over $F \in \mathcal{L}^I_{p_v,p_x}$ and $x \in I$ as $N \to \infty, p_v = p_v(N) \to \infty$, $p_x = p_x(N) \to \infty$, $K = K(N) \to \infty$.

Consider the first sum and observe that it does not depend on F. Consider the function $g_N(k)$ defined as $g_N(k) = (k!)^{-1}\sup_{x\in I}|\tilde{\mu}_k^2(x) - \mu_k^2(x)|$ for $k \leq K$ and $g_N(k) = 0$ for $k > K$. Define the event $B_k = \{g_N(k) \xrightarrow{a.s.} 0\}$. By the assumption of the theorem, $P(B_k) = 1$ for all k. On the event $B = \bigcap_{k=1}^{\infty} B_k$, the function g_N converges to the zero function pointwise. Observe that $|g_N(k)| \leq 2C_\mu^k(k!)^{-1/2}$ for all N by assumption 5.2 and definition of $\tilde{\mu}_k(x)$ (above eq. (32)). As $\sum_{k=0}^{\infty} C_\mu^k(k!)^{-1/2} < \infty$, by the dominated convergence theorem on B it holds that

$$\sup_{F\in\mathcal{L}^I_{p_v,p_x}} \max_{x\in I}\left|\sum_{k=1}^{K-1} \frac{1}{k!}\left(\tilde{\mu}_k^2(x) - \mu_k^2(x)\right)\right| \leq \sum_{k=1}^{\infty} g_N(k) \to 0.$$

Since $P(B) = 1$, convergence is a.s.

Now consider the last term. Observe the support of any $F \in \mathcal{L}^I_{p_v,p_x}$ satisfies $\text{supp}(F) \subset [-C_{\mathcal{F}}\log(p_v), C_{\mathcal{F}}\log(p_v)])$, where $C_{\mathcal{F}}$ is a constant that does not depend on F or (p_v, p_x). Thus, for any $F \in \mathcal{L}^I_{p_v,p_x}$ its kth conditional moment $\int v^k F(dv|x)$ lies in the interval $[-(C_{\mathcal{F}}\log(p_v))^k, (C_{\mathcal{F}}\log(p_v))^k]$ for all $x \in I$. It follows that

$$\begin{aligned}
\sup_{F\in\mathcal{L}^I_{p_v,p_x}} \sum_{k=K}^{\infty} \frac{1}{k!}\left[\int v^k F(dv|x)\right]^2 &\leq \sum_{k=K}^{\infty} \frac{(C_{\mathcal{F}}\log(p_v))^{2k}}{k!} \\
&= 2\left[e^{(C_{\mathcal{F}}\log(p_v))^2} - \sum_{k=0}^{K-1} \frac{(C_{\mathcal{F}}\log(p_v))^{2k}}{k!}\right] \\
&\lesssim \frac{(C_{\mathcal{F}}\log(p_v))^{2K}}{K!} e^{(C_{\mathcal{F}}\log(p_v))^2} \\
&\sim \frac{e^K (C_{\mathcal{F}}\log(p_v))^{2K} e^{(C_{\mathcal{F}}\log(p_v))^2}}{\sqrt{K}K^K},
\end{aligned}$$

where the second line follows from (1.1) in Pritsker and Varga (1997) and where we apply Stirling's approximation in the last line. As $p_v = o(\exp(K^{1/2}))$ by the assumption of the theorem, the above expression tends to zero.

The remaining terms can be handled by applying logic similar to that of the above arguments. Again using the fact that $\sup_{F\in\mathcal{F}^I_{p_v,p_x}}\int v^k F(dv|x) \leq 2(C_{\mathcal{F}}\log(p_v))^k$, we obtain the following bound for the second term:

$$\sup_{f\in\mathcal{L}^I_{p_v,p_x}}\max_{x\in I}\left|\sum_{k=1}^{K-1}\frac{1}{k!}\left[(\tilde{\mu}_k(x)-\mu_k(x))\int v^k F(v|x)dt\right]\right| \leq \sum_{k=1}^{K-1}\frac{|C_{\mathcal{F}}\log(p_v)|^k}{k!}\sup_{x\in I}|\tilde{\mu}_k(x)-\mu_k(x)|$$

