Nonparametric Local Rank Test and Reverse Demand Modeling Strategy

January, 2008

By

Pian Chen

Email: Pian.chen@vuw.ac.nz

Phone: +64 – 27 – 6207728

School of Economics and Finance

Victoria University of Wellington

PO Box 600, Wellington, New Zealand

Abstract: I develop a consistent nonparametric local rank test to study demand systems with many commodities. The test is applied to the China Living Standard Survey data and suggests that rank two demand models are sufficient for the data. However, the estimated nonparametric budget share Engel curves indicate that the popular rank two Almost Ideal Demand System has incorrect specification of price effects. This is the consequence of its misspecified PIGLOG cost function, no matter it is locally or globally flexible. To solve the misspecification problem, I propose a new reverse demand modeling strategy. Rather than deriving budget share equations from a cost function using Shepard's Lemma, I model the response of budget shares to prices nonparametrically and then recover the PIGLOG cost function from the estimated nonparametric budget share Engel curves. The reverse modeling strategy can generate a demand model that is consistent with both demand theory and the data under study.

Keywords: Nonparametric Engel curve, local rank test, *U*-statistics, curse of dimensionality, nearest neighbor inverse regression, basis function visualization, demand model specification, PIGLOG cost function, compensating variation

1. Introduction

In this paper, I consider several important and related econometric problems in the demand literature, including nonparametric local rank test, Engel curve estimation, and demand system model specification. The demand system rank theorem, originated by Gorman (1981) and generalized by Lewbel (1989a, 1991), suggests that testing the local rank of a demand system can provide useful information for specifying parsimonious demand models. But because of the "*curse of dimensionality*", price variables are often discarded when testing the local rank (see Donald 1997) and estimating nonparametric Engel curves (see Banks, Blundell, Lewbel 1997). Omitting the price variables results in a rank estimate that is not in line with Lewbel's definition of local rank as well as incorrect estimates of nonparametric Engel curves, which may be misleading for demand model specification.

To illustrate the problem, I recapitulate Lewbel's definitation of global rank and local rank. Lewbel (1989a) shows that for any demand system budget shares can be expressed as

$$w_i = A(p_i)H(p_i, c_i) \quad (i = 1, \cdots, n), \tag{1}$$

where i represents an individual household, n is the sample size, w, p, and c denote budget shares, log prices, and log total expenditure on G commodities, $A(p)$ is a $G \times L$ random matrix with rank $L \le G$, $H(p,c)$ represents L unknown functions of log prices and log total expenditure.[1] Lewbel (1991) defines the global rank of a demand system to be the supremum rank of the matrix $A(p)$ over all possible price vectors. The matrix $A(p)$ may be degenerate at some price vectors and Lewbel terms the rank of $A(\tilde{p})$ the

[1] Throughout the paper, we use the lower cases p and c for log prices and log total expenditure and use the upper cases P and C for prices and total expenditure.

local rank at a given price vector $\tilde{p}$, denoted by $\tilde{L}$. Given the price vector $\tilde{p}$, $A(\tilde{p})$ is a constant matrix, and we can express budget shares Engel curves as linear combinations of the $\tilde{L}$ functions $H(\tilde{p},c_i)=\left[H_1(\tilde{p},c_i) \quad \cdots \quad H_{\tilde{L}}(\tilde{p},c_i)\right]'$. Similar to the vectors spanning a vector space, $H(\tilde{p},c_i)$ contains the $\tilde{L}$ basis functions spanning the budget share Engel curve function space. The local rank $\tilde{L}$ and the functional forms of $H(\tilde{p},c_i)$ typically depend on the price vector $\tilde{p}$. If a dataset has a single price regime, the local rank is the same as the global rank. So, it is not necessary to include price variables when testing the rank. However, if a dataset has considerable price variations across households, we need to explore the information in prices and estimate the local rank for all price vectors in the sample.

