Demand Analysis with Many Prices*

Victor Chernozhukov† Jerry A. Hausman‡ Whitney K. Newey§
MIT *MIT* *MIT*

August 2019

Abstract

From its inception, demand estimation has faced the problem of "many prices." While some aggregation across goods is always necessary, the problem of many prices remains even after aggregation. Although objects of interest may mostly depend on a few prices, many prices should be included to control for omitted variables bias.

This paper uses Lasso to mitigate the curse of dimensionality in estimating the average expenditure share from cross-section data. We estimate bounds on consumer surplus (BCS) using a novel double/debiased Lasso method. These bounds allow for general, multidimensional, nonseparable heterogeneity and solve the "zeros problem" of demand by including zeros in the estimation.

We also use panel data to allow for prices paid to be correlated with preferences. We average ridge regression individual slope estimators and bias correct for the ridge regularization.

We find that panel estimates of price elasticities are much smaller than cross section elasticities in the scanner data we consider. Thus, it is very important to allow correlation of prices and preferences to correctly estimate elasticities. We find less sensitivity of consumer surplus bounds to this correlation.

Keywords: Demand analysis, machine learning, panel data.

*Research for this paper was supported by NSF Grant 1757140. Helpful comments were provided by R. Blundell, B. Deaner, Y. Gao, M. Harding, S. Hoderlein, M. Keene, and J. Shapiro. B. Deaner, Y Gao, M. Hardy, and K. Quist provided excellent research assistance. The empirical work here is researchers own analyses based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researchers and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

†Department of Economics, MIT, Cambridge, MA 02139, U.S.A E-mail: vchern@mit.edu.

‡Department of Economics, MIT, Cambridge, MA 02139, U.S.A E-mail: jahausman@gmail.com

§Department of Economics, MIT, Cambridge, MA 02139, U.S.A E-mail: wnewey@mit.edu.

1 Introduction

Estimation of demand models has a long history in econometrics. Beginning in the 1950s during the "Stone age" of econometrics at Cambridge University and elsewhere, applied researchers began estimating systems of demand equations as computer power increased.[1] From its inception, demand estimation faced the problem of "many prices." A demand system of m goods for person i takes the form $q_i = q(p_i, y_i, w_i, \eta_i)$, where q_i is an vector of quantities demanded, p_i is an m-vector of prices, y_i is a measure of income or expenditure, w_i is a conditioning variable for the individual and potentially for variables that are time specific, and η_i is a vector of unknown dimension (potentially an infinite vector) of an unobserved heterogeneity term and stochastic terms. Since m has high dimension, the presence of many goods and many prices creates problems for demand estimation.[2]

While some aggregation across goods is always necessary, the problem of many prices remains even after aggregation. Omission of relevant prices creates an omitted-variables problem, which leads to biased estimates. A number of approaches have been used on the many-prices problem. The Hicks-Leontief composite commodity theorem states that if prices move together for a group of commodities, i.e. relative prices are constant, they can be treated as a single good.[3] This type of aggregation is often used, at least implicitly. Another common approach is to assume some form of separability of preferences. If demand for some goods is independent of the demand and prices of other goods, no omitted variables problems will exist. A more sophisticated approach using this idea is to assume two-stage budgeting. At the top stage the consumer determines how much to spend on, say, food using price indices for food and other groups of expenditure. At the second stage the consumer determines the price index for food and demand for individual food products using only food prices and food expenditure. The required conditions for exact two-stage budgeting can be weakened and further results by Gorman (1959, 1981) used.[4] Separability can also be tested using specification tests. However, all the approaches to separability place strong assumptions on the demand system, such as approximate homotheticity, where budget shares do not depend on expenditure, or quasi-homotheticity, where budget shares are linear functions of expenditure so that Engel curves are linear. A last approach not using separability assumptions uses statistical aggregation for the many prices to a price index which, however, is independent of consumer preferences.[5]

[1] Named after Richard Stone, director of the Department of Applied Economics and co-inventor of the Stone-Geary demand system.

[2] For example, a typical supermarket has approximately 50,000 individual products.

[3] See e.g. Deaton and Muellbauer (1980) for a discussion of separability approaches.

[4] Blundell and Robin (2000) introduce latent separability in demand systems but, again, the required conditions are quite strong in terms of the cross price effects among goods.

[5] A recent approach is by Holderlein and Lewbel (2012), who use principal components to reduce the dimension of prices.

An observation arising from economic theory is that often, but not always, the policy question of interest depends on only one, or a very few, price effects. For example, estimation of consumer surplus and deadweight loss typically depend only on the own price effect since, all other prices are held constant.[6] Another example is merger analysis, where the price effect of the merger depends mostly on the cross-price effect of the merging goods, so if single-product firms are merging, only the cross price effect between the two goods which, by Slutsky, are equal in both directions, matters.[7] Another common feature is that cross-price effects for goods that are not closely related tend to be small, e.g. of an order of magnitude smaller than own-price effects.

These observations suggest that machine learning (ML) methods could be used to mitigate the curse of dimensionality this type of situation, e.g. Lasso. Exact sparsity (zero cross-price effects) is not required for estimating the objects of interest, only approximate sparsity. Thus, to estimate consumer surplus (CS) and the bounds on consumer surplus (BCS) along with the associated deadweight loss (DWL), we employ modern machine-learning methods to first estimate average demand for products in the presence of many substitute and complementary products. Here the average demand function is "high-dimensional" in that it may depend on a high-dimensional price vector of prices and other features such as expenditure and consumer characteristics. ML methods can perform well for estimating such high-dimensional demand by employing regularization to reduce variance and trading off regularization bias with overfitting in practice. The regularization used by Lasso mitigates the curse of dimensionality by setting many coefficients equal to zero.

Both regularization bias and overfitting in estimating demand cause a heavy bias in estimators of BCS that are obtained by naïvely plugging ML estimators of demand into the definition of BCS. This bias results in the naïve estimator failing to be root- n consistent, where n is the sample size. The impact of regularization bias and overfitting on estimation of the parameter of interest can be removed by using double/debiased machine learning (DML), which relies on two critical ingredients:

1. using debiased/locally robust/orthogonal moments/scores that have reduced sensitivity with respect to unknown functions (the average expenditure share in our case) to estimate CS/BCS and
2. making use of cross-fitting, which provides an efficient form of data-splitting.

By constructing the debiased moment equation, our DML estimators of BCS are root (n)-consistent and are approximately unbiased and normally distributed, which allows us to construct valid confidence statements. We focus our estimates of high-dimensional average share based on Lasso, although the overall strategy can also be used in conjunction with other ML

[6] See e.g. Hausman (1981) and Hausman and Newey (1995).

[7] See e.g. Hausman, Leonard and Zona (1994) and Hausman, Morisi, and Rainey (2010) and more generally the 2010 DOJ and FTC Horizontal Merger Guidelines.

methods. This approach to estimate BCS when demand depends on many prices is a main contribution of our paper.

This paper allows for general consumer heterogeneity through the multidimensional η_i in the demand function $q(p_i, y_i, w_i, \eta_i)$. Hausman and Newey (2016) find that in a single cross section of individuals that the demand functions are not identified. If the demand functions are linear, parameter estimates may find the mean preference of the "representative consumer." However, typically, demand systems are non-linear because of the presence of the budget constraint and non-homotheticity as demonstrated in the often-used AIDs demand system of Deaton and Muellbauer (1980).[8] Thus, Hausman and Newey (2016) developed the BCS approach based on the average demand $\bar{q}(p_i, y_i, w_i) = \int q(p_i, y_i, w_i, \eta) G(d\eta)$ where $G(\eta)$ denotes the distribution of heterogeneity. Empirically, the bounds are found to be close to each other. This paper also allows for η_i to be correlated with income y_i, which can occur because y_i is set equal to total expenditure. We control for this source of endogeneity by using a control variable v_i such that preferences are independent of prices, total expenditure, and covariates conditional on v_i, similarly to Hausman and Newey (2016).

Another methodological contribution is to treat the "zero problem" of demand estimation as a demand choice, not as a result of a stochastic disturbance. For example, consumer data which considers alcohol or tobacco consumption will have many individuals with zero consumption. In typical demand estimation, where identical parameters are assumed across individuals, zero consumption must occur because of a stochastic disturbance, since similar individuals both consume and do not consume the same good, e.g. alcohol. However, with a vector of disturbances η_i of unknown dimension, we allow zero purchases to be the outcome of a demand choice rather than the outcome of a stochastic disturbance. Thus, some consumers have preferences such that they will not consume alcohol or tobacco. Allowing for preference variation and including the zeros in the demand estimation is the correct econometric approach for estimating average demand. Similarly, including the zero-consumption outcomes in the estimate of CS and BCS is the correct approach for policy analysis. This approach to zero consumption outcomes greatly simplifies the analysis and estimation of demand systems.

An additional contribution is to use panel data to control for prices and total expenditure that may be correlated with preferences. Such correlation could result from consumers with higher elasticities searching more intensively for lower prices. We estimate separate own price and income effects for each individual and then average them to obtain average price effects. We regularize using ridge regression for each individual and debias to correct for ridge regularization on average. The resulting average slope estimators are unbiased if individual coefficients are independent of regressors and otherwise are a weighted average of individual coefficients with more strongly identified individual coefficients weighted more heavily. We give inference theory,

[8] Other demand systems such as the translog are also non-linear in the absence of homotheticity.

including primitive conditions for large enough, fixed number of time periods.

We compare these methods in estimating average share regressions using scanner data for soda and other commodities. We use these estimates to bound average welfare effects of an increase in the price of soda, as would occur if soda were taxed more heavily. For the cross-section estimators we use share regression specifications that allow for nonlinearity in log prices and log income. For the panel results we consider a share regression that is linear in log prices and income. This functional form is more parsimonious than our cross-section models, motivated by the few numbers of observations for each individual. We find panel elasticities are substantially smaller than the cross section estimates, strongly suggesting that prices are correlated with preferences. We also find less striking differences between cross-section and panel estimates of average surplus bounds.

Choice models with general heterogeneity have previously been considered. In their analysis of nonlinear taxes, Burtless and Hausman (1978) allowed heterogenous income effects. Lewbel (2001) considered the implications of such models for conditional mean regressions. McFadden (2005) allowed for general heterogeneity in a revealed preference framework. The approach here specializes the revealed preference work in imposing single valued, smooth demands to facilitate estimation, as in Hausman and Newey (2016). Blomquist and Newey (2002) derived the form of average demand with nonparametric, nonseparable, scalar heterogeneity and nonlinear taxes and Blomquist, Kumar, Liang, and Newey (2014) showed the same form for general heterogeneity. Hoderlein and Stoye (2014) showed how to impose the weak axiom of revealed preference. Dette, Hoderlein, and Neumeyer (2016) proposed tests of downward sloping compensated demands. Bhattacharya (2015) derived average surplus for discrete demand and general heterogeneity. Kitamura and Stoye (2018) gave tests of the revealed preference hypothesis.

The double machine learning estimator is novel in the use of a minimum distance Lasso method to debias the estimator when using a control variable. The estimator and theory build on that of Chernozhukov, Newey, and Robins (2018) and Chernozhukov, Newey, and Singh (2018) for minimum distance Lasso bias correction without a control function. This work in turn builds on Belloni et al. (2012) and Belloni, Chernozhukov and Hansen (2013) on debiased machine learning.

For panel data Chamberlain (1982, 1992), Pesaran and Smith (1995), Wooldridge (2005), Arellano and Bonhomme (2012), Chernozhukov, Fernandez-Val, Hahn, and Newey (2013), and Graham and Powell (2012) have considered averaging individual slope estimates. The bias corrected average ridge estimator given here appears to be novel as does the associated inference theory.

Harding and Lovenheim (2017) analyze the role of prices in determining food purchases and nutrition and estimate the impact of taxes on nutrition and individual welfare. Allcott, H., B. B. Lockwood, and D. Taubinsky (2019) and Dubois, P., R. Griffith, and M. O'Connell (2019)

have also considered the welfare effects of taxing soda. Our results are complementary to theirs in the use of grocery store scanner data, allowance for nonparametric, general heterogeneity in the cross-section, including zeros in regressions, and in the comparison of cross-section and panel results.

2 Demand and Weighted Average Surplus

We consider a demand model where the form of heterogeneity is unrestricted. To describe the model let q denote the quantity of a vector of goods, a the quantity of a numeraire good, p the price vector for q relative to a, and y the individual income level relative to the numeraire price. The unobserved heterogeneity will be represented by a vector η of unobserved disturbances of unknown dimension. We think of each value of η as corresponding to a consumer but do allow η to be continuously distributed.

For each consumer η the demand function $q(p, y, \eta)$ will be obtained by maximizing a utility function $U(q, a, \eta)$ that is monotonic increasing in q and a, subject to the budget constraint, with

$$q(p, y, \eta) = \arg \max_{q \geq 0, a \geq 0} U(q, a, \eta) \text{ s.t. } p'q + a \leq y. \tag{2.1}$$

Here we assume that demand is single valued and not a correspondence. This assumption is essentially equivalent to strict quasi-concavity of the utility function. We impose no form on the way η enters the utility function U, and hence the form of heterogeneity is completely unrestricted.

For analyzing the effect of price changes on welfare we focus on equivalent variation. Let $e(p, u, \eta) = \min_{q \geq 0, a \geq 0}\{p'q + a \text{ s.t. } U(q, a, \eta) \geq u\}$ be the expenditure function and $V(y, \eta) = y - e(p^0, u^1, \eta)$ be the equivalent variation for individual η for a price change from p^0 to p^1 with income y and u^1 the utility at price p^1. The corresponding deadweight loss is $D(y, \eta) = V(y, \eta) - (p^1 - p^0)'q(p^1, y, \eta)$.