Let the event B be as above. Each term on the right hand side converges to zero on B by the assumption that $(\log(p_v))^k = o(\delta_{k,N})$ for each k. By proceeding as with the first term, we now obtain that the this term converges to zero on B. For the third term, note that by assumption 5.2 the true distribution F_0 has bounded support, bounded uniformly in $x \in I$. Correspondingly, there exists some constant M such that $\text{supp}(F_0(v|x)) \subset [-M, M]$ for all $x \in I$. Then

$$\sup_{F\in\mathcal{L}^I_{p_v,p_x}}\sup_{x\in I}\sum_{k=K}^{\infty}\mu_k^2(x) \leq \sum_{k=K}^{\infty}\frac{M^{2k}}{k!} \lesssim \frac{M^{2K}}{K!}e^{M^2},$$

which tends to zero as $K \to \infty$.

Finally, for the fourth term use both the support bound on members of $\mathcal{L}^I_{p_v,p_x}$ and the bounded support assumption on F_0 to obtain

$$\sup_{F\in\mathcal{L}^I_{p_v,p_x}}\sup_{x\in I}\left|\sum_{k=K}^{\infty}\frac{1}{k!}\mu_k(x)\int v^k F(dv|x)\right| \leq \sum_{k=K}^{\infty}\frac{(MC_{\mathcal{F}})^k(\log(p_v))^k}{k!},$$

which tends to zero by the same argument as for the fifth term. □

Proof of theorem 5.4. Throughout, $\mathcal{F}$ and $\mathcal{L}_p$ are equipped with the topology generated by $d_{2,\mu}$; $\mathcal{F}^I$ and $\mathcal{L}^I_{p_v,p_x}$ are equipped with the topology generated by $d^{\pi}_{2,\mu}$.

Consider the first assertion. Consistency is established by verifying the conditions of lemma L.1:

(1) Holds by the assumption of the theorem.

(2) Holds by lemma L.3

(3) Identify $\mathcal{L}_p$ with the unit simplex Δ^p in $\mathbb{R}^p$ via the map $\Gamma_p : \mathcal{L}_p \to \Delta^p$ that sends $\sum_{j=1}^p \gamma_j \Psi\left(\sigma_p^{-1}(v - v_{j,p})\right) \to (\gamma_1, \ldots, \gamma_p) =: \boldsymbol{\gamma}$. Γ_p is well-defined. To see this, let $F \in \mathcal{L}_p$ and suppose that F can be represented by $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$, so that $\sum_{j=1}^p(\gamma_j - \delta_j)\Psi\left(\sigma_p^{-1}(v - v_{j,p})\right) = 0$. Then $\sum_{j=1}^p(\gamma_j - \delta_j)\int v^k\sigma_p^{-1}\psi\left(\sigma_p^{-1}(v - v_{j,p})\right)dv = 0$ for $k = 0, \ldots, p-1$. We show by induction that $\sum_{i=1}^v (v_{j,p})^k(\gamma_i - \delta_i) = 0$ for $k = 0, \ldots, p-1$. For $k = 0$

$$\sum_{j=1}^p(\gamma_j - \delta_j)\int t^0\sigma_p^{-1}\psi\left(\sigma_p^{-1}(v - v_{j,p})\right) = \left[\int \sigma_p^{-1}\psi\left(\sigma_p^{-1}(v - v_{j,p})\right)dt\right]\sum_{j=1}^p(\gamma_j - \delta_j) = 0$$

Since the integral in brackets is equal to 1, it holds that $\sum_{j=1}^{p}(\gamma_j-\delta_j)(v_{j,p})^0=0$. Suppose that we have shown that $\sum_{j=1}^{p}(\gamma_-\delta_j)(v_{j,p})^l=0$ for $l=0,\dots,k-1$. Then