The local rank has important implications for empirical demand modeling. For example, the popular Almost Ideal Demand System (AIDS) has rank two (Deaton and Muellbauer 1980); the Quadratic Almost Ideal Demand System (QUAIDS) has rank three (Banks, Blundell, and Lewbel 1997); a rational rank four model has also been proposed (Lewbel 2003). A model with higher rank requires a larger number of basis functions. Without knowing the true rank, misspecification is likely, and the resulting estimates of price and income elasticities may be inconsistent. Overfitting avoids that problem, but it is inefficient, especially with scarce degrees of freedom. A good starting point for demand analyses, therefore, is to test the local rank of a demand system, because knowledge about its local rank can guide model specification.

To estimate the local rank, I design a second order *U*-statistic to test the number of zero eigenvalues of a covariance matrix when evaluated at $\tilde{p}$. Computing the test

statistic requires estimating both budget share equations and budget share Engel curves nonparametrically. To do so, I employ a dimension reduction technique called Nearest Neighbor Inverse Regression (NNIR) proposed by Hsing (1999). This dimension reduction technique enables us to include the price variables in testing the local rank and estimating the basis functions $H(\tilde{p}, c_i)$.

I apply the nonparametric local rank test to study the China Living Standard Survey (CLSS) data. The dataset contains detailed food consumption information of 786 rural households in the year of 1995. There are considerable price variations across the households because of market segmentation. The local rank test indicates that rank two demand models are sufficient for the data. In addition, basis function visualization suggests that the budget shares are linear functions of log total expenditure, conditional on prices.

The linear budget share Engel curves imply that households' preferences can be represented by the Price Independent Generalized Logarithm (PIGLOG) cost function

(2) $$c(p_i, u_i) = a(p_i) + b(p_i)\, u_i ,$$

where u denotes indirect utility (Muellbauer 1975, 1976). Different specifications of $a(p_i)$ and $b(p_i)$ result in different parametric demand models. For example, Piggott (2003) nests 13 parametric specifications, including the AIDS. The CLSS data, however, rejects both the locally and globally flexible AIDS because it misspecifies the response of budget shares to prices (price effect for short). The misspecification of price effect comes from the misspecified PIGLOG cost function in the first place, particularly $b(p_i)$. The misspecification problem is not unique to the AIDS and other parametric demand models,

but caused by the conventional practice of specifying a cost function *a priori* and then deriving the budget share equations using Shepard's lemma.

As opposed to the conventional approach, I estimate the budget share Engel curves nonparametrically and then retrieve the cost function from the estimated budget share Engel curves. The reverse modeling strategy ensures that the resulting cost function is consistent with both the demand theory and the data under study, which is essential for computing compensating variations in welfare analyses.

The rest of the paper is organized as follows. In Section 2, I develop the nonparametric local rank test and prove its asymptotic properties. Section 3 discusses how to specify a parsimonious parametric model for budget share Engel curves via the local rank test and basis function visualization. In Section 4, I demonstrate why the widely used AIDS is not an appropriate model for the CLSS data and propose the reverse demand modeling strategy. Section 5 concludes the paper, and the appendix contains the proofs of all theorems and detailed estimation procedures.

2. A Nonparametric Local Rank Test

In this section, I present a nonparametric test that can consistently estimate the local rank of a demand system. For notation simplicity, let Y_i denote a G-vector of budget shares for household i, $X_i = (p_i, c_i)$, and $Y_i = F_0(X_i) + U_i$, where U_i denotes a G-vector of error terms satisfying $E(U_i \mid X_i) = \underline{0}$. The budget share equations can be expressed as $F_0(X_i) = A(p_i)H(p_i, c_i)$, and therefore the budget share Engel curves at the given price vector $\tilde{p}$ can be written as $F_0(\tilde{X}_i) = A(\tilde{p})H(\tilde{p}, c_i) \equiv \tilde{A}\tilde{H}(c_i)$, where