In the remainder of this paper we focus on the case where the first price p_1 changes from $\check{p}_1$ to a higher value $\bar{p}_1$ and the other prices p_2 in $p = (p_1, p_2')'$ are fixed. In that case the equivalent variation $V(p_2, y, \eta)$ will also depend on the other prices p_2. Also, in many applications it may be useful to allow for covariates. Covariates w represent observed sources of heterogeneity in preferences that are allowed to be correlated with prices and income and are independent of preference heterogeneity η. In that case the equivalent variation $V(p_2, y, w, \eta)$ will also depend on w. For notation we will find it convenient to put the prices, income, and covariates into one vector $x = (p', y, w')'$ and partition as $x = (p_1, x_2')'$. Also we denote the equivalent variation and demand for the first good as $V(x_2, \eta)$ and $q_1(x, \eta)$.

Our object of interest is the average equivalent variation (AEV) weighted by a function $\omega(x_2)$

that depends on observed variables x_2 other than p_1, given by

$$V_0 = E[\omega(x_{2i})V(x_{2i}, \eta_i)].$$

Following Hausman and Newey (2016) we can use bounds on income effects to construct an identified set for V_0 using expected demand. The bound on income effects takes the following form.

ASSUMPTION 1: *There are $b \leq \bar{b}$ such that for all $p_1 \in [\check{p}_1, \bar{p}_1]$, (x_2, η), and $\Delta y \in [0, V(x_2, \eta)]$,*

$$\omega(x_2) \cdot b \cdot \Delta y \leq \omega(x_2)[q_1(p, y, w, \eta) - q_1(p, y - \Delta y, w, \eta)] \leq \omega(x_2) \cdot \bar{b} \cdot \Delta y.$$

This condition places upper and lower Lipschitz bounds on income effects, as we will further discuss below.

The bound on income effects leads to a BCS of the form

$$V_B(x_2, \eta) = \int_{\check{p}_1}^{\bar{p}_1} q_1(u, x_2, \eta) \exp(-B[u - \check{p}_1])du, \tag{2.2}$$

where u is a scalar variable of integration for the first price and B is equal to b or $\bar{b}$ from Assumption 1. If $B = b$ (or $B = \bar{b}$) then $V_B(x_2, \eta)$ is an upper (lower) bound on the equivalent variation for a price change from $\check{p}_1$ to $\bar{p}_1$ at x_2, for an individual indexed by η. This bound can be integrated over η to obtain a BCS for average equivalent variation based on average demand. Taking the expectation over the marginal distribution G of η and interchanging the order of integration we obtain

$$\begin{aligned} V_B(x_2) &= \int \{ \int_{\check{p}_1}^{\bar{p}_1} q_1(u, x_2, \eta) \exp(-B[u - \check{p}_1])du \} G(d\eta) \\ &= \int_{\check{p}_1}^{\bar{p}_1} \bar{q}_1(u, x_2) \exp(-B[u - \check{p}_1])du, \quad \bar{q}_1(x) = \int q_1(x, \eta) G(d\eta). \end{aligned} \tag{2.3}$$

If $B = b$ (or $B = \bar{b}$) then $V_B(x_2)$ is an upper (or lower) bound on the equivalent variation for a price change from $\check{p}_1$ to $\bar{p}_1$ averaged over the unobserved individual heterogeneity η.

The bound $V_B(x_2)$ will be identified from data where individual heterogeneity η is distributed independently of x. Under such independence $\bar{q}_1(x)$ will be the conditional expectation of q_1 given x in the data, i.e. the nonparametric regression of q_1 on x. Thus the BCS $V_B(x_2)$ can be obtained by integrating the nonparametric regression $\bar{q}_1(x)$ of q_1 on x as in equation (2.3). A corresponding BCS for V_0 can be constructed from the weighted expectation of $V_B(x_2)$ over x, as

$$V_B = E[\omega(x_{2i})V_B(x_{2i})].$$

If $B = b$ (or $B = \bar{b}$) then V_B is an upper (lower) bound on the AEV V_0.

One example of a weight function is an average BCS where income varies over a range. In that example we could take

$$\omega(x_2) = 1(Q_y(\tau_1) \leq y \leq Q_y(\tau_2))/(\tau_2 - \tau_1),$$

where $Q_y(\tau)$ is the quantile function for y. As τ_1 and τ_2 vary V_B will give a BCS for different income groups. In the application we will consider the case where $(\tau_1, \tau_2) = (0, .25)$ for one bound and $(\tau_1, \tau_2) = (.75, 1)$ for another bound. In this case the two bounds will give BCS over the lower and upper quartiles of income.

Applications of demand models often involve estimating expenditure share equations rather than demand equations, for the reasons discussed in Deaton and Muellbauer (1980). We will follow that practice in this paper. For this reason we restate the BCS in terms of expenditure share. Let $s(x, \eta) = y^{-1} p_1 q_1(x, \eta)$ denote the share of income spent on the first good. The average share is given by

$$\bar{s}(x) = \int s(x, \eta) G(d\eta) = \frac{p_1 \bar{q}_1(x)}{y}.$$

The BCS in terms of expected share is

$$\beta_0 = E[\omega(x_{2i}) \int_{\check{p}_1}^{\bar{p}_1} (\frac{y_i}{u}) \bar{s}(u, x_{2i}) \exp(-B[u - \check{p}_1]) du]. \tag{2.4}$$

Throughout the remainder of the paper we will carry out the analysis in terms of share equations to maintain a close link with applied demand analysis. Thus we will take β_0 to be one object of interest, a BCS stated in term of the regression of share on prices, income, and other covariates. We could also consider a corresponding bound on deadweight loss (BDL) given by

$$d_0 = \beta_0 - \bar{p}_1^{-1} (\bar{p}_1 - \check{p}_1) E[\omega(x_{2i}) y_i s(\bar{p}_1, x_{2i})]. \tag{2.5}$$

In this paper we will focus on the BCS β_0.

An important feature of these bounds is that they allow for individuals to choose zeros for some goods. Intuitively, if $q_1(x, \eta) = 0$ over the range of change for p_1 then the price change does not effect the welfare of the individual (e.g. see equation (2.2)). The BCS β_0 simply includes these zeros in the average. Similarly if quantity demanded is not zero over the price range then the positive part will also be included in the average. In addition, the form of the income effect bound in Assumption 1 implicitly allows for zeros. A demand function will generally not be differentiable in income at a point where demand begins to become positive. The bound in Assumption 1 allows for nondifferentiable demand functions.

The unstructured nature of heterogeneity also provides intuition for the absence of a zeros problem. Any disturbance can affect any quantity demanded. Also, the presence of specific disturbances that determine when specific goods are zero is allowed for. The dimension of η and the way in which η affects demand is completely unrestricted. The average share $\bar{s}(p_1, x_2)$ takes

all this into account as it integrates over possible values of η. In this way the BCS overcomes the "zeros problem."

The BCS depends on bounds on the income effect. Simple bounds are available when all goods are normal goods, that is when all income effects are nonnegative.

LEMMA 1: *If preferences satisfy local nonsatiation and all goods are normal then Assumption 1 is satisfied with* $b = 0$ *and* $\bar{b} = 1/\check{p}_1$.

This result is intuitive: If all of the income addition Δy is spent on the first good and $p_1 \geq \check{p}_1$ then the individual can purchase no more than $1/\check{p}_1$ and no other income will be available because all of the goods are normal goods. This bound is quite coarse because we would expect purchases from additional income to be spread across all goods, at least to some degree. Hausman and Newey (2016) consider finer bounds obtained as some large multiple of the maximum of quantile derivatives over several quantiles.

This bound does depend crucially on all goods being normal. Normal goods seems a reasonable assumption for scanner data if goods are aggregated into groups of goods. For individual goods which differ primarily in quality, e.g. standard and premium orange juice, it seems unlikely that the normal goods assumption would hold. Consumers might choose less of the lower quality good as income increases.

For a normal good the upper BCS will be approximate average surplus obtained from equation (2.4) with $B = 0$. We can also obtain a simple lower bound from choosing $B = 1/\check{p}_1$ in equation (2.4). For $u \in [\check{p}_1, \bar{p}_1]$,

$$\exp(-\frac{u - \check{p}_1}{\check{p}_1}) \geq \exp(-\frac{\bar{p}_1 - \check{p}_1}{\check{p}_1}) \geq 2 - [\bar{p}_1/\check{p}_1].$$

It follows from the this equation and the form of β_0 in equation (2.4) that the lower BCS will be $2-[\bar{p}_1/\check{p}_1]$ times the upper bound. For example, for a 10 percent price increase this lower bound would be 90 percent of the upper bound. We emphasize that this is an even coarser bound than that for $B = 1/\check{p}_1$. In the application we will consider both this coarse bound and finer bounds based on quantile derivatives.

3 Learning the BCS from Cross-Section Data

For learning (estimating) the weighted average BCS it is helpful to modify the formula to allow simulation to be used in estimating the integral in the BCS. For this purpose let u_i denote a random variable that is independent of the data and uniformly distributed on $(\check{p}_1, \bar{p}_1)$, $\tilde{x}_i = (u_i, x_{2i})$, and $\zeta(\tilde{x}) = \omega(x_2)(\bar{p}_1 - \check{p}_1)(y/u)\exp(-B[u - \check{p}_1])$. The BCS is then

$$\beta_0 = E[\zeta(\tilde{x}_i)\bar{s}(\tilde{x}_i)], \quad \bar{s}(x) = \frac{p_1 \int q_1(x, \eta)G(d\eta)}{y}. \tag{3.1}$$

When η is independent of prices p, income y, and covariates w the average share $\bar{s}(x)$ will equal the conditional expectation $E[s|x]$ of observed share s given $x = (p, y, w)$.

In scanner data independence of η and x will be problematic because the variable y will be total expenditure on the goods considered. Total expenditure depends on η so y will be endogenous and $\bar{s}(x) \neq E[s|x]$. A control variable can be used to correct for this endogeneity. A control variable is an observed or estimable variable v such that x and η are independent conditional on v. Averaging over v controls for endogeneity in nonseparable models with general heterogeneity, see Chamberlain (1984), Blundell and Powell (2003), Wooldridge (2002), and Imbens and Newey (2009). Demand with general heterogeneity is such a model, as shown in Hausman and Newey (2016) and Kitamura and Stoye (2018). For demand analysis a control variable can be constructed when there is a first stage equation for expenditure as a function of earnings, other exogenous variables, and a scalar disturbance, with earnings acting as an instrument for total expenditure. Any strictly monotonic function of the scalar disturbance will be a control variable when it and the demand heterogeneity are jointly independent of earnings. Blundell, Duncan, and Pendakur (1998) used such a specification to control for endogeneity of total expenditure.

Independence of x and η conditional on the control variable v and an identification condition will imply that

$$\bar{s}(x) = \int \gamma_0(x, v) F_0(dv), \quad \gamma_0(x, v) = E[s|x, v]. \tag{3.2}$$

Here the average share $\bar{s}(x)$ is the average structural function of Blundell and Powell (2003) and Wooldridge (2002). The average structural function will be identified if $s(x, \eta) = \theta(x)'\eta$ for a known vector of functions $\theta(x)$ and $E[\theta(x)\theta(x)'|v]$ is nonsingular with probability one, as shown by Masten and Torgovitsky (2015). This nonsingularity condition allows for a discrete instrument as long as the instrument has as many points of support as the dimension of $\theta(x)$. Allowing for a discrete instrument will be important in our application. With average share $\bar{s}(x)$ in equation (3.2) the BCS will be given by equation (3.1). We will develop an estimator of this object.

In scanner data x can be high dimensional because prices of many goods may affect the share of a particular good. Cross price elasticities tend to be quite small suggesting machine learning methods that make use of approximate sparseness might be useful. Here we consider Lasso estimation of the share regression in order to do this. A natural approach would be to "plug in" a Lasso share regression into sample analogs of equations (3.2) and (3.1). That approach does not work in general. It may be so biased that it is not root-n consistent, as discussed in Chernozhukov et al. (2018a,b). An alternative method that will give a root-n consistent estimator is debiased/double machine learning (DML).

DML modifies the plug-in estimator by adding the influence adjustment for the presence of an unknown conditional expectation and unknown distribution of the control function, as

in Chernozhukov et al. (2018b), Newey (1994), and Newey and McFadden (1994). Adding the adjustment gives second order error from estimating the conditional expectation and distribution of the control function. The adjustment term does depend on additional, unknown high-dimensional objects and so these have to be estimated. Here we do so using an automatic method that depends only the integral in equation (2.4) and not on knowing the form of the adjustment. That automatic method is a minimum distance Lasso that builds on Chernozhukov, Newey, and Singh (2018) and Chernozhukov, Newey, and Robins (2018) and is novel in accounting for the distribution of the control function.

To describe the DML estimator let z_i denote a data observation that includes the share s_i, prices p_i, income y_i, covariates w_i, and control variable v_i for observation $i = 1, ..., n$. We will use a Lasso estimator of the share regression. To describe that estimator let $b(x, v) = (b_1(x, v), ..., b_J(x, v))'$ be a dictionary of functions that will be used to approximate the share regression. A Lasso estimator of $\gamma_0(x, v)$ is given by

$$\hat{\gamma}(x, v) = b(x, v)'\hat{\delta}, \quad \hat{\delta} = \arg\min_{\delta}\{\frac{1}{n}\sum_{i=1}^{n}[s_i - b(x_i, v_i)'\delta]^2 + \frac{r_\delta}{2}\sum_{j=1}^{J}|\delta_j|\},$$

where we assume that the elements of $b(x, v)$ have already been scaled so that $\sum_i b(x_i, v_i)^2/n = 1$. The coefficient vector $\hat{\delta}$ will often have some zero elements corresponding to a sparse approximation to the conditional mean. The term r_δ is a regularization degree that controls how much sparsity (number of zero coefficients) there are in $\hat{\delta}$. In the application we will use cross-validation to choose r_δ.