$$\begin{aligned}
0&=\sum_{j=1}^{p}(\gamma_j-\delta_j)\int v^k\sigma_p^{-1}\psi\left(\sigma_p^{-1}(v-v_{j,p})\right)dv\\
&=\sum_{j=1}^{p}(\gamma_j-\delta_j)\int((v-v_{j,p})+v_{j,p})^k\sigma_p^{-1}\psi\left(\sigma_p^{-1}(v-v_{j,p})\right)dv\\
&=\sum_{j=1}^{p}(\gamma_j-\delta_j)\sum_{l=0}^{k}\binom{k}{l}(v_{j,p})^l\int(v-v_{j,p})^{k-l}\sigma_p^{-1}\psi\left(\sigma_p^{-1}(v-v_{j,p})\right)dv\\
&=\sum_{l=0}^{k}\binom{k}{l}\sigma_p^{k-l}\mu_{k-l,\psi}\sum_{j=1}^{p}(\gamma_j-\delta_j)(v_{j,p})^l\\
&=\mu_{0,\psi}\sum_{j=1}^{p}(\gamma_j-\delta_j)(v_{j,p})^k.
\end{aligned}$$

where $\mu_{j,\psi}\coloneqq\int v^j\psi(v)dv$, and the last equality holds by the inductive assumption. Since $\mu_{0,\psi}=1$, we conclude that $\sum_{i=1}^{p}(\gamma_j-\delta_j)(v_{j,p})^k=0$ as desired.

Treated as a system in $(\gamma_j-\delta_j)$, the equations $\sum_{i=1}^{v}(v_{j,p})^k(\gamma_i-\delta_i)=0$ for $k=0,\dots,p-1$ define a full-rank Vandermonde system of linear equations. It follows that $\gamma_j=\delta_j$.

For $F\in\mathcal{L}_p$, define $P_p(F)=\|\Gamma_p(F)\|_2^2$. $P_p(F)$ is well-defined since Γ_p is.

Then by eq. (18), $\tilde{F}_N(\cdot|x_0)$ minimizes $\tilde{Q}_N(F|x_0)+\lambda_N P_{p(N)}(F)$ over $\mathcal{L}_{p(N)}$.

(4) Holds by lemma K.1.

(5) By definition of $\mathcal{L}_p$, for any $F\in\mathcal{L}_p$ it holds that $\|\Gamma(F)\|_1=1$. Then $\|\Gamma(F)\|_2\le\|\Gamma(F)\|_1=1$, and hence $P_p(F)=\|\Gamma(F)\|_2^2\le 1$. Since $\lambda_N\to 0$ by the assumption of the theorem, $\lambda_N P_p\left(\Pi_p^{(x_0)}F_0\right)\to 0$.

(6) Each moment estimate $\tilde{\mu}_k(x_0)$ is a continuous function of the data. Then by a standard argument the estimated mixture coefficients $\tilde{\boldsymbol{\gamma}}$ of eq. (19) are a measurable function of the data (with respect to the Borel σ-algebra induced by the norm topology on $\Delta^{p(N)}$). Let Γ_p be as above and note that Γ_p is a bijection between $\mathcal{L}_p$ and Δ^p. Observe that $\tilde{F}_N(v|x_0)=\Gamma_p^{-1}(\tilde{\boldsymbol{\gamma}})$. The map $\Gamma_p^{-1}:\Delta^p\to\mathcal{L}_p$ is continuous if $\mathcal{L}_p$ is equipped with the L^1 topology.[10] Since the L^1 topology is stronger than the metric topology of $d_{2,\mu}$, Γ_p^{-1} is also continuous with respect to the latter topology. Measurability of $\tilde{F}_N(v|x_0)$ follows.

Thus, conditions of lemma L.1 hold and $d_{2,\mu}\left(\tilde{F}_N(\cdot|x_0),F_0(\cdot|x_0)\right)\xrightarrow{a.s.}0$. By lemma 5.3 we conclude that $\sup_{v\in\mathbb{R}}\left|\tilde{F}_N(v|x_0)-F(v|x_0)\right|\xrightarrow{a.s.}0$.

[10] Norm Δ^p with the ∞-norm. Then $\left\|\Gamma_p^{-1}(\boldsymbol{\delta})-\Gamma_p^{-1}(\boldsymbol{\gamma})\right\|_{L^1}$ is bounded by $\int\left|\sum_{=1}^{p}(\gamma_j-\delta_j)\sigma_p^{-1}\psi\left(\sigma_p^{-1}(t-v_{j,p})\right)\right|\le\sum_{i=1}^{p}|\gamma_j-\delta_j|\int\sigma_p^{-1}\psi\left(\sigma_p^{-1}(t-v_{j,p})\right)dt\le p\left\|\boldsymbol{\gamma}-\boldsymbol{\delta}\right\|_\infty$.