$\tilde{X}_i = (\tilde{p}, c_i)$. Because the local rank $\tilde{L}$ corresponds to the rank of matrix $\tilde{A}$, I can estimate $\tilde{L}$ by testing the number of zero eigenvalues of the weighted covariance matrix

$$\Gamma = E\left[p(\tilde{X}_i)^2 F_0(\tilde{X}_i) F_0(\tilde{X}_i)' \right] = \tilde{A} E\left[p(\tilde{X}_i)^2 \tilde{H}(c_i) \tilde{H}(c_i)' \right] \tilde{A}', \tag{3}$$

where $p(\tilde{X}_i)$ represents the density function $p(X_i)$ evaluated at $\tilde{X}_i$. I choose such a Γ matrix for three reasons. First, Γ has the right rank, i.e., the same rank as the matrix $\tilde{A}$. Second, by construction, Γ is a symmetric matrix. A nice property of symmetric matrices is that their eigenvalues are always real numbers, which avoids dealing with complex numbers. Third, the weighting function $p(\tilde{X}_i)^2$ yields an estimate whose asymptotic behavior can be derived in a convenient fashion.

To estimate Γ, I use

$$\tilde{\Gamma} = \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{ik} Y_j Y_k', \tag{4}$$

where $\tilde{\Upsilon}_{ij} = h^{-J} K\left(h^{-1}(X_j - \tilde{X}_i) \right)$, $\tilde{\Upsilon}_{ik} = h^{-J} K\left(h^{-1}(X_k - \tilde{X}_i) \right)$, h is a bandwidth, and J is the number of X variables (i.e., the number of commodities plus one). The estimate in (4) is in fact a consistent plug-in estimate of Γ, i.e., $n^{-1} \sum_{i=1}^{n} \hat{p}(\tilde{X}_i)^2 \hat{F}_0(\tilde{X}_i) \hat{F}_0(\tilde{X}_i)'$, where the first $\hat{p}(\tilde{X}_i) = n^{-1} \sum_{j=1}^{n} \tilde{\Upsilon}_{ij}$, the second $\hat{p}(\tilde{X}_i) = (n-1)^{-1} \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ik}$, $\hat{F}_0(\tilde{X}_i) = \sum_{j=1}^{n} \tilde{\Upsilon}_{ij} Y_j \Big/ \sum_{j=1}^{n} \tilde{\Upsilon}_{ij}$, and $\hat{F}_0(\tilde{X}_i)' = \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ik} Y_k' \Big/ \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ik}$. Note that $k \neq j$ permits a test statistic that does not need to be recentered to have asymptotic mean zero. In addition, because $\tilde{\Gamma}$ is only symmetric between the indices *j* and *k*, it is not a third-order *U*-statistic but a second-order *U*-statistic.

To compute $\tilde{\Gamma}$, we encounter the "*curse of dimensionality*" if the demand system has a large number of commodities. In Appendix A, I show that the common practice of discarding price variable results in erroneous estimates of nonparametric budget Engel curves $F_0(\tilde{X}_i)$ and therefore incorrect test statistic $\tilde{\Gamma}$. To include the price variables, I replace high dimensional X_i with a few NNIR variates X_iB (see Hsing 1999) and assume that $F_0(X_i) = F_1(X_iB)$ (see Chen and Smith 2007). I then calculate $\tilde{\Gamma}$ using $\tilde{\Upsilon}_{ij} = h^{-d} K\left(h^{-1}(X_j - \tilde{X}_i)B\right)$ and $\tilde{\Upsilon}_{ik} = h^{-d} K\left(h^{-1}(X_k - \tilde{X}_i)B\right)$, where $d \le J$ is the number of NNIR variates.

To obtain a test statistic with a null distribution free of nuisance parameters, I standardize $\tilde{\Gamma}$ by $\Sigma = E(U_iU_i' \mid X_iB)$ as in Donald (1997). I then construct the test statistic using the sum of $G - \tilde{L}$ smallest eigenvalues of $\tilde{\Gamma}\Sigma^{-1}$. The test statistic is

$$T(\tilde{L} \mid \tilde{p}) = nh^{d/2} V \sum_{g=1}^{G-\tilde{L}} \lambda_g\left(\tilde{\Gamma}\Sigma^{-1}\right), \tag{5}$$

where λ_g represents the g^{th} smallest eigenvalues, n is the sample size, h is the bandwidth, d is the number of NNIR variates, $V = \left[2(G-\tilde{L}) \| K \|_4^3 \, E\left(p(\tilde{X}_iB)^3\right)\right]^{-1/2}$ is a rescaling factor, where $\| K \|_4^3 = \iiint K(\varphi_1) K(\varphi_2) K(\varphi_1 - \varphi_3) K(\varphi_2 - \varphi_3) d\varphi_1 d\varphi_2 d\varphi_3$. Theorem 1, soon to be presented, shows that the test statistic has the standard normal asymptotic distribution if the true local rank equals $\tilde{L}$.