We next describe the plug-in estimator. It is convenient to combine equations (3.1) and (3.2) to obtain

$$\beta_0 = E[\int \zeta(\tilde{x}_i)\gamma_0(\tilde{x}_i, v)F_0(dv)].$$

The plug in estimator can be constructed by substituting $\hat{\gamma}$ and $\hat{F}_v$ in this equation and replacing the expectation with the sample average. To help reduce bias and to obtain root-n consistency and asymptotic normality under weak regularity conditions we will use cross-fitting where $\hat{\gamma}$ and $\hat{F}_v$ come from different observations than those being averaged over. To do this cross-fitting we divide the data into L about equal sized groups. Let I_ℓ denote the index of observations in group ℓ. Let $\hat{\gamma}_\ell$ be the Lasso estimator computed from all observations not in I_ℓ. The cross-fit plug in estimator is

$$\tilde{\beta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\zeta(\tilde{x}_i)\left[\frac{1}{n - n_\ell}\sum_{j\notin I_\ell}\hat{\gamma}_\ell(\tilde{x}_i, v_j)\right],$$

where n_ℓ is the number of observations in I_ℓ.

As previously noted such a plug-in estimator can have large bias. We debias by adding the influence adjustment that corrects for the presence of an unknown conditional expectation and

marginal distribution. The adjustment is the influence function of $\int\int \zeta(\tilde{x})\gamma(\tilde{x},v,F)F_0(d\tilde{x})F_v(dv)$ where $\gamma(x,v,F)$ denotes the conditional expectation of s given (x,v) and F_v the marginal distribution of v when F is the true distribution. As in Newey (1994, p. 1357) the influence adjustment will be the sum of two terms, one being the adjustment for γ and the other for F_v. The influence adjustment for γ depends on a Riesz representer $\alpha_0(x,v)$ such that

$$E[\int \zeta(\tilde{x}_i)\gamma(\tilde{x}_i,v)F_v(dv)] = E[\alpha_0(x_i,v_i)\gamma(x_i,v_i)], \quad \alpha_0(x,v) = \zeta(x)\frac{f_{x_20}(x_2)f_{v0}(v)}{f_{x,v0}(x,v)},$$

for all $\gamma(x_i,v_i)$ with finite second moment, where $f_{x_20}(x)$ and $f_{v0}(v)$ are the marginal pdf's of x_{2i} and v_i and $f_{x,v0}(x,v)$ the joint pdf. The adjustment for γ is

$$\phi_1(z,\alpha,\gamma) = \alpha(x,v)[s-\gamma(x,v)],$$

where γ and α represent a possible conditional mean and Riesz representer, as shown by Newey (1994). Also, the adjustment for F_v is

$$\phi_2(v,\gamma,F_{\tilde{x}},F_v) = \int \zeta(\tilde{x})\gamma(\tilde{x},v)F_{\tilde{x}}(d\tilde{x}) - \int\int \zeta(\tilde{x})\gamma(\tilde{x},v)F_{\tilde{x}}(d\tilde{x})F_v(dv)],$$

where $F_{\tilde{x}}$ and F_v are possible CDF's of $\tilde{x}_i$ and v_i respectively, as shown in Newey and McFadden (1994). Plugging in estimators $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ and taking $\hat{F}_v$ and $\hat{F}_{\tilde{x}}$ to be the empirical distributions over observations not in I_ℓ gives the estimated adjustment term

$$\hat{\phi}_\ell(z) = \hat{\phi}_{\ell 1}(z) + \hat{\phi}_{\ell 2}(v), \quad \hat{\phi}_{\ell 1}(z) = \hat{\alpha}_\ell(x,v)[s-\hat{\gamma}_\ell(x,v)],$$
$$\hat{\phi}_{\ell 2}(v) = \frac{1}{n-n_\ell}\sum_{j\notin I_\ell}\zeta(\tilde{x}_j)\hat{\gamma}_\ell(\tilde{x}_j,v) - \left(\frac{1}{n-n_\ell}\right)^2\sum_{j,j'\notin I_\ell}\zeta(\tilde{x}_j)\hat{\gamma}_\ell(\tilde{x}_j,v_{j'}).$$

The DML estimator with cross-fitting for the influence adjustment is then

$$\hat{\beta} = \tilde{\beta} + \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\phi}_\ell(z_i).$$

This estimator depends on the estimator $\hat{\alpha}_\ell$ of the Riesz representer α_0. It is not necessary to use the form of α_0 to estimate it. We can construct a Lasso minimum distance estimator that automatically estimates α_0 using only $\zeta(\tilde{x})$ without knowing the form of α_0. Let $\hat{M}_\ell$ denote a $J\times 1$ vector with k^{th} component

$$\hat{M}_{\ell k} = \left(\frac{1}{n-n_\ell}\right)^2\sum_{i\notin I_\ell}\sum_{j\notin I_\ell}\zeta(\tilde{x}_i)b_k(\tilde{x}_i,v_j), (k=1,...,J), \ \hat{G}_\ell = \frac{1}{n-n_\ell}\sum_{i\notin I_\ell}^{n} b(x_i,v_i)b(x_i,v_i)'.$$

The estimator $\hat{\alpha}_\ell$ is

$$\hat{\alpha}_\ell(x,v) = b(x,v)'\hat{\rho}_\ell, \ \hat{\rho}_\ell = \arg\min_\rho\{-2\hat{M}_\ell'\rho + \rho'\hat{G}_\ell\rho + r_\rho\sum_{k=1}^{J}|\rho_k|\}.$$

The coefficients $\hat{\rho}_\ell$ minimize a L_1 penalized minimum distance objective function. The $\hat{M}_\ell$ here has a novel form in accounting for endogeneity through averaging over the control function. The objective function is like that considered in Chernozhukov, Newey, and Singh (2018) with the novel form of $\hat{M}_\ell$.

We can estimate the asymptotic variance of $\hat{\beta}$ using the fact that to first order $\hat{\beta}$ is a sample average. For $i \in I_\ell$ let

$$\hat{\psi}_{i\ell} = \frac{1}{n - n_\ell} \sum_{j \notin I_\ell} [\zeta(\tilde{x}_i)\hat{\gamma}_\ell(\tilde{x}_i, v_j) + \zeta(\tilde{x}_j)\hat{\gamma}_\ell(\tilde{x}_j, v_i)] - \left(\frac{1}{n - n_\ell}\right)^2 \sum_{j,j' \notin I_\ell} \zeta(\tilde{x}_j)\hat{\gamma}_\ell(\tilde{x}_j, v_{j'}) - \hat{\beta}$$
$$+ \hat{\alpha}_\ell(x_i, v_i)[s_i - \hat{\gamma}_\ell(x_i, v_i)].$$

This $\hat{\psi}_{i\ell}$ is an estimator of the influence function of $\hat{\beta}$. The asymptotic variance of $\hat{\beta}$ will be estimated by the sample second moment of $\hat{\psi}_{i\ell}$ as

$$\hat{V} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_{i\ell}^2.$$

We now give regularity conditions for asymptotic normality of $\hat{\beta}$ and consistency of $\hat{V}$. The first condition specifies that the dictionary is multiplicatively separable in the regressors and control variable and bounded.

ASSUMPTION 2: *There is $C > 0$ and for every j there is $b_j^x(x)$ and $b_j^v(x)$ such that*

$$b_j(x, v) = b_j^x(x) b_j^v(v), \quad \left|b_j^x(x)\right| \leq C, \quad \left|b_j^v(v)\right| \leq C.$$

The multiplicative form of the dictionary terms b_j is useful in analyzing the double averages on which $\hat{\beta}$ depends. We also require that the joint pdf of x_2 and v dominates the product of marginal densities.

ASSUMPTION 3: *There is $C > 0$ such that $|\omega_0(x)| \leq C$, $J \leq Cn^C$ for some $C > 0$, and the (x_i, v_i) are absolutely continuous with respect to a product measure with joint pdf $f_{x,v0}(x, v)$ and marginal pdf's $f_{x_2 0}(x_2)$, and $f_{v0}(v)$ satisfying*

$$1(\check{p}_1 \leq x_1 \leq \bar{p}_1) f_{x_2 0}(x_2) f_{v0}(v) \leq C f_{x,v0}(x, v).$$

We note that this condition requires that the pdf of x, v is bounded away from zero over the price range $\check{p}_1 \leq x_1 \leq \bar{p}_1$, at all x_2 and v with $f_{x_2 0}(x_2) > 0$ and $f_{v0}(v) > 0$. This condition also includes the full support condition for the control function by virtue of the joint distribution dominating the product of marginals. In Appendix A we also give more technical conditions that involve sparse eigenvalue and rate of approximation conditions in Assumptions A1 and A2.

THEOREM 2: *If Assumptions 1, 2, 3, A1, and A2 are satisfied,* $\sqrt{\ln(J)/n} = o(r_\delta)$, $\sqrt{\ln(J)/n} = o(r_\alpha)$, *and for* $\bar{s}_\delta$ *from Assumption A2,* $\bar{s}_\delta r_\delta^2 \longrightarrow 0$, $r_\rho \longrightarrow 0$, *and* $\sqrt{n}(\bar{s}_\delta)^{1/2} r_\delta r_\rho^{1/2} \longrightarrow 0$ *then there is* $V > 0$ *with*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V), \ \hat{V} \xrightarrow{p} V.$$

4 Estimation for Panel Data

Panel data has the potential to control for time invariant endogenous unobserved individual heterogeneity. Such endogeneity could arise from correlation of expenditure on a group of commodities with preferences that determine share purchases. Also, preferences could be correlated with prices due to search behavior of consumers. In this Section we give a panel estimator for a linear model with individual specific coefficients that may be correlated with prices and income. Individual coefficients are estimated in a ridge regression to allow for the possibility that these coefficients are not well identified for some individuals. We average the individual ridge estimates and bias correct the averages for ridge regularization. We rely on the number of time periods and the within individual variation being large enough that the inverse of the individual regressor second moment matrices have finite expectation.

One important feature of our data is that not every time period is observed for every individual. We will assume for convenience that the first $T_i \leq T$ observations are available for individual i. The vector of observations on shares and prices, income, and other covariates will then be

$$s_i = (s_{i1}, ..., s_{i,T_i})', \ x_i = (x_{i1}', ..., x_{iT_i}')'.$$

We continue to assume that the data are independently distributed across i. We also assume that time series observations are missing at random so that our results are not affected by having different number of observations for different individuals. Assumption 4 given below implicitly includes this condition. Our conditions are like Wooldridge (2018). Also, Hausman and Leibtag (2007) tested the missing at random assumption and found it was not rejected.

For panel data we assume that the budget share s_{it} of individual i in time period t is

$$s_{it} = s(x_{it}, \eta_{it}), \ (t = 1, ..., T_i; \ i = 1, ..., n),$$

where η_{it} denotes period specific preferences. Here each individual is allowed to have different preferences in each time period. Such idiosyncratic preference variation should help fit data better because it is often found that individuals make different choices when faced with the same choice sets. If preferences of individuals change over time in unrestricted ways then panel data provides no more information than cross-section data. Panel data does provide information when time variation is restricted. We will consider individual preferences where the distribution of η_{it} given $x_i = (x_{i1}', ..., x_{iT_i}')'$ is the same in each time period. This assumption can be thought of

as time homogeneity of preferences, with the preference being drawn from the same distribution in each time period conditional on x_i. Time homogeneity of preferences corresponds to time homogeneity of disturbances, an econometric condition that has proven useful in recent work on nonlinear panel data models, such as Chernozhukov et al. (2013), Graham and Powell (2012), Hoderlein and White (2011), Chernozhukov et al. (2015), and Chernozhukov, Fernandez-Val, and Newey (2017). Here time homogeneity will allow us to identify the average share, as needed for BCS, under conditions that we will describe.

We will impose the condition that the share is a linear combination of known functions of x_{it}. Specifically, we assume that there is a known vector of functions $b(x)$ that includes a constant and η_{it} are coefficients, with individual shares given by

$$s_{it} = s(x_{it}, \eta_{it}) = b(x_{it})'\eta_{it}.$$

As discussed in Hausman and Newey (2016) this specification can be interpreted as a series approximation to a general nonseparable share equation where the $b(x)$ is a vector of approximating functions. In this paper we ignore the approximation error and treat $b(x_{it})'\eta_{it}$ as a correct specification of individual shares.

The next condition imposes our basic panel data identifying assumption.

ASSUMPTION 4: $s_{it} = b(x_{it})'\eta_{it}$, $E[\eta_{it}|x_i]$ *does not depend on* t, *and* $\bar{\eta} = E[\eta_{it}]$ is finite and does not depend on i.

Here we require that the conditional mean $E[\eta_{it}|x_i]$ does not vary with t. In requiring that $E[\eta_{it}]$ does not vary with i we also impose that unbalanced panel data, where T varies with i, does not affect $E[\eta_{it}]$. This assumption is weaker than those of Graham and Powell (2012) in only imposing time homogeneity on conditional means rather than conditional distributions.

The average share continues to be the object of interest for learning the BCS. In the panel model here the average share will be

$$\bar{s}(x) = b(x)'\bar{\eta} = \int b(x)'\eta_t G(d\eta_t) = \int s(x, \eta_t) G(d\eta_t).$$

Here the integration over η is done for a single time period, with the time homogeneity hypothesis of Assumption 4 making the average surplus not depend on the time period. The corresponding BCS will be an average over different time periods and individuals of the bound on the equivalent variation. The average over individuals will not depend on the time period by virtue of Assumption 4.

The time stationarity condition helps to identify average coefficients when η_{it} and x_i are correlated, similarly to Chamberlain (1982, 1992). To explain let $\bar{\eta}_i = E[\eta_{it}|x_i]$ and $\varepsilon_{it} = b(x_{it})'[\eta_{it} - \bar{\eta}_i]$. Adding and subtracting $b(x_{it})'\bar{\eta}_i$ gives

$$s_{it} = b(x_{it})'\bar{\eta}_i + \varepsilon_{it}, \ (t = 1, ..., T_i; \ i = 1, ..., n).$$

By the time stationarity condition of Assumption 4 the disturbance ε_{it} will have conditional mean zero,

$$E[\varepsilon_{it}|x_i] = 0,$$

leading to unbiasedness of ordinary least squares. Let $b_i = [b(x_{1i}), ..., b(x_{iT_i})]'$ and $Q_i = b_i' b_i / T_i$. By Assumption 4 $E[\varepsilon_{it}|x_i] = 0$, so the usual least square properties will imply that when Q_i is nonsingular the least squares estimator $\tilde{\eta}_i = Q_i^{-1} b_i' s_i / T_i$ will be a conditionally unbiased estimator of $\bar{\eta}_i$,

$$E[\tilde{\eta}_i|x_i] = \bar{\eta}_i.$$

Thus using just the within individual variation in $b(x_{it})$ a conditionally unbiased estimator of $\bar{\eta}_i$ can be constructed by least squares regression for individual i. If Q_i is nonsingular for every i then the sample average $\sum_{i=1}^n \tilde{\eta}_i / n$ of individual least squares estimators is a conditionally unbiased estimator of $\sum_{i=1}^n \bar{\eta}_i / n$.