The second assertion follows analogously by lemma L.2. We highlight the relevant changes:

(3) To replace Γ_p let $\mathcal{L}^I_{p_v,p_x}$ be as in eq. (27). Identify $\mathcal{L}^I_{p_v,p_x}$ with $(\Delta^{p_v})^{p_x}$ via the map $\Gamma_{p_v,p_x} : \sum_{j=1}^{p_v}\sum_{l=0}^{p_x}\gamma_{j,l}b_{l,p_x}(x)\sigma_p^{-1}\psi\left(\sigma_p^{-1}(v-v_{j,p_v})\right) \to \boldsymbol{\gamma}$. Γ_{p_v,p_x} is well-defined. To see this, suppose that the same $F \in \mathcal{L}^I_{p_v,p_x}$ can be represented by $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, that is, that

$$\sum_{j=1}^{p_v}\sum_{l=0}^{p_x}(\gamma_{j,l}-\delta_{j,l})b_{l,p_x}(x)\Psi\left(\sigma_p^{-1}(v-v_{j,p_v})\right) = 0.$$

Evaluate $\int v^k \sum_{j=1}^{p_v}\sum_{l=0}^{p_x}(\gamma_{j,l}-\delta_{j,l})b_{l,p_x}(x)\sigma_p^{-1}\psi\left(\sigma_p^{-1}(v-v_{j,p_v})\right)dv$ for $k=0,\dots,p_v-1$. As above, we obtain that $\sum_{j=1}^{p_v}\sum_{l=0}^{p_x}(\gamma_{j,l}-\delta_{j,l})b_{l,p_x}(x)v_{j,p_v}^k = 0$. Rearranging,

$$\sum_{l=0}^{p_x} b_{l,p_x}(x)\sum_{j=1}^{p_v}(\gamma_{j,l}-\delta_{j,l})v_{j,p_v}^k = 0, \quad k=0,\dots,p_v-1.$$

Since $\{b_{0,p_x},\dots,b_{p_x,p_x}\}$ is a linearly independent collection of functions, we conclude that for each $l=0,\dots,p_x$ it holds that $\sum_{j=1}^{p_v}(\gamma_{j,l}-\delta_{j,l})v_{j,p_v}^k = 0$. This is a full-rank Vandermonde system of equations in $(\gamma_{1,l}-\delta_{1,l},\dots,\gamma_{p_v,l}-\delta_{p_v,l})$. It has the unique solution $\gamma_{j,l}=\delta_{j,l}$ for $j=1,\dots,p_v$. The same holds for all $l=0,\dots,p_x$. Thus $\boldsymbol{\gamma}=\boldsymbol{\delta}$.
Now we define a suitable penalty. Define $P^I_{p_v,p_x}(F) = \sum_{j=1}^{p_v}\sum_{l=0}^{p_x}\gamma_{j,l}^2$ where $\gamma_{j,l}$ be the (j,l)th element of $\Gamma_{p_v,p_x}(F)$.

(4) We use lemma K.2 in place of lemma K.1.

(5) Let $F \in \mathcal{L}^I_{p_v,p_x}$. Let $\gamma_{j,l} = \Gamma_{p_v,p_x}(F)$ for $f \in \mathcal{L}^I_{p_v,p_x}$. Observe that $\gamma_{j,l} \in [0,1]$ for all j,l. Then $P^I_{p_v,p_x}(F) = \sum_{j=1}^{p_v}\sum_{l=0}^{p_x}\gamma_{j,l}^2 \le \sum_{j=1}^{p_v}\sum_{l=0}^{p_x}\gamma_{j,l} = p_x$ where the final equality holds by definition of $\mathcal{L}^I_{p_v,p_x}$. Now by the assumption on $\lambda_N^I = o(p_x^{-1})$ we conclude that $\sup_{F\in\mathcal{L}_{p_v,p_x}}\lambda_N^I P^I_{p_v,p_x}(F) \to 0$.

By lemma L.2 $d^{\pi}_{2,\mu}\left(\hat{F}_N, F_0\right) \xrightarrow{a.s.} 0$. The second conclusion follows from lemma 5.3. □