The test statistic in equation (5) depends on unknown quantities $\tilde{\Gamma}$, Σ, and V. To calculate the test statistic, I use consistent estimates of $\tilde{\Gamma}$, Σ, and V as follows.

$$\hat{\tilde{\Gamma}} = \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \ne j}^{n} \hat{\tilde{\Upsilon}}_{ij} \hat{\tilde{\Upsilon}}_{ik} Y_j Y_k' \tag{6}$$

$$= \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{n} \hat{\tilde{\Upsilon}}_{ij} Y_j \right) \left(\sum_{j=1}^{n} \hat{\tilde{\Upsilon}}_{ij} Y_j \right)' - \left(\sum_{j=1}^{n} \hat{\tilde{\Upsilon}}_{ij}^2 Y_j Y_j' \right) \right],$$

where $\hat{\tilde{\Upsilon}}_{ij} = h^{-d} K\left(h^{-1}(X_j - \tilde{X}_i)\hat{B}\right)$, $\hat{\tilde{\Upsilon}}_{ik} = h^{-d} K\left(h^{-1}(X_k - \tilde{X}_i)\hat{B}\right)$, $\hat{B}$ is the NNIR estimate of B (see Appendix B for the estimation procedure of NNIR), and the second formula is provided for computational convenience.

(7) $$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{F}_1(X_i\hat{B})\right)\left(Y_i - \hat{F}_1(X_i\hat{B})\right)',$$

where $\hat{F}_1(X_i\hat{B}) = \sum_{j=1}^{n} \hat{\Upsilon}_{ij} Y_j \Big/ \sum_{j=1}^{n} \hat{\Upsilon}_{ij}$ and $\hat{\Upsilon}_{ij} = h^{-d} K\left(h^{-1}(X_j - X_i)\hat{B}\right)$.

(8) $$\hat{V} = \left[\frac{2(G-\tilde{L})}{n} \sum_{i=1}^{n} \hat{p}(\tilde{X}_i\hat{B})^3 \right]^{-1/2},$$

where $\hat{p}(\tilde{X}_i\hat{B}) = \frac{1}{n} \sum_{j=1}^{n} \hat{\tilde{\Upsilon}}_{ij}$.

Let $\tilde{L}_0$ denote the true local rank. To characterize the asymptotic distribution of $T(\tilde{L} \mid \tilde{p}) = nh^{d/2} V \sum_{g=1}^{G-\tilde{L}} \lambda_g\left(\tilde{\Gamma}\Sigma^{-1}\right)$ and $\hat{T}(\tilde{L} \mid \tilde{p}) = nh^{d/2} \hat{V} \sum_{g=1}^{G-\tilde{L}} \lambda_g\left(\hat{\tilde{\Gamma}}\hat{\Sigma}^{-1}\right)$ under the null hypothesis $H_0 : \tilde{L}_0 = \tilde{L}$, I make the following assumptions.

Assumptions:

A1: $(X_i', Y_i')' \in R^{J+G}$ are *i.i.d.* random vectors for $i = 1, \cdots, n$. $Y_i = F_0(X_i) + U_i$, where $Y_i = w_i$, $X_i = (p_i, c_i)$, and $E(U_i \mid X_i) = \underline{0}$ with probability one.