The problem with the average of individual least squares estimators $\sum_{i=1} \tilde{\eta}_i / n$ is that Q_i could be close to singular or even singular for some individuals, so that average least squares may not be well behaved. Singularity of Q_i can occur when there is not enough variation in $b(x_{it})$ over time for some individuals and will result in $\bar{\eta}_i$ not being identified for those individuals. Also, even when Q_i is nonsingular for every individual the average of individual least squares estimators may not be unconditionally unbiased nor consistent because moments of Q_i^{-1} may not exist, as pointed out by Graham and Powell (2012).

We deal with this problem by averaging ridge regularized individual estimates and correcting for the average bias of the regularization. We also partial out each individual specific constant before the ridge regularization. We suppose that $b(x) = (1, b_2(x)')'$ and partition $\eta = (\eta_1, \eta_2')'$ conformably, so that η_2 is the vector of coefficients of the nonconstant elements of $b(x)$. Let $b_{2i} = [b_2(x_{i1}), ..., b_2(x_{iT_i})]'$ be the matrix of observations on nonconstant regressors with rows corresponding to time periods and columns to variables. Let $e_{T_i} = (1, ..., 1)'$ be a $T_i \times 1$ vector of ones and $\tilde{b}_{2i}$ be the matrix of deviations from time means given by

$$\tilde{b}_{2i} = b_{2i} - e_{T_i} \bar{b}_{2i}', \ \bar{b}_{2i} = b_{2i}' e_{T_i} / T_i.$$

Let $Q_i = \tilde{b}_{2i}' \tilde{b}_{2i} / T_i$. A ridge regression estimator of the individual coefficients η_{2i} of nonconstant variables is

$$\hat{\eta}_{2i} = \Lambda_i \tilde{b}_{2i}' s_i / T_i, \ \Lambda_i = (Q_i + \lambda I_i)^{-1},$$

where λ is a positive scalar and I_i is a T_i dimensional identity matrix. The estimator of the average coefficients $\bar{\eta}_2$ that we consider is

$$\hat{\eta}_2 = \hat{B} \left(\frac{1}{n} \sum_{i=1}^n \hat{\eta}_{2i} \right), \quad \hat{B} = \left(\frac{1}{n} \sum_{i=1}^n \Lambda_i Q_i \right)^{-1}. \tag{4.1}$$

The matrix $\hat{B}$ in the estimator $\hat{\eta}_2$ corrects for average regularization bias from the individual ridge regressions. We can see this by noting that when Q_i^{-1} exists for each i the estimator $\hat{\eta}_2$ is a matrix weighted average of the individual least squares slope estimators $\tilde{\eta}_{2i} = Q_i^{-1}\tilde{b}_{2i}'s_i/T_i$. Since $\hat{\eta}_{2i} = \Lambda_i Q_i \tilde{\eta}_{2i}$ we have

$$\hat{\eta}_2 = \left(\sum_{i=1}^n W_i\right)^{-1} \sum_{i=1}^n W_i \tilde{\eta}_{2i}, \quad W_i = \Lambda_i Q_i.$$

The matrix W_i is closer to the identity the larger is Q_i in the positive semidefinite sense. Larger Q_i corresponds to slope coefficients being more strongly identified and W_i closer to the identity corresponds to less shrinkage. Thus, $\hat{\eta}_2$ can be interpreted as a matrix weighted average where more strongly identified individuals receive weight with less shrinkage. We can also estimate the average of the constant coefficients while correcting for regularization bias as

$$\hat{\eta}_1 = \frac{1}{n}\sum_{i=1}^n \hat{\eta}_{1i}, \ \hat{\eta}_{1i} = \bar{s}_i - \bar{b}_{2i}'(\hat{\eta}_{2i} + \lambda\Lambda_i\hat{\eta}_2). \tag{4.2}$$

To show how $\hat{B}$ corrects for regularization bias we can give an explicit expression for the expectation of $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2')'$ conditional on the regressors for all individuals.

THEOREM 3: *If Assumption 4 is satisfied then for* $\bar{\eta}_i = E[\eta_{it}|x_i]$,

$$E[\hat{\eta}_2|x_1, ..., x_n] = \left(\sum_{i=1}^n W_i\right)^{-1} \sum_{i=1}^n W_i \bar{\eta}_{2i},$$

$$E[\hat{\eta}_1|x_1, ..., x_n] = \frac{1}{n}\sum_{i=1}^n \{\bar{\eta}_{1i} + \lambda\bar{b}_{2i}'\Lambda_i(\bar{\eta}_{2i} - E[\hat{\eta}_2|x_1, ..., x_n])\}$$

Also, if $E[\eta_{it}|x_i]$ *does not depend on* x_i *then* $\hat{\eta}$ *is an unbiased estimator of* $\bar{\eta} = E[\eta_{it}]$.

Thus we see that the conditional expectation of the slope estimator is a matrix weighted average of the expectations of the individual slopes, with weights $W_i = \Lambda_i Q_i$. Also we see that $\hat{\eta}$ is corrected for regularization in that it is an unbiased estimator for the expectation of individual coefficients when they are conditional mean independent of the regressors.

It is straightforward to estimate the asymptotic variance of $\hat{\eta}$ while accounting for estimated bias corrections. An estimator is

$$\hat{V}_\eta = \frac{1}{n}\sum_{i=1}^n \hat{\psi}_{\eta i}\hat{\psi}_{\eta i}', \ \hat{\psi}_{\eta i} = (\hat{\psi}_{1i}, \hat{\psi}_{2i}')', \ \hat{\psi}_{2i} = \hat{B}[\hat{\eta}_{2i} - \Lambda_i Q_i \hat{\eta}_2], \tag{4.3}$$

$$\hat{\psi}_{1i} = \bar{s}_i - \bar{b}_{2i}'\hat{\eta}_{2i} - \frac{1}{n}\sum_{j=1}^n \left(\bar{s}_j - \bar{b}_{2j}'\hat{\eta}_{2j}\right)$$

$$- \lambda(\bar{b}_{2i}'\Lambda_i - \frac{1}{n}\sum_{j=1}^n \bar{b}_{2j}'\Lambda_j)'\hat{\eta}_2 - \lambda(\frac{1}{n}\sum_{j=1}^n \bar{b}_{2j}'\Lambda_j)\hat{\psi}_{2i}.$$

In panel data we can construct a BCS estimator from an estimator for average share similar to the cross-section case. In panel data the estimator of the average share will be

$$\hat{s}(x) = b(x)'\hat{\eta}.$$

To describe a corresponding BCS estimator let $\tilde{x}_{it} = (u_{it}, x_{2it}')'$ where u_{it} is uniformly distributed on $(\check{p}_1, \bar{p}_1)$ independently of the data. Also let $\overline{\zeta b}_i = \sum_{t=1}^{T_i} \zeta(\tilde{x}_{it}) b(\tilde{x}_{it})/T_i$. A BCS estimator is

$$\hat{\beta} = \overline{\zeta b}'\hat{\eta},\ \overline{\zeta b} = \frac{1}{n}\sum_{i=1}^{n} \overline{\zeta b}_i. \tag{4.4}$$

An estimator of the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_0)$ will be

$$\hat{V} = \frac{1}{n}\sum \hat{\psi}_i^2,\ \hat{\psi}_i = (\overline{\zeta b}_i - \overline{\zeta b})'\hat{\eta} + \overline{\zeta b}'\hat{\psi}_{\eta i} \tag{4.5}$$

Graham and Powell (2012) have given a regularized estimator that is an average of individual least squares estimators over individuals where the determinant of $b_i'b_i$ is larger than some cutoff. This is a hard thresholding regularization where individual data with $b_i'b_i$ close to singular are not used in the estimator. The ridge regularization involves shrinkage where all individuals are used with individual coefficients weighted by the strength of identification for the individual. Varying λ for the ridge regularization is useful because that changes how much the strength of identification affects the weights. This feature of ridge will be useful for the demand application where variation in λ helps quantify how fixed effect demand elasticities differ from average individual elasticities.

The ridge and Graham and Powell (2012) estimators are special cases of a general class of bias corrected regularized estimators. Let Λ_i denote some regularization of Q_i^{-1}. For the ridge and Graham and Powell (2012) estimators we would have

$$\Lambda_i = (Q_i + \lambda I)^{-1} \text{ (ridge)};\ \Lambda_i = 1(\det(Q_i) \geq \lambda)Q_i^{-1} \text{ (Graham and Powell, 2012)}.$$

A general regularized estimator is

$$\hat{\eta}_2 = \hat{B}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\eta}_{2i}\right),\quad \hat{B} = [\frac{1}{n}\sum_{i=1}^{n} \Lambda_i Q_i]^{-1},\ \hat{\eta}_{2i} = \Lambda_i \tilde{b}_{2i}' s_i/T_i,$$

where for convenience we have not changed the notation for $\hat{B}$ and $\hat{\eta}_i$. It follows exactly as in Theorem 3 that

$$E[\hat{\eta}_2 | x_1, ..., x_n] = \left(\sum_{i=1}^{n} W_i\right)^{-1} \sum_{i=1}^{n} W_i \bar{\eta}_{2i},$$

so that $\hat{\eta}_2$ will be unbiased when $\bar{\eta}_{2i}$ does not depend on i. This general class of estimators will be considered in more detail in future work.

For consistency and asymptotic normality we will impose that slightly more than the first moment of the trace $tr(Q_i^{-1})$ of Q_i exists.

ASSUMPTION 5: *There is C, $\delta > 0$ such that for all i, $\|\bar{\eta}_i\| \leq C$, $\|b_{2i}\| \leq C$*

$$E\left[\text{tr}\left(Q_i^{-1}\right)^{1+\delta}\right] \leq C.$$

Also, $E[s_i s_i'|x_i] \geq C^{-1}I$ with probability one for every i.

The existence of the $1+\delta$ moment of $tr(Q_i^{-1})$ will be implied by more primitive conditions. One example is where $b_{2i} \sim N(\mu_i, \Sigma_i)$ conditional on unobserved $\alpha_i = (\mu_i, \Sigma_i)$. In this case Q_i has a Wishart distribution conditional on α_i and hence Assumption 5 will be satisfied when T is sufficiently large relative to the dimension $\dim(b_{2i})$ of b_{2i}.

THEOREM 4: *If $T_i \geq \dim(b_{2i}) + 5$ for all i, there are unobserved random vectors μ_i and matrices Σ_i such that $b_{2i} \sim N(\mu_i, \Sigma_i)$ conditional on $\alpha_i = (\mu_i, \Sigma_i)$, and Σ_i is bounded and has smallest eigenvalue bounded away from zero uniformly in i, then there are $C, \delta > 0$ such that* $E\left[\text{tr}\left(Q_i^{-1}\right)^{1+\delta}\right] \leq C$.

This result gives primitive conditions for Assumption 5 when T is large enough relative to the number of regressors. The conditional Gaussian assumption allows for a wide range of possible distributions of the observed b_{2i} that could be nonsymmetric and vary across time periods in general ways. Boundedness of the smallest eigenvalue of Σ_i away from zero does mean that there is variation in the regressors after conditioning on individual effects.

The next result shows asymptotic normality of the average ridge coefficients and BCS and that using the estimated variances results in correct inference in large samples. Let $\bar{\eta} = E[\eta_{it}]$ from Assumption 4 and $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2')'$ be as defined in equations (4.1) and (4.2).

THEOREM 5: *If Assumptions 4 and 5 are satisfied and $\sqrt{n}\lambda \longrightarrow 0$ then there is $C > 0$ with $Var(\hat{\eta}) \geq CI$ for all n large enough and for any $s \times p$ matrix D with $rank(D) = s$,*

$$[DVar(\hat{\eta})D']^{-1/2} D\sqrt{n}(\hat{\eta} - \bar{\eta}) \xrightarrow{d} N(0, I_p), \ [D\hat{V}_\eta D']^{-1/2} D\sqrt{n}(\hat{\eta} - \bar{\eta}) \xrightarrow{d} N(0, I_p).$$

Also if $\beta_0 = \lim_{n\longrightarrow\infty}[\sum_{i=1}^n \sum_{t=1}^{T_i} E[b(x_{it})']/(nT_i)]'\bar{\eta}$ exists then

$$\hat{V}^{-1/2}\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, 1)$$

This result differs from previous work in applying to the debiased average ridge estimator, in having T big enough relative to p so that the estimator is root-n consistent (unlike most of Graham and Powell, 2012), and in the parameter of interest being the average of coefficients over the whole population rather than over a subset where $\det(b_i'b_i)$ is large enough (unlike Arellano and Bonhomme, 2012). In our empirical work we will use this result to make inference about average elasticities for income and prices and for BCS.

5 Application to Scanner Data

The data we used is a subset of the Nielsen Homescan Panel like that of Burda, Harding, and Hausman (2008, 2012). The data include 1483 households from the Houston-area zip codes for the years 2004-2006. The number of monthly observations for each household ranges from 12 to 36, with some households being added and taken away throughout the 3 years we covered and 609 households being included the entire time. At several points in the empirical analysis we checked for differences in results between using all households and the 609 that were always present and found no statistically significant differences. This lack of sensitivity to the length of panels used in estimation suggests an absence of attrition and selection bias in this data, similar to Wooldridge (2018).

Expenditures are total over all purchases of the household bought in each month. The original data had timestamps for purchases. If a household purchased something more than once in a month, the "monthly price" is the average price that the household purchased (i.e. total amount spent on good/total quantity purchased).

Including zero expenditures makes it necessary to impute prices for those times periods where an individual purchased none of a particular good. We tried two ways of imputing the missing prices. We tried replacing the missing price with the price paid last time the good was purchased, or the price next paid if there was no past purchase, or the average price paid for that good that month at stores frequented by the consumer if there was no purchases. We also tried just imputing the price to be the average price paid that month at stores frequented by the consumer. These two imputation methods produced similar results for soda demand, so we used the first method of imputing price, involving past or future purchase prices.

We included prices for 15 groups of goods; soda, milk, soup, water, butter, cookies, eggs, orange juice, ice cream, bread, chips, salad, yogurt, coffee, and cereal. As in Burda, Harding, and Hausman (2008, 2012) we chose these groups because they made up a relatively large proportion of total expenditure. The data also includes demographics such as race, marital status, household composition, as well as male and female employment status. We use family size as a covariate, after finding that other family composition covariates are not important.

We also use income as an instrument for total expenditure in linear share regressions and to construct a control variable for total expenditure in the Lasso estimates. The income variable is the integer denoting which of 20 categories the household income fell in, excluding one category.. Such a category counter will be a valid instrument whenever income is independent of the disturbance in the share equation disturbance. Also, as shown by Masten and Torgovitzky (2015), categorical instruments are allowed for in the specification of a control variable.

We first give standard share regression results. Table 1 gives the soda and milk expenditure and own price elasticities from regressing the share of (soda and milk) expenditure on the natural log of prices and of total expenditure. The results in the table are for OLS and IV

where total expenditure is instrumented by the income variable. We find that instrumenting for total expenditure has very little effect on price elasticities and the total expenditure elasticity for milk, but does have some effect on the expenditure elasticity for soda. The IV coefficient .932 is quite different than the OLS .683, with the $N(0,1)$ Hausman test statistic for the difference of OLS and IV expenditure elasticities for soda being -1.41. Although this is not significant at conventional levels the economic reasons for endogeneity of total expenditure are important enough that we will correct for endogeneity for our BCS estimates by using a control variable.. All the standard errors for the cross-section estimates correct for clustering that could arise from correlation of individual observations over time.

Table 1: Cross Section Elasticities For Each Good

	OLS				IV			
	Exp	S.E.	Own P.	S.E.	Exp	S.E.	Own P.	S.E.
soda	.683	.015	-.855	.020	.932	.177	-.867	.052
milk	.539	.024	-1.416	.020	.570	.154	-1.412	.064

Table 2 gives the soda share cross price elasticities and their standard errors as well as the soda results from Table 1. We do find that cross-price elasticities are much smaller than own price elasticities, which motivates our use of Lasso in the cross section estimation of BCS.

Table 2: Cross-price Elast for Soda

	OLS		IV	
	Elast.	S.E.	Elast.	S.E.
exp	.683	.015	.932	.177
soda	-.855	.020	-.867	.052
soup	.028	.022	.040	.056
water	.034	.012	.032	.039
butter	-.177	.013	-.177	.040
cookies	-.028	.014	-.046	.035
eggs	-.074	.022	-.080	.049
oj	.040	.027	.015	.089
ice cream	.177	.023	.174	.076
bread	-.127	.019	-.153	.055
chips	-.006	.023	-.016	.055
milk	-.076	.026	-.039	.079
salad	-.084	.014	-.092	.041
yogurt	.032	.025	.025	.075
coffee	-.074	.011	-.068	.032
cereal	.087	.021	.041	.059

We estimated the BCS for a 10% price increase in soda with starting price being the sample mean of soda price observed in the data. We did Lasso regression of soda expenditure share on ln(prices), ln(total expenditure), powers of logs of own price, total expenditure, and the control function up to order 4, quadratic terms in own other prices, an interaction of the log of total expenditure and household size, and an interaction of the control function with log total expenditure. We estimated the BCS by total expenditure groups, obtaining estimates for those in the lowest and highest quartile of total expenditure as well as overall averages. In the estimation we use the cross-validated choice of penalty for the share regression and vary the penalty for the estimation of the Riesz representer. The results for all households and for the lower quartile did not change much with the penalty. For the upper quartile the BCS changed by slightly less than 10 percent as we varied the penalty. For the BCS the income effect lower bound was taken to be zero and upper bound to be 20 times the maximum of the income effect over .1, .25, .5, .75, .9 quantile effect regressions, similar to Hausman and Newey (2016). In the results we only report one bound because the lower and upper estimated bounds were equal to four significant digits. Wider bounds based on assuming that all goods are normal would have upper bounds equal to the reported one and lower bounds being 90 percent of the reported bounds, corresponding to the 10 percent price change we are considering.

The BCS estimates in terms of annual dollars and their standard errors are given in Table 3.

Table 3: Cross Section Estimates of Surplus Bounds			
	All households	Lower Quartile	Upper Quartile
BCS	12.97	5.43	17.28
S.E.	.34	.14	.44

Despite the high dimensional potentially nonparametric specification the BCS are very precisely estimated from this data, consistent with the averaging over many individuals and time periods. As with the elasticity estimates from Tables 1 and 2 we accounted for clustering by individuals in the standard errors. The average monthly total food expenditure for all individuals is 621, for upper quartile 1265, for lower is 190. The ratio of the BCS to the average total food expenditure is $17.28/1265 = .01366$ for the upper quantile and $5.43/190 = 0.02858$ for the lower quartile. Thus we find that the average surplus, i.e. average welfare cost, of a 10 percent soda price increase relative to average expenditure, is estimated to be much higher for individuals in the lowest quartile of the total food expenditure distribution. Using the more conservative bounds based on all goods being normal does not change this conclusion. The lower bound for the average surplus relative to expenditure in the lowest expenditure quartile would be $.9 * (.02858) = .02572$ which is still much larger than the upper bound .01366 for the upper quartile. We also estimated the BCS using many more regressors including interactions among all different prices and the results were found to be very close to those reported here.

Turning to the panel data results, Table 4 gives standard fixed effects estimates (allowing an individual specific constant) of soda and milk own price and total expenditure elasticities, without instrumenting for total expenditure. We found no evidence for time trends in this data, so we report results for fixed effects without any treands and for the time homogenous model we have considered. The standard errors here allow for general dependence over time for each individual. We find that the fixed effect own price elasticities are much smaller than the cross-section. For example, the panel milk price elasticity is about half the size of the cross-section, with the panel estimate being more reasonable in size. Standard errors are not very much larger than the cross-section elasticities.

Table 4: Fixed Effects Elasticities				
	Exp	S.E.	Own P.	S.E.
soda	.638	.018	-.689	.022
milk	.430	.042	-.699	.033

Table 5 gives the elasticities corresponding to bias corrected averages of individual ridge coefficient estimates for $\lambda = .05$ and $\lambda = .005$. We find that these elasticities are much smaller than the fixed effects estimates. The $-.364$ own price elasticity from the average coefficient for soda is slightly more than 1/3 the size of the corresponding cross-section elasticity and somewhat more than half the size of the fixed effects estimate. Evidently allowing for individual price and expenditure coefficients that can be correlated with prices and expenditure has strong effects on elasticity estimates.

Table 5: Average Ridge Elasticities								
λ	.05				.0005			
	Exp	S.E.	Own P.	S.E.	Exp	S.E.	Own P.	S.E.
soda	.615	.012	-.558	.017	.595	.021	-.364	.056
milk	.445	.014	-.652	.013	.437	.020	-.508	.052

Table 6 gives panel estimates of the BCS for all individuals, the highest quartile of the total expenditure, and the lowest quartile.

Table 6: Panel Estimates of Surplus Bounds, $\lambda = .05$			
	All households	Lower Quartile	Upper Quartile
BCS	8.70	3.90	15.43
S.E.	.49	.21	1.25

The income grouped panel estimates are quite similar to the corresponding cross-section, although BCS for all households is quite a bit smaller for the panel estimates. The ratio of the BCS to the average total food expenditure is $15.43/1265 = .01220$ for the highest quartile

and $3.90/190 = 0.02053$ for the lowest quartile. Here the discrepancy of average surplus/total expenditure between income groups is not as large as in the cross-section. This discrepancy does still persist when we consider the wider bounds based on all goods being normal, as it does for the cross-section estimates.

A particularly striking empirical finding is the small size of the elasticity average relative to the elasticities obtained in the cross section data. The way that panel estimates change with the regularization parameter λ helps describe these differences. The decrease in the price elasticity as λ decreases is consistent with lower elasticities being associated with individuals with less price variation. As we have discussed, the bias corrected ridge estimator gives more weight to individuals where there is more variation in the regressors. As λ decreases there is less variation in the weights, approaching an equal weighted average as λ goes to zero. Observations with less variation in prices receive more weight for smaller lamdas. Thus, the elasticity reduction as λ goes from .05 to .0005 is consistent with individuals with lower elasticities also having less price variation.

The difference between the fixed effects and bias corrected ridge estimates are also consistent with this pattern. For simplicity we explain when there is a single nonconstant regressor X_{it} and the number of time periods is the same for each individual. Wooldridge (2005) used a similar calculation to get conditions for consistency of the fixed effects estimator when slopes are varying. Our purpose is to find conditions for the fixed effects slope estimator to be downward biased, so that fixed effects elasticities are more negative than, i.e. larger in absolute value, than the elasticity for the average slope.

A model with varying slopes and a single regressor is

$$s_{it} = \alpha_i + \beta_i X_{it} + \varepsilon_{it},\ E[\varepsilon_{it}|X_i] = 0,\ (t = 1, ..., T;\ i = 1, ..., n).$$

Taking deviations from individual means gives

$$\tilde{s}_{it} = \beta_i \tilde{X}_{it} + \tilde{\varepsilon}_{it},$$

where $\tilde{s}_{it} = s_{it} - \bar{s}_i$, $\tilde{X}_{it} = X_{it} - \bar{X}_i$, $\tilde{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$, and $\bar{s}_i$, $\bar{X}_i$, and $\bar{\varepsilon}_i$ are time means. Adding and subtracting $\bar{\beta}\tilde{X}_{it}$, for $\bar{\beta} = E[\beta_i]$, gives

$$\tilde{s}_{it} = \bar{\beta}\tilde{X}_{it} + \tilde{\varepsilon}_{it} + \tilde{X}_{it}(\beta_i - \bar{\beta}),$$

The usual regression analysis implies that when the data is i.i.d. across individuals

$$p\lim(\hat{\beta}_{FE}) = \bar{\beta} + \frac{E[(\sum_t \tilde{X}_{it}^2/T)(\beta_i - \bar{\beta})]}{E[(\sum_t \tilde{X}_{it}^2/T)]}.$$

Thus, the fixed effects estimator of the price elasticity being larger in magnitude that the average elasticity across individuals (i.e. $p\lim(\hat{\beta}_{FE}) < \bar{\beta}$) is associated with

$$E[(\sum_t \tilde{X}_{it}^2/T)(\beta_i - \bar{\beta})] < 0,$$

that is the sample time variance of log of prices is larger when the elasticity is larger in magnitude. This bias characterization for the fixed effects estimator seems consistent with search behavior of consumers, where those with higher elasticities have high dispersion of prices over time due to more search.

A potential alternative explanation for the decrease in elasticities when going from least squares to fixed effects is measurement error as in Griliches and Hausman (1986). There is some possibility of measurement error in prices because of missing price data for zeros and because some averaging over prices at different stores is done in data collection. To see if accounting for measurement error changed the estimates we tried estimating individual coefficients while instrumenting the own price by the four month lag of the own price, a type of instrument considered by Griliches and Hausman (1986). We report averages of individual, ridge regularized two-stage least squares estimates, where we bias correct the average. The estimates are given in Table 6. There is a slight increase in the estimated average price elasticity of soda although the increase is small and not enough to explain the large differences between cross-section and panel elasticities.

Table 6: Ridge Elasticities with Four Month Own Lag Price as IV

λ	.05				.0005			
	Exp	S.E.	Own P.	S.E.	Exp	S.E.	Own P.	S.E.
soda	.615	.013	-.653	.011	.585	.029	-.392	.055
milk	.446	.014	-.705	.008	.422	.021	-.448	.048

These results suggest that the large differences between cross section and panel elasticity estimates cannot be explained by measurement error.

We also conducted some small Monte Carlo experiments to determine whether the smaller elasticities could result from finite sample bias. We did find small finite sample bias in the bias corrected average ridge estimates with endogenous X_{it} but not large enough to explain the small average elasticities. We conclude that there is strong evidence in the data that prices are correlated with preferences, and that the large differences in cross-section and panel elasticities are not explained by either measurement error or bias in the average ridge estimator.

6 Conclusion

In this paper we have found large differences between cross-section and panel price elasticity estimates, where the panel estimates allow coefficients to vary over individuals. These findings provide strong evidence that individual preferences are correlated with prices in the Nielsen scanner data. The BCS estimates by expenditure group appear to be less sensitive, with cross-section and panel estimates being similar.

The estimators solve the zeros problem by simply including zero expenditure values for the left had side variable. The allowance for general, nonparametric heterogeneity facilitates this solution to the zeros problem. We allow for many prices by using debiased machine learning in the cross-section and bias corrected ridge regularization for panel data. The cross-section allows endogeneity of total expenditure and potentially of prices via inclusion of control functions. In these ways we provide useful approaches to estimation of demand models in large data sets with many prices, where prices may be correlated with preferences.

7 Appendix A: Assumptions A1 and A2.

In this brief Appendix we give Assumptions A1 and A2 which are used in the result for the DML estimator of Section 3. We give only a brief discussion here. A more extensive discussion can be found in Chernozhukov, Newey, and Singh (2018). Assumption A1 gives an approximation rate hypothesis for both the regression $\gamma_0(x,v)$ and the Riesz representer $\alpha_0(x,v)$.

ASSUMPTION A1: *There exists* $C > 0$, $\bar{\rho}$, *and* $\bar{\delta}$ with $\bar{s}_\delta$ *nonzero elements such that* $\sum_{j=1}^J |\bar{\rho}_j| \leq C$, $\sum_{j=1}^J \left|\bar{\delta}_j\right| \leq C$, and $\|\alpha_0 - b'\bar{\rho}\|^2 \leq C\sqrt{\ln(J)/n}$, $\left\|\gamma_0 - b'\bar{\delta}\right\|^2 \leq C\bar{s}_\delta \ln(J)/n$.