A2: $F_0(X_i) = A(p_i)H(p_i, c_i)$ with probability one. For a given price vector $\tilde{p}$, $\tilde{X}_i = (\tilde{p}, c_i)$, $F_0(\tilde{X}_i) \in \Im(\tilde{L}_0)$, where $\Im(\tilde{L}_0)$ defines a set of functions $\left\{F : R^J \to R^G\right\}$

that satisfies $F_0(\tilde{X}_i) = \tilde{A}\tilde{H}(c_i)$, where $rank(\tilde{A}) = \tilde{L}_0 \le G$ and the elements of $\tilde{H}(c_i)$ are functionally uncorrelated.

A3: $F_0(X_i) = F_1(X_iB)$ with probability one, where $\dim(X_iB) = d \le J$. Each element of F_1 has an extension to domain R^d that has $s > 0$ continuous bounded derivatives.

A4: The support of X_iB is the Cartesian product of compact intervals $[a_v, b_v]$ for $v = 1, \cdots, d$. X_iB is continuously distributed with the density function $p(X_iB)$ that has $s \ge k$ continuous bounded derivatives (k is the order of kernel in A6) and is bounded and bounded below by a positive constant.

A5: $E(U_iU_i' \mid X_iB) = \Sigma$ for all i with probability one and Σ is finite and positive definite.

A6: $K(\varphi)$ is a symmetric bounded kernel of order k such that

(i) $K(\varphi)$ has compact support in R^d;

(ii) $\int K(\varphi)d\varphi = 1$;

(iii) $K(\varphi)$ is differentiable of order $q \ge 0$ and the q^{th} order derivative is *Lipschitz*;

(iv) $\int \varphi_1^{\theta_1} \cdots \varphi_d^{\theta_d} K(\varphi)d\varphi = \pi(|\theta|)$, where $|\theta| = \sum_{l=1}^{d} \theta_l$ with each θ_l being a nonnegative integer, and $\pi(|\theta|) = 0$ when $|\theta| < k$.

A7: $nh^d \to \infty$ and $h \to 0$ as $n \to \infty$.

A8: $E(Xb \mid X\beta_1, X\beta_2, \cdots, X\beta_d)$ is linear in $X\beta_1, X\beta_2, \cdots, X\beta_d$ for any vector $b \in R^J$.

Assumption A1 describes the data generating process for (X_i, Y_i). Assumption A2 defines the relevant set $\Im$ in which the local rank exists with probability one. Assumption

A3 states that the first moment of Y depends only on $d \le J$ linear combinations of X, which validates nonparametric regressions with J-dimensional X replaced by d-dimensional NNIR variates XB. In addition, it imposes a smooth condition on the nonparametric regression function F_1. Assumption A4 imposes a smooth condition on the density function $p(X_i B)$. It requires that $p(X_i B)$ is at least as smooth as the order of kernel to enable precise analysis of the bias terms. Assumption A5 implies that the covariance matrix of the error terms is constant across observations. Assumptions A6-A7 are useful in controlling the convergence rate of $\hat{\tilde{\Gamma}} \xrightarrow{p} \Gamma$ and $\hat{\Sigma} \xrightarrow{p} \Sigma$. Assumption A8 is the linear design condition, which is crucial for applying inverse regression to estimate B (see Li 1991, 2000).

Under assumptions A1-A7, I establish the asymptotic normality of $T(\tilde{L} \mid \tilde{p})$.

Theorem 1: $T(\tilde{L} \mid \tilde{p}) = nh^{d/2} V \sum_{g=1}^{G-\tilde{L}} \lambda_g \left(\tilde{\Gamma} \Sigma^{-1} \right) \xrightarrow{d} N(0,1)$ under $H_0 : \tilde{L}_0 = \tilde{L}$.

Proof: See Appendix C

Under assumptions A1-A8, I establish the asymptotic normality of $\hat{T}(\tilde{L} \mid \tilde{p})$.

Corollary 1: $\hat{T}(\tilde{L} \mid \tilde{p}) = nh^{d/2} \hat{V} \sum_{g=1}^{G-\tilde{L}} \lambda_g \left(\hat{\tilde{\Gamma}} \hat{\Sigma}^{-1} \right) \xrightarrow{d} N(0,1)$ under $H_0 : \tilde{L}_0 = \tilde{L}$.

Corollary 1 follows immediately from Theorem 1, the Slutsky Theorem, and Lemma 4.7 in White (2001), given the consistent estimates of $\tilde{\Gamma}$, Σ, and V and the fact that eigenvalues are continuous functions of the elements in a matrix. Assumptions A3

and A8 ensure that the estimate $\hat{B}$ generated by inverse regression is root-*n* consistent for *B*. In addition, using *B* or $\hat{B}$ does not affect the asymptotic properties of nonparametric estimates because $\hat{B}$ has a faster convergence rate than nonparametric estimates (see Chen and Smith 2007 for proof). Given assumptions A4 and A6-A8, $\hat{p}(\tilde{X}_i\hat{B})$ is consistent for $p(\tilde{X}_iB)$, which implies that $\hat{V}$ is consistent for *V* using the Slutsky Theorem and the Law of Large Numbers. Under assumptions A1 and A3-A8, $\hat{\Sigma} \xrightarrow{p} \Sigma$ can be shown in a similar fashion to the proof of Lemma 2 in Donald (1997). Given these consistent estimates, $\hat{T}(\tilde{L} \mid \tilde{p})$ and $T(\tilde{L} \mid \tilde{p})$ are asymptotically equivalent and therefore have the same limiting distribution using Lemma 4.7 in White (2001).

Based on the asymptotic distribution of $\hat{T}(\tilde{L} \mid \tilde{p})$, $\tilde{L}_0$ can be consistently estimated via sequential testing of the hypotheses $H_0: \tilde{L}_0 = 0$ vs $H_1: \tilde{L}_0 > 0$, $H_0: \tilde{L}_0 \leq 1$ vs $H_1: \tilde{L}_0 > 1$, $H_0: \tilde{L}_0 \leq 2$ vs $H_1: \tilde{L}_0 > 2$, and so on until the null is not rejected. The sequential test is one-sided in nature, therefore it is undersized but consistent as discussed in Theorem 2 of Donald (1997). Its consistency prevents us from underestimating the true local rank.

3. Basis Function Visualization and Engel Curve Specification

In this section, I discuss how to specify a parsimonious parametric model for budget share Engel curves via the local rank test and basis function visualization. The estimated local rank indicates that how many basis functions are needed to span the budget share Engel curve function space. If we can visualize the basis functions $H(\tilde{x}B)$, it will be much easier for us to specify a parametric model for the budget share Engel curves. So, the first task after the local rank test is to estimate the basis functions.

But an important fact is that the basis functions are not identifiable because

(9) $$F_1(\tilde{x}B) = \tilde{A}H(\tilde{x}B) = \tilde{A}\Pi^{-1}\cdot\Pi H(\tilde{x}B) = \tilde{A}^*H^*(\tilde{x}B)$$

holds for any $\tilde{L}\times\tilde{L}$ full rank matrix Π. Although $H(\tilde{x}B)$ and $H^*(\tilde{x}B) = \Pi H(\tilde{x}B)$ may appear to be different functions, they in fact span the same function space because $H^*(\tilde{x}B)$ is just a linear transformation of $H(\tilde{x}B)$. Consequently, we can identify the function space using $H(\tilde{x}B)$ or $H^*(\tilde{x}B)$. This fact implies that we can calculate the basis functions using

(10) $$H(\tilde{x}B) = \Omega F_1(\tilde{x}B),$$

where Ω denotes any $\tilde{L}\times G$ matrix with rank $\tilde{L}$. Some choices of Ω may be better than others for visualization purpose. To capture interesting features of high-dimensional data, I recommend experimenting with several different Ω matrices and comparing the resulting basis functions.

In the case of budget share Engel curves, we have some prior information on the basis functions. The adding-up constraint from demand theory suggests a constant basis function. We expect one of the estimated basis functions to reflect this prior information. But if we simply pre-multiply the estimated $F_1(\tilde{x}B)$ by an arbitrary Ω matrix, we would have little chance to obtain such a constant basis function. To obtain a constant basis function, we need to impose some restrictions on Ω.