The next Assumption gives a sparse eigenvalue conditions that is common in the Lasso literature

ASSUMPTION A2: $G = E[b(w_i)b(w_i)']$ *is nonsingular and has largest eigenvalue uniformly bounded in* n. *Also there is* $k > 3$ *such that for* $\tilde{\delta} = \arg\min \left\{\|\gamma_0 - b'\delta\|^2 + r_\delta |\delta|_1\right\}$ *and* $\mathcal{J} = \{j : \tilde{\delta}_j \neq 0\}$,

$$\inf_{\{\delta:\delta\neq 0, \sum_{j\in\mathcal{J}^c}|\delta_j| \leq k\sum_{j\in\mathcal{J}}|\delta_j|\}} \frac{\delta' G\delta}{\sum_{j\in\mathcal{J}}\delta_j^2} > 0.$$

8 Appendix B: Proofs of Theorems

Proof of Lemma 1: In this proof let J denote the number of goods and $q_j(p_1, y)$ denote the demand for the jth good as a function of p_1 and y holding all other prices p_2 and covariates w fixed. Then by local nonsatiation,

$$y = \sum_{j=1}^J p_j q_j(p_1, y), y - \Delta y = \sum_{j=1}^J p_j q_j(p_1, y - \Delta y).$$

Subtracting the second equation from the first gives

$$\Delta y = \sum_{j=1}^J p_j[q_j(p_1,y) - q_j(p_1, y-\Delta y)] \geq p_1[q_1(p_1,y) - q_1(p_1,y-\Delta y)] \geq \check{p}_1[q_1(p_1,y) - q_1(p_1,y-\Delta y)],$$

where the first inequality follows by all goods being normal goods and the second by $p_1 \in [\check{p}_1, \bar{p}_1]$. Dividing through by $\check{p}_1$ gives $\bar{b} = 1/\check{p}_1$. Also $b = 0$ follows by q_1 being a normal good. *Q.E.D.*.

The following two results are useful in the proof of Theorem 2. Let $a_j^x(\tilde{x}) = \zeta(\tilde{x}) b_j^x(\tilde{x})$, $\bar{a}_j^x = E[a_j^x(\tilde{x}_i)]$, $\bar{b}_j^v = E[b_j^v(v_i)]$, and $M_j = \bar{a}_j^x \bar{b}_j^v$, $(j = 1, ..., J)$.

LEMMA A1: *If Assumption 3 is satisfied then for any two empirical CDF's $\hat{F}$ and $\tilde{F}$ for subsamples with sample sizes greater than Cn for some C,*

$$\max_{j \leq J} \left| \hat{M}_j - M_j \right| = O_p(\sqrt{\ln(J)/n}), \quad \int \zeta(\tilde{x}) \hat{\gamma}(\tilde{x}, v) (\hat{F} - F_0)(d\tilde{x})(\tilde{F} - F_0)(dv) = o_p(n^{-1/2}).$$

Proof: By Assumptions 2 and 3 $\zeta(\tilde{x})$, $b_j^x(\tilde{x})$, and $b_j^v(v)$ are each bounded uniformly in j, $\tilde{x}$, and v so that $a_j^x(\tilde{x})$ is also. Then by Assumption 2 and standard maximal inequality arguments,

$$\max_{k \leq J} \left| \int a_k^x(\tilde{x})(\hat{F} - F_0)(d\tilde{x}) \right| = O_p(\sqrt{\frac{\ln(J)}{n}}), \quad \max_{k \leq J} \left| \int b_k^v(v)(\hat{F} - F_0)(dv) \right| = O_p(\sqrt{\frac{\ln(J)}{n}}).$$

Then we have

$$\begin{aligned} \left| \hat{M}_j - M_j \right| \leq & \left| \int a_j^x(\tilde{x})(\hat{F} - F_0)(d\tilde{x}) \right| \left| \int b_j^v(v)(\hat{F} - F_0)(dv) \right| \\ & + \left| \int a_j^x(\tilde{x})(\hat{F} - F_0)(d\tilde{x}) \right| \left| \bar{b}_j^v \right| + \left| \bar{a}_j^x \right| \left| \int b_j^v(v)(\hat{F} - F_0)(dv) \right|, \end{aligned}$$

so the first conclusion follows by $\left| \bar{b}_j^v \right|$ and $\left| \bar{a}_j^x \right|$ uniformly bounded in j. Also by Lemma A3 of Chernozhukov, Newey, and Singh (2018), $\sum_{k=1}^J \left| \hat{\delta}_k \right| = \left| \hat{\delta} \right|_1 = O_p(1)$. Then by Assumptions 2 and 3 it follows that

$$\begin{aligned} & \left| \int \zeta(x) \hat{\gamma}(x, v) (\hat{F} - F_0)(dx)(\tilde{F} - F_0)(dv) \right| = \left| \sum_{k=1}^J \hat{\delta}_k [\int a_k^x(x) (\hat{F} - F_0)(dx)][\int b_k^v(v)(\tilde{F} - F_0)(dv)] \right| \\ & \leq \left| \hat{\delta} \right|_1 \max_{k \leq J} \left| \int a_k^x(\tilde{x})(\hat{F} - F_0)(d\tilde{x}) \right| \max_{k \leq J} \left| \int b_k^v(v)(\tilde{F} - F_0)(dv) \right| = O_p(\frac{\ln(J)}{n}) = o_p(n^{-1/2}). Q.E.D. \end{aligned}$$

Proof of Theorem 2: We fix I_ℓ and let $\hat{F}$ be the empirical distribution over observations not in I_ℓ and $\tilde{F}$ be the empirical distribution over observations in I_ℓ. Also define $a_0(\tilde{x}, v) = \zeta(\tilde{x}) \gamma_0(\tilde{x}, v)$, $\hat{a}(\tilde{x}, v) = \zeta(\tilde{x}) \hat{\gamma}(\tilde{x}, v)$,

$$\begin{aligned} T_1 &= \int \hat{a}(\tilde{x}, v) \left(\hat{F} - F_0 \right)(d\tilde{x}) \left(\hat{F} - F_0 \right)(dv), \\ T_2 &= \int \hat{a}(\tilde{x}, v) \left(\tilde{F} - F_0 \right)(d\tilde{x}) \hat{F}(dv) - \int a_0(\tilde{x}, v) \left(\tilde{F} - F_0 \right)(d\tilde{x}) F_0(dv), \\ T_3 &= \int \hat{a}(\tilde{x}, v) \hat{F}(d\tilde{x}) \left(\tilde{F} - F_0 \right)(dv) - \int a_0(\tilde{x}, v) F_0(d\tilde{x}) \left(\tilde{F} - F_0 \right)(dv). \end{aligned}$$

Lemma A1 implies that $T_1 = o_p(n^{-1/2})$. Also, for $\tilde{T}_2 = \int [\hat{a}(\tilde{x}, v) - a_0(\tilde{x}, v)] \left(\tilde{F} - F_0\right)(d\tilde{x}) F_0(dv)$ we have that

$$T_2 = \tilde{T}_2 + \int \hat{a}(\tilde{x}, v) \left(\tilde{F} - F_0\right)(d\tilde{x}) \left(\hat{F} - F_0\right)(dv) = \tilde{T}_2 + o_p(n^{-1/2}),$$

where the last equality follows by Lemma A1. Furthermore, for the estimation sample $\hat{I}$ for $\hat{\gamma}$,

$$\begin{aligned} E\left[n\tilde{T}_2^2 \mid \hat{I}\right] &\leq \int \left\{[\hat{a}(\tilde{x}, v) - a_0(\tilde{x}, v)] F_0(dv)\right\}^2 F_0(d\tilde{x}) \leq \int [\hat{a}(\tilde{x}, v) - a_0(\tilde{x}, v)]^2 F_0(dv) F_0(d\tilde{x}) \\ &\leq C \int [\hat{\gamma}(\tilde{x}, v) - \gamma_0(\tilde{x}, v)]^2 F_0(dv) F_0(d\tilde{x}) \leq C \int [\hat{\gamma}(x, v) - \gamma_0(x, v)]^2 F_0(dx, dv) \xrightarrow{p} 0, \end{aligned}$$

where the last inequality follows by Assumption 3. Therefore, by the conditional Markov inequality, $\tilde{T}_2 = o_p\left(n^{-1/2}\right)$. It then follows by the triangle inequality that $T_2 = o_p\left(n^{-1/2}\right)$. An analogous argument also gives $T_3 = o_p\left(n^{-1/2}\right)$.

Let $\hat{\beta}_\ell$ be the average over I_ℓ, so that

$$\hat{\beta} = \sum_{\ell=1}^{L} \frac{n_\ell}{n} \hat{\beta}_\ell.$$

In this notation,

$$\begin{aligned} \hat{\beta}_\ell &= \int \hat{a}(\tilde{x}, v) \tilde{F}(d\tilde{x}) \hat{F}(dv) + \int \hat{a}(\tilde{x}, v) \hat{F}(d\tilde{x}) \tilde{F}(dv) - \int \hat{a}(\tilde{x}, v) \hat{F}(d\tilde{x}) \hat{F}(dv) \\ &\quad + \int \hat{\alpha}(x, v) [y - \hat{\gamma}(x, v)] \tilde{F}(dz) \\ &= \int \hat{a}(\tilde{x}, v) F_0(d\tilde{x}) \hat{F}(dv) + \int \hat{a}(\tilde{x}, v) (\tilde{F} - F_0)(d\tilde{x}) \hat{F}(dv) \\ &\quad + \int \hat{a}(\tilde{x}, v) \hat{F}(d\tilde{x}) F_0(dv) + \int \hat{a}(\tilde{x}, v) \hat{F}(d\tilde{x}) (\tilde{F} - F_0)(dv) - \int \hat{a}(\tilde{x}, v) \hat{F}(d\tilde{x}) \hat{F}(dv) \\ &\quad - \int \hat{a}(\tilde{x}, v) F_0(d\tilde{x}) F_0(dv) + \int \hat{a}(\tilde{x}, v) F_0(d\tilde{x}) F_0(dv) + \int \hat{\alpha}(x, v) [y - \hat{\gamma}(x, v)] \tilde{F}(dz). \\ &= T_1 + T_2 + T_3 + \int a_0(\tilde{x}, v) \left(\tilde{F} - F_0\right)(d\tilde{x}) F_0(dv) + \int a_0(\tilde{x}, v) F_0(d\tilde{x}) \left(\tilde{F} - F_0\right)(dv) \\ &\quad + \int \hat{a}(\tilde{x}, v) F_0(d\tilde{x}) F_0(dv) + \int \hat{\alpha}(x, v) [y - \hat{\gamma}(x, v)] \tilde{F}(dz). \end{aligned}$$

It follows from the above reasoning and the triangle inequality that $T_1 + T_2 + T_3 = o_p(n^{-1/2})$.

Next, for $w = (x, v)$ let

$$\begin{aligned} T_1' &= \int [\hat{\alpha}(w) - \alpha_0(w)] [y - \gamma_0(w)] \tilde{F}(dz), T_2' = \int [\hat{\alpha}(w) - \alpha_0(w)] [\gamma_0(w) - \hat{\gamma}(w)] \tilde{F}(dz), \\ T_3' &= \int \alpha_0(w) [\gamma_0(w) - \hat{\gamma}(w)] (\tilde{F} - F_0)(dz). \end{aligned}$$

Then we have

$$\begin{aligned}\int \hat{\alpha}(w)\left[y-\hat{\gamma}(w)\right]\tilde{F}(dz) &= \int \hat{\alpha}(w)\left[y-\gamma_0(w)\right]\tilde{F}(dz)+\int \hat{\alpha}(w)\left[\gamma_0(w)-\hat{\gamma}(w)\right]\tilde{F}(dz)\\ &= \int \alpha_0(w)\left[y-\gamma_0(w)\right]\tilde{F}(dz)+T_1'+\int \alpha_0(w)\left[\gamma_0(w)-\hat{\gamma}(w)\right]\tilde{F}(dz)+T_2'\\ &= \int \alpha_0(w)\left[y-\gamma_0(w)\right]\tilde{F}(dz)+\int \alpha_0(w)\left[\gamma_0(w)-\hat{\gamma}(w)\right]F_0(dz)\\ &+T_1'+T_2'+T_3'.\end{aligned}$$

Note that Assumption 1 Chernozhukov et al. (2018, CNS) is satisfied with $b_j(x,v)$ bounded uniformly in j, Assumption 2 of CNS holds by the first part of Lemma A1, and Assumption 3 of CNS holds by Assumption A1. Then by Theorem 2 of CNS,

$$\|\hat{\alpha}-\alpha_0\|^2 = O_p(r_\rho) \xrightarrow{p} 0,$$

where $\|\hat{\alpha}-\alpha_0\|^2 = \int[\hat{\alpha}(x,v)-\alpha_0(x,v)]^2F_0(dx,dv)$. This result implies $T_1' = o_p\left(n^{-1/2}\right)$ as in CNS. Also, by Assumptions A1 and A2 and Theorem 3 of CNS,

$$\|\hat{\gamma}-\gamma_0\|^2 = O_p(\bar{s}_\delta r_\delta^2) \xrightarrow{p} 0,$$

implying $T_3' = o_p\left(n^{-1/2}\right)$ as in CNS. It also follows that

$$\sqrt{n}\,\|\hat{\alpha}-\alpha_0\|\,\|\hat{\gamma}-\gamma_0\| = O_p(\sqrt{n}(\bar{s}_\delta)^{1/2}r_\delta r_\rho^{1/2}) \xrightarrow{p} 0,$$

so that $T_2' = o_p\left(n^{-1/2}\right)$. We also have

$$\begin{aligned}\beta_0 &= \int a_0(\tilde{x},v)F_0(d\tilde{x})\,F_0(dv) = \int \alpha_0(w)\gamma_0(w)F_0(dw),\\ \int \hat{a}(\tilde{x},v)F_0(d\tilde{x})\,F_0(dv) &= \int \alpha_0(w)\hat{\gamma}(w)F_0(dw).\end{aligned}$$