Suppose that we have prior information on $m < \tilde{L}$ basis functions. I partition Ω and $H(\tilde{x}B)$ as follows:

(11) $$\begin{bmatrix} H_{(m)}(\tilde{x}B) \\ H_{(\tilde{L}-m)}(\tilde{x}B) \end{bmatrix} = \begin{bmatrix} \Omega_{m\times G} \\ \Omega_{(\tilde{L}-m)\times G} \end{bmatrix} F_1(\tilde{x}B),$$

where $H_{(m)}(\tilde{x}B) = \left[H_1(\tilde{x}B), \cdots, H_m(\tilde{x}B)\right]'$ denotes the basis functions known from the prior information and $\Omega_{(\tilde{L}-m)\times G}$ is a user-specified $(\tilde{L}-m)\times G$ constant matrix with rank $\tilde{L}-m$. Given $\hat{F}_1(\tilde{x}B)$, $H_{(m)}(\tilde{x}B)$, and $\Omega_{(\tilde{L}-m)\times G}$, I can solve for the submatrix $\Omega_{m\times G}$ and the unknown basis functions $H_{(\tilde{L}-m)}(\tilde{x}B) = \left[H_{m+1}(\tilde{x}B), \cdots, H_{\tilde{L}}(\tilde{x}B)\right]'$ using

$$\hat{\Omega}_{m\times G} = H_{(m)}(\tilde{x}B)\hat{F}_1(\tilde{x}B)'\left(\hat{F}_1(\tilde{x}B)\hat{F}_1(\tilde{x}B)'\right)^{-1} \tag{12}$$

and

$$\hat{H}_{(\tilde{L}-m)}(\tilde{x}B) = \Omega_{(\tilde{L}-m)\times G}\hat{F}_1(\tilde{x}B). \tag{13}$$

$\hat{\Omega}_{m\times G}$ contains the restrictions on Ω, using which we can obtain the m basis functions $\hat{H}_{(m)}(\tilde{x}B) = \hat{\Omega}_{m\times G}\hat{F}_1(\tilde{x}B)$ that are closest to those suggested by the prior information; $\hat{H}_{(\tilde{L}-m)}(\tilde{x}B)$ contains the remaining $\tilde{L}-m$ basis functions if $\begin{bmatrix} \hat{\Omega}_{m\times G} \\ \Omega_{(\tilde{L}-m)\times G} \end{bmatrix}$ has rank $\tilde{L}$.

To visualize the basis functions, I can plot the estimated $H(\tilde{x}B)$ against the estimated NNIR variates $\tilde{x}B$. In the example of budget share Engel curves, the NNIR variates $\tilde{x}B$ only vary with log total expenditure, holding the prices constant. So, I only need to plot the basis functions against log total expenditure.

Next, I demonstrate the local rank test and basis function visualization using a dataset of food consumption in north China. The data is from the Living Standards Measurement Study (LSMS): 1995-1997 China Living Standards Survey (CLSS) conducted by the World Bank.[2] It contains information on purchased, self-produced, and traded quantities of 25 food items consumed by 786 households in Hebei and Liaoning

[2] http://www.worldbank.org/LSMS/guide/select.html

Provinces in the year of 1995. The expenditure information is for the purchased quantities only. To calculate the prices faced by each household, I divide its expenditures by the corresponding purchased quantities. I then use the derived prices to recalculate the expenditures on the purchased, self-produced, and traded quantities combined. Using the recalculated expenditures, I compute the total expenditure and budget shares for each household.

To focus on main food consumption, I select ten food items with the largest budget shares to form a demand system, including (1) Vegetables, (2) Milled rice, (3) Flour, (4) Pork, (5) Vegetable oil, (6) Eggs, (7) Corn, corn flour, (8) Lard, (9) Potatoes, and (10) Fruit. The expenditures on the ten food items account for 87.53% of the total expenditure on the 25 food items. Table 1 provides summary statistics on budget shares and prices of the ten-good demand system. An interesting feature of this dataset is that it has considerable price variations across households, as indicated by the large standard deviations of the price variables. The results presented below indicate that the price variables play a significant role in estimating nonparametric Engel curves and determining the local rank of a demand system.