Therefore we have

$$\begin{aligned}\hat{\beta}_\ell &= \int a_0(\tilde{x},v)\left(\tilde{F}-F_0\right)(d\tilde{x})\,F_0(dv)+\int a_0(\tilde{x},v)F_0(d\tilde{x})\left(\tilde{F}-F_0\right)(dv)+\int \hat{a}(\tilde{x},v)F_0(d\tilde{x})\,F_0(dv)\\ &+\int \alpha_0(w)\left[y-\gamma_0(w)\right]\tilde{F}(dz)+\int \alpha_0(w)\left[\gamma_0(w)-\hat{\gamma}(w)\right]F_0(dz)+o_p\left(n^{-1/2}\right)\\ &= \int a_0(\tilde{x},v)\left(\tilde{F}-F_0\right)(d\tilde{x})\,F_0(dv)+\int a_0(\tilde{x},v)F_0(d\tilde{x})\left(\tilde{F}-F_0\right)(dv)+\beta_0\\ &+\int \alpha_0(w)\left[\hat{\gamma}(w)-\gamma_0(w)\right]F_0(dw)\\ &+\int \alpha_0(w)\left[y-\gamma_0(w)\right]\tilde{F}(dz)+\int \alpha_0(w)\left[\gamma_0(w)-\hat{\gamma}(w)\right]F_0(dz)+o_p\left(n^{-1/2}\right)\\ &= \beta_0+\int a_0(\tilde{x},v)\left(\tilde{F}-F_0\right)(d\tilde{x})\,F_0(dv)+\int a_0(\tilde{x},v)F_0(d\tilde{x})\left(\tilde{F}-F_0\right)(dv)\\ &+\int \alpha_0(w)\left[y-\gamma_0(w)\right]\tilde{F}(dz)+o_p\left(n^{-1/2}\right).\end{aligned}$$

Then it follows that

$$\hat{\beta} = \beta_0 + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(z_i) + o_p(n^{-1/2}),$$
$$\psi(z_i) = \int a_0(\tilde{x}_i, v) F_0(dv) + \int a_0(\tilde{x}, v_i) F_0(d\tilde{x}) - 2\beta_0 + \alpha_0(w_i)\left[s_i - \gamma_0(w_i)\right].$$

The remainder of Theorem 2 follows similarly to CNS. Q.E.D.

Proof of Theorem 3: Note that for $\bar{\eta}_i = E[\eta_{it}|x_i]$,

$$E[s_{it}|x_i] = b(x_{it})' E[\eta_{it}|x_i] = b(x_{it})' \bar{\eta}_i.$$

Therefore

$$E[\hat{\eta}_{2i}|x_i] = \Lambda_i \tilde{b}_{2i}' E[s_i|x_i]/T_i = \Lambda_i Q_i \bar{\eta}_{2i} = W_i \bar{\eta}_{2i}.$$

The first conclusion then follows from independence of the observations which gives

$$E[\hat{\eta}_2|x_1, ..., x_n] = \left(\sum_{i=1}^{n} W_i\right)^{-1} \sum_{i=1}^{n} E[\hat{\eta}_{2i}|x_1, ..., x_n] = \left(\sum_{i=1}^{n} W_i\right)^{-1} \sum_{i=1}^{n} E[\hat{\eta}_{2i}|x_i].$$
$$= \left(\sum_{i=1}^{n} W_i\right)^{-1} \sum_{i=1}^{n} W_i \bar{\eta}_{2i}.$$

Similarly, for $\bar{\eta}_2 = E[\hat{\eta}_2|x_1, ..., x_n]$ it follows from $E[\bar{s}_i|x_i] = \bar{b}_i' \bar{\eta}_i$ that

$$E[\hat{\eta}_1|x_1, ..., x_n] = \frac{1}{n} \sum_{i=1}^{n} \{E[\bar{s}_i|x_i] - \bar{b}_{2i}'(E[\hat{\eta}_{2i}|x_i] + \lambda \Lambda_i \bar{\eta}_2)\}$$
$$= \frac{1}{n} \sum_{i=1}^{n} [\bar{b}_i' \bar{\eta}_i - \bar{b}_{2i}'(\Lambda_i Q_i \bar{\eta}_{2i} + \lambda \Lambda_i \bar{\eta}_2)]$$
$$= \frac{1}{n} \sum_{i=1}^{n} [\bar{\eta}_{1i} + \bar{b}_{2i}' \{(I - \Lambda_i Q_i) \bar{\eta}_{2i} - \lambda \Lambda_i \bar{\eta}_2\}]$$
$$= \frac{1}{n} \sum_{i=1}^{n} [\bar{\eta}_{1i} + \lambda \bar{b}_{2i}' \Lambda_i (\bar{\eta}_{2i} - \bar{\eta}_2)]..$$

Also, if $\bar{\eta}_{2i}$ does not depend on i so that $\bar{\eta}_{2i} = \bar{\eta}_2$ for all i then

$$E[\hat{\eta}_2|x_1, ..., x_n] = \left(\sum_{i=1}^{n} W_i\right)^{-1} \sum_{i=1}^{n} W_i \bar{\eta}_2 = \bar{\eta}_2.$$

Similarly we will have $E[\hat{\eta}_1|x_1, ..., x_n] = \bar{\eta}_1$. *Q.E.D.*.

Proof of Theorem 4: By boundedness of Σ_i and its smallest eigenvalue bounded away from zero uniformly in i it follows that there is C such that for $\Psi_i = \Sigma_i^{-1}$

$$tr(\Psi_i) \leq C$$

for all i with probability one (wp1). By b_{2i} Gaussian conditional on $\alpha_i = (\mu_i, \Sigma_i)$ is follows that $\mathcal{W}_i = T_i Q_i$ is Wishart with degrees of freedom $v_i = T_i - 1$ and variance matrix Σ_i, all conditional on α_i. Then by moment formulas for the inverse of a Wishart (Press, 1982) it follows that for $p = \dim(b_{2i})$ and each $j \in \{1, ..., p\}$,

$$\begin{aligned} E[(Q_i^{-1})_{jj}^2 | \alpha_i] &= Var((Q_i^{-1})_{jj} | \alpha_i) + E[(Q_i^{-1})_{jj}^2 | \alpha_i] \\ &= \frac{2\Psi_{i,jj}}{[v_i - p - 1]^2 [v_i - p - 3]} + \frac{\Psi_{i,jj}^2}{[v_i - p - 1]^2} \\ &= \frac{2\Psi_{i,jj}}{[T_i - p - 2]^2 [T_i - p - 4]} + \frac{\Psi_{i,jj}^2}{[T_i - p - 2]^2} \\ &\leq \frac{2C}{3^2} + \frac{C^2}{3^2} \leq C. \end{aligned}$$

By iterated expectations it follows that $E[(Q_i^{-1})_{jj}^2] = E[E[(Q_i^{-1})_{jj}^2 | \alpha_i]] \leq C$ for all i and j. Then $E[tr(Q_i^{-1})^{1+\delta}] \leq C$ for all i for $\delta = 1$. $Q.E.D.$

Proof of Theorem 5: By Assumption 5, Q_i^{-1} exists with probability one. Also, by $s_i s_i' \leq \|s_i\|^2 I$ and s_i bounded,

$$\|\hat{\eta}_{2i}\|^2 = \mathrm{tr}\left(\hat{\eta}_{2i}\hat{\eta}_{2i}'\right) = \mathrm{tr}\left(\Lambda_i \tilde{b}_{2i}' s_i s_i' \tilde{b}_{2i} \Lambda_i / T_i^2\right) \leq \mathrm{tr}\left(\Lambda_i Q_i \Lambda_i\right) \|s_i\|^2 / T_i \leq C \mathrm{tr}\left(\Lambda_i Q_i \Lambda_i\right).$$

For a symmetric square root $Q_i^{1/2}$ of Q_i it follows by $\Lambda_i \leq Q_i^{-1}$ that

$$\begin{aligned} \mathrm{tr}\left(\Lambda_i Q_i \Lambda_i\right) &= \mathrm{tr}\left(Q_i^{1/2} \Lambda_i^2 Q_i^{1/2}\right) \leq \mathrm{tr}\left(Q_i^{1/2} \Lambda^{1/2} Q_i^{-1} \Lambda_i^{1/2} Q_i^{1/2}\right) \\ &\leq \mathrm{tr}\, Q_i^{-1} \mathrm{tr}\left(Q_i^{1/2} \Lambda_i^{1/2} \Lambda_i^{1/2} Q_i^{1/2}\right) \leq \mathrm{tr}\, Q_i^{-1} \mathrm{tr}(Q_i^{1/2} Q_i^{-1} Q_i^{1/2}). \end{aligned}$$

Therefore

$$\|\hat{\eta}_{2i}\|^2 \leq C \left(\mathrm{tr}\, Q_i^{-1}\right).$$

Then for the δ of Assumption 5 it follows that

$$E[\|\hat{\eta}_{2i}\|^{2+2\delta}] < C.$$

Next, it follows by time stationarity that

$$E[\hat{\eta}_{2i} | x_i] = \Lambda_i \tilde{b}_{2i}' E[s_i | x_i] / T_i = \Lambda_i \tilde{b}_{2i}' b_i \bar{\eta}_i / T_i = \Lambda_i Q_i \bar{\eta}_{2i}.$$

By Λ_i positive definite it follows that $|(\Lambda_i)_{jk}| \leq ((\Lambda_i)_{jj} + (\Lambda_i)_{kk})/2$. Then by $\bar{\eta}_{2i}$ bounded we have

$$|\Lambda_i \bar{\eta}_{2i}|_\infty \leq C \max_j \sum_{k=1}^{T_i} |(\Lambda_i)_{jk}| \leq tr(\Lambda_i).$$

Also by $\Lambda_i Q_i - I = \lambda\Lambda_i$, it follows by the Holder and triangle inequalities that

$$\left|\frac{1}{\sqrt{n}}\sum_i (E[\hat{\eta}_{2i}|x_i] - \bar{\eta}_{2i})\right|_\infty = \left|\frac{1}{\sqrt{n}}\sum_i (\Lambda_i Q_i - I)\right|_\infty = \left|\sqrt{n}\lambda\frac{1}{n}\sum_i \Lambda_i\bar{\eta}_{2i}\right|_\infty$$
$$\leq o(1)\frac{1}{n}\sum_i |\Lambda_i\bar{\eta}_{2i}|_\infty \leq o(1)\frac{1}{n}\sum_i E[tr(\Lambda_i)]$$
$$\leq o(1)\frac{1}{n}\sum_i tr(Q_i^{-1}) = o(1)O_p(1) = o_p(1).$$

It follows similarly that

$$\hat{B} = I + o_p(n^{-1/2}).$$

It then follows that

$$\sqrt{n}(\hat{\eta}_2 - \bar{\eta}_2) = \sqrt{n}[I + o_p(n^{-1/2})]\frac{1}{n}\sum_i (\hat{\eta}_{2i} - \bar{\eta}_2)$$
$$= \frac{1}{\sqrt{n}}\sum_i (\hat{\eta}_{2i} - E[\hat{\eta}_{2i}]) + \frac{1}{\sqrt{n}}\sum_i (E[\hat{\eta}_{2i}] - \bar{\eta}_2) + o_p(1)$$
$$= \frac{1}{\sqrt{n}}\sum_i (\hat{\eta}_{2i} - E[\hat{\eta}_{2i}]) + o_p(1).$$

Next, by s_{it} and $b(x_{it})$ bounded there is C such that for all i,
$\hat{\eta}_{1i} = \bar{s}_i - \bar{b}_{2i}'(\hat{\eta}_{2i} + \lambda\Lambda_i\hat{\eta}_2)$

$$E[|\hat{\eta}_{1i}|^{2+2\delta}] \leq C(1 + E[\|\hat{\eta}_{2i}\|^{2+2\delta}] + \lambda) < C.$$

Also, by $E[\bar{s}_i|x_i] = \bar{b}_i'\bar{\eta}_i = \bar{\eta}_{1i} + \bar{b}_{2i}'\bar{\eta}_{2i}$,

$$\left|\frac{1}{\sqrt{n}}\sum_i (E[\hat{\eta}_{1i}] - \bar{\eta}_1)\right| \leq \frac{1}{\sqrt{n}}\sum_i \left|E[(\bar{s}_i - \bar{\eta}_{1i} - \bar{b}_{2i}'\hat{\eta}_{2i})]\right| = \frac{1}{\sqrt{n}}\sum_i \left|E[E[\bar{s}_i|x_i] - \bar{\eta}_{1i} - \bar{b}_{2i}'E[\hat{\eta}_{2i}|x_i]]\right|$$
$$= \frac{1}{\sqrt{n}}\sum_i \left|E[\bar{b}_{2i}'(I - \Lambda_i Q_i)\bar{\eta}_{2i}]\right| = \sqrt{n}\lambda\frac{1}{n}\sum_i \left|E[\bar{b}_{2i}'\Lambda_i\bar{\eta}_{2i}]\right| = o_p(1).$$

It then follows similarly to the above that

$$\sqrt{n}(\hat{\eta}_1 - \bar{\eta}_1) = \frac{1}{\sqrt{n}}\sum_i (\hat{\eta}_{1i} - E[\hat{\eta}_{1i}]) + o_p(1).$$

For $0 < \varepsilon < 1$ let $\mathcal{A}_i$ denote the event that $tr(Q_i^{-1}) \leq (1-\varepsilon)/(\varepsilon\lambda)$. Note that $\lambda_{\max}(Q_i) \leq C$ by b_i bounded. Then we have

$$tr(Q_i^{-1}) \leq \frac{1-\varepsilon}{\varepsilon\lambda} \Longrightarrow \lambda_{\max}(Q_i^{-1}) \leq \frac{1-\varepsilon}{\varepsilon\lambda} \Longrightarrow \lambda_{\min}(Q_i) \geq \frac{\varepsilon\lambda}{1-\varepsilon}$$
$$\Longrightarrow Q_i \geq \frac{\varepsilon\lambda}{1-\varepsilon}I \Longrightarrow Q_i \geq \varepsilon\Lambda_i^{-1}$$
$$\Longrightarrow \Lambda_i Q_i \Lambda_i \geq \varepsilon\Lambda_i \geq \varepsilon^2 Q_i^{-1} \geq \varepsilon^2\lambda_{\min}(Q_i^{-1})I = \varepsilon^2\frac{1}{\lambda_{\max}(Q_i)}I \geq CI.$$

Therefore there exists c such that

$$1(\Lambda_i Q_i \Lambda_i \geq cI) \geq 1(\mathcal{A}_i).$$

Furthermore, note that by the Markov inequality and Assumption 5,

$$\Pr(\mathcal{A}_i) = 1 - \Pr(\mathcal{A}_i^c) \geq 1 - \frac{\lambda\varepsilon E[tr(Q_i^{-1})]}{1-\varepsilon} \geq 1 - C\lambda \tag{8.1}$$

Next consider $\hat{\eta}_i = (\hat{\eta}_{1i}, \hat{\eta}_{2i}')'$. For convenience we will neglect the $\bar{b}_{2i}'\lambda\Lambda_i\hat{\eta}_2$ term in $\hat{\eta}_{1i}$, which will be asymptotically negligible due to $\lambda \longrightarrow 0$. Note that

$$\hat{\eta}_i = F_i s_i,\ F_i = [F_{1i}', F_{2i}']',\ F_{1i} = T_i^{-1}[e_{T_i}' - \bar{b}_{2i}'\Lambda_i\tilde{b}_{2i}'],\ F_{2i} = T_i^{-1}\Lambda_i\tilde{b}_{2i}'.$$

Then by Assumption 5, for $B_i' = [-\bar{b}_{2i}, I]$

$$\begin{aligned} Var(\hat{\eta}_i|x_i) = F_i Var(s_i|x_i)F_i' \geq cF_iF_i' &= T_i^{-1}\begin{bmatrix} 1+\bar{b}_{2i}'\Lambda_i Q_i\Lambda_i\bar{b}_{2i} & -\bar{b}_{2i}'\Lambda_i Q_i\Lambda_i \\ -\Lambda_i Q_i\Lambda_i\bar{b}_{2i} & \Lambda_i Q_i\Lambda_i \end{bmatrix} \\ &= T_i^{-1}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + T_i^{-1}B_i\Lambda_i Q_i\Lambda_i B_i' \geq C\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + c1(\mathcal{A}_i)B_iB_i' \\ &\geq C1(\mathcal{A}_i)\begin{bmatrix} 1+\bar{b}_{2i}'\bar{b}_{2i} & -\bar{b}_{2i}' \\ -\bar{b}_{2i} & I \end{bmatrix} \geq C1(\mathcal{A}_i)I, \end{aligned}$$

where the last inequality follows by $\bar{b}_{2i}$ uniformly bounded. Also we have

$$Var(\hat{\eta}_i) = E[Var(\hat{\eta}_i|x_i)] + Var(E[\hat{\eta}_i|x_i]) \geq E[Var(\hat{\eta}_i|x_i)] \geq c\Pr(\mathcal{A}_i)I.$$

It then follows from equation (8.1) and $\lambda \longrightarrow 0$ that for large enough n,

$$\frac{1}{n}\sum_{i=1}^{n} Var(\hat{\eta}_i) \geq cI\frac{1}{n}\sum_{i=1}\Pr(\mathcal{A}_i) \geq cI.$$

Therefore $\sum_{i=1}^{n} Var(\hat{\eta}_i)/n$ is uniformly nonsingular.

The remainder of the proof is entirely standard. Q.E.D.

9 Appendix C: Solving the Zeros Problem of Demand

In this Appendix we briefly explain how the bounds on average surplus we consider solve the zeros problem of demand. First, allowing a vector of disturbances to enter the demand function in a nonseparable, nonlinear way allows zeros as the outcome of a very general choice specification. For example, an underlying demand model where zeros occur as a censored outcome is allowed because disturbances are not constrained to be additively separable in our specification.

Second, the economics of demand helps explain how zeros are correctly accounted for. An individual who chooses zero over the price range being considered does not care about the price change and so has zero surplus. Average demand includes zero choices and the average and average surplus averages over the same zero surplus individuals. Including zeros in the average demand accounts for individuals who have zero surplus because they do not choose to purchase the good with changing price. This is why including zeros is the correct econometric approach.

To elaborate on the economics suppose only one price p is varying. For simplicity we consider quantity rather than share. Let $q(p, y, \eta)$ be the demand function for the good with price varying (holding all other prices constant) for one type η of individual preferences. Let $s(p, \eta)$ be the equivalent variation EV for a price change from p to p_1 for type η. Hausman and Newey (2016) showed that if the income effect for every is bounded below and above by B_ℓ and B_u respectively then

$$\int_p^{p_1} q(t, y, \eta) \exp(-B_u[t-p])dt \leq s(p, \eta) \leq \int_p^{p_1} q(t, y, \eta) \exp(-B_\ell[t-p])dt.$$

If $q(p, y, \eta)$ is zero over $[p, p_1]$ then upper and lower bounds coincide at zero. Itegrating over the distribution of η gives

$$\int_p^{p_1} \bar{q}(t, y) \exp(-B_u[t-p])dt \leq \bar{s}(p) \leq \int_p^{p_1} \bar{q}(t, y) \exp(-B_\ell[t-p])dt.$$

where $\bar{q}(p, y) = \int q(p, y, \eta)G(d\eta)$ is average demand. Here $\bar{s}(p)$ includes the zero surplus individuals as it must do to be an average surplus over all individuals. Also the average over quantities $\bar{q}(p, y)$ includes the zeros, as it must do to include in the average surplus those who do not purchase any of the good. Thus we get correct bounds for average EV by including the zeros in estimation of average demand.

It also follows from Battacharya (2015) that averaging over zeros leads to the correct calculation in multinomial discrete choice when the price of one good is changing. Let $\bar{q}(p, y)$ denote choice probability for the good with changing price. Bhattacharya (2015) shows average equivalent variation is

$$\bar{s}(p) = \int_p^{p_1} \bar{q}(t, y)dt,$$

As usual the choice probability is the average across individuals of the choice of 0 or 1. Thus the choice probability is an average demand that includes zeros. Average surplus also includes zeros. Thus we get correct average EV by including zeros. We also note that we do not have bounds. With discrete choice the income effects are not important so the integral gives exact EV.

10 References

Allcott, H., B. B. Lockwood, and D. Taubinsky (2019): "Should we tax soda? an overview of theory and evidence," *Journal of Economic Perspectives* 33 (2).

Arellano, M. and S. Bonhomme (2012), "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," *Review of Economic Studies* 79, 987–1020.

Belloni, A., V. Chernozhukov, D. Chen, C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica* 80, 2369-2429.

Belloni, A., V. Chernozhukov, C. Hansen (2013): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies* 81, 608–650,

Bhattacharya, D. (2015): "Nonparametric Welfare Analysis for Discrete Choice," *Econometrica* 83, 617–649.

Blomquist, S. and W.K. Newey (2002): "Nonparametric Estimation with Nonlinear Budget Sets," *Econometrica* 70, 2455-2480.

Blomquist, S., A Kumar, Che-Yuan Liang, and W.K. Newey (2014): "Individual Heterogeneity, Nonlinear Budget Sets, and Taxable Income," CEMMAP working paper 21/14.

Blundell, R., A. Duncan, and K. Pendakur (1998): "Semiparametric Estimation and Consumer Demand", *Journal of Applied Econometrics* 13, 435-462.

Blundell, R. and J.M. Robin (2000): "Latent Separability: Grouping Goods without weak Separability," *Econometrica* 68, 53-84.

Blundell, R. and J.L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in M. Dewatripont, L.P. Hansen and S.J. Turnovsky, eds. *Advances in Economics and Econonometrics: Theory and Applications, Eighth World Congress, Vol. II*, Cambridge: Cambridge University Press.

Blundell, R., D. Kristensen, and R. Matzkin (2014): "Bounding Quantile Demand Functions Using Revealed Preference Inequalities," *Journal of Econometrics* 179, 112–127.

Blundell, R. and R. Matzkin (2014): "Control Functions in Nonseparable Simultaneous Equations Models," *Quantitative Economics* 5, 271–295.

Burda, M., M. Harding, J.A. Hausman (2008): "A Bayesian Mixed Logit Probit model for Multinomial Choice," *Journal of Econometrics* 147, 232-246.

Burda, M., M. Harding, J.A. Hausman (2012): "A Poisson Mixture Model of Discrete Choice," *Journal of Econometrics* 166, 184-203.

Burtless, G. and J.A. Hausman (1978): "The Effect of Taxation on Labor Supply: Evaluating the Gary Negative Income Tax Experiments," *Journal of Political Economy* 86, 1103-1130.

Chamberlain, G. (1982): "Multivariate Regression Models for Panel Data," *Journal of Econometrics* 18, 5-46.

Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics, Volume 2*, eds. Z. Griliches and M. Intriligator,Amsterdam: North-Holland, 1984, 1247–1318.

Chamberlain, G. (1992): "Efficiency Bounds for Semiparametric Regression," *Econometrica* 60, 567-596.

Chernozhukov, V., I. Fernandez-Val, J. Hahn, W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica* 81, 535–580.

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

Chernozhukov, V., I. Fernandez-Val, and W.K. Newey (2017): "Nonseparable Multinomial Choice Models in Cross-Section and Panel Data," *Journal of Econometrics,* forthcoming.

Chernozhukov, V., W.K. Newey, and J. Robins (2018): "Double/De-Biased Machine Learning Using Regularized Riesz Representers," https://arxiv.org/abs/1802.08667.

Chernozhukov, V., W.K. Newey, and R. Singh (2018): "Learning L2-Continuous Regression Functionals via Regularized Riesz Representers," https://arxiv.org/pdf/1809.05224.

Deaton, A. and J. Muellbauer (1980): "Economics of Consumer Behavior," Cambridge: Cambridge University Press.

Dette, H., S. Hoderlein, and N. Neumayer (2016): "Testing Multivariate Economic Restrictions Using Quantiles: The Example of Slutsky Negative Semidefiniteness," *Journal of Econometrics* 191, 129-144.

Dubois, P., R. Griffith, and M. O'Connell (2019): "How well targeted are soda taxes?" working paper.

Gorman, W.M. (1959): "Separable Utility and Aggregation," *Econometrica* 27, 469-481.

Gorman, W.M. (1981): "Some Engel Curves," in *Essays in the Theory and Measurement of Consumer Behaviour in Honor of Sir Richard Stone,* ed. by Angus Deaton. Cambridge: Cambridge University Press.

Graham, B and J.L. Powell (2012): "Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models," *Econometrica* 80, 2105-2152.

Griliches, Z. and J.A. Hausman (1986): "Errors in Variables in Panel Data," *Journal of Econometrics* 31, 93-118.

Harding, M. and M. Lovenheim (2017): "The Effect of Prices on Nutrition: Comparing the Impact of Product-and Nutrient-Specific Taxes," *Journal of Health Economics* 53, 53-71.

Hausman, J.A. (1981): "Exact Consumer Surplus and Deadweight Loss," *American Economic Review* 71, 662-676.

Hausman, J.A., G. Leonard, and J.D. Zona (1994): "Competitive Analysis with Differentiated Products," *Annales d'Économie et de Statistique* 34, 159-180

Hausman, J.A. and W. K. Newey (1995): "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss," *Econometrica* 63, 1445-1476.

Hausman, J.A. and E. Leibtag (2007): "Consumer benefits from increased competition in shopping outlets: Measuring the effect of Wal-Mart," *Journal of Applied Econometrics* 22, 1157-

1177.

Hausman, J.A., S. Morisi, and M. Rainey (2010): "Unilateral Effects of Mergers with General Linear Demand," working paper.

Hausman, J.A. and W.K. Newey (2016): "Individual Heterogeneity and Average Welfare," *Econometrica* 84, 1225-1248.

Hoderlein, S. and A. Lewbel (2012): "Regressor Dimension Reduction with Economic Constraints: The Example of Demand Systems with Many Goods," *Econometric Theory* 28, 1087-1120.

Hoderlein, S. and J. Stoye (2014): "Revealed Preferences in a Heterogeneous Population," *Review of Economics and Statistics* 96, 197-213.

Imbens, G.W. and W.K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica* 77, 1481–1512.

Kitamura, Y. and J. Stoye (2018): "Nonparametric Analysis of Random Utility Models," *Econometrica* 86, 1883-1909.

Lewbel, A. (2001): "Demand Systems With and Without Errors," *American Economic Review* 91, 611-618.

Masten, M. A., and A. Torgovitsky (2016): "Identification of Instrumental Variable Correlated Random Coefficients Models," *Review of Economics and Statistics* 98, 1001–5.

McFadden, D.L. (2005): "Revealed Stochastic Preference: A Synthesis," *Economic Theory* 26, 245-264.

McFadden, D.L. and M. Richter (1991): "Stochastic Rationality and Revealed Stochastic Preference," in J. Chipman, D. McFadden, and M. Richter (eds.) *Preferences, Uncertainty and Optimality: Essays in Honour of Leonid Hurwicz.* Boulder, Co.: Westview Press.

Newey, W.K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62, 1349-1382.

Newey, W.K. and D.L. McFadden: (1994): "Large Sample Estimation and Hypothesis Testing," in R. Engel and D. McFadden, *Handbook of Econometrics* Vol 4., Amsterdam: North-Holland.

Pesaran, H. and R.P. Smith (1995): "Estimating Long-Run Relationships from Dynamic Heterogeneous Panels," *Journal of Econometrics* 68, 79-113.

Press, S.J. (1982): *Applied Multivariate Analysis*, New York: Dover Publications.

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

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

Wooldridge, J.M. (2018): "Correlated Random Effects Models with Unbalanced Panels,"

Journal of Econometrics, forthcoming.