Using three NNIR variates and Epanechnikov kernel[3] with a bandwidth $h = 1.6$ suggested by cross validation (see Appendix A), I perform the local rank test conditional on each of the 786 price realizations in the sample. To satisfy assumption A5, I drop the fruit equation to ensure the covariance matrix Σ is positive definite. The estimated rank is two for 480 price vectors, and it is one for 306 price vectors. To summarize the testing results, I list only six price vectors in Table 2. In Table 3, I report the *p*-values of the local

[3] For Epanechnikov kernel, $\| K \|_4^3 = 0.345$.

rank tests conditional on the six price vectors. At the 5% confidence level, the first three price vectors result in rank two and the last three price vectors result in rank one. The first three price vectors are chosen because their *p*-values for rank $\tilde{L}_0 = 1$ correspond to the 25%, 50%, and 75% quantiles of the *p*-values that indicate rank two; the last three price vectors are chosen because their *p*-values for rank $\tilde{L}_0 = 1$ correspond to the 25%, 50%, and 75% quantiles of the *p*-values that indicate rank one. Among the six price vectors, the local rank at the price vector 6 is closest to rank zero, while the local rank at the price vector 1 is farthest from rank one. These local rank test results indicate that the local rank is up to two for all 786 price realizations in the sample.

In Figure 1, I plot the estimated budget share Engel curves evaluated at the six price vectors. Specifically, I present the nonparametric Engel curves for the first three price vectors in panel (a) and those for the last three price vectors in panel (b). In the two panels, I use solid, dashed, and dotted curves for the price vectors 1-3 and 4-6, respectively. It is evident that the nonparametric Engel curves of the food items 1-5 are quite different across the price vectors. For example, in panel (a), as the total expenditure increases, the Engel curve of flour goes up when evaluated at the price vector 3 (the dotted curve) but it goes down when evaluated at the price vectors 1 and 2 (the solid and dashed curves). The difference between the Engel curves evaluated at different price vectors can be better understood using the estimated basis functions.

In Figure 2, I plot the estimated basis functions evaluated at the six price vectors. Panel (a) presents the two basis functions for the price vectors 1-3: one is approximately constant as suggested by the adding-up constraint, the other appears to be a linear function. However, the linear functions have different slope across the three price vectors.

Clearly, the slope should be a function of prices, which accounts for the different Engel curves in Figure 1. Panel (b) presents the one basis function for the price vectors 4-6. Not surprisingly, this basis function is close to be constant.

The local rank test and basis function visualization together can facilitate Engel curve specification. Lewbel (1991) shows that a demand system has rank two if and only if the budget share Engel curves have the Generalized Linear form (GL, see Muellbauer 1975), $w_g(\tilde{p},c_i) = A_{1g}(\tilde{p}) + A_{2g}(\tilde{p})\xi(\tilde{p},c_i)$ $(g = 1,\cdots,G)$, where $\xi(\tilde{p},c_i)$ is an unknown function of log prices $\tilde{p}$ and log total expenditure c_i, and both A_{1g} and A_{2g} are functions of $\tilde{p}$ satisfying $\sum_{g=1}^{G} A_{1g} = 1$ and $\sum_{g=1}^{G} A_{2g} = 0$. Knowledge of the local rank does restrict one's attention to certain types of demand models, but it provides no information on the functional form of $\xi(\tilde{p},c_i)$. To select an appropriate functional form for $\xi(\tilde{p},c_i)$, I turn to basis function visualization. The estimated basis functions in Figure 2 suggest that $\xi(\tilde{p},c_i)$ may be linear in c_i conditional on $\tilde{p}$. Therefore, the budget share Engel curves may be well approximated by a linear function of log total expenditure plus a constant term, i.e.,

$$E(w_{ig} \,|\, p_i = \tilde{p}, c_i) = F_{0g}(\tilde{p},c_i) \approx \alpha_{1g}(\tilde{p}) + \alpha_{2g}(\tilde{p})\, c_i \qquad (g = 1,\cdots,G). \tag{14}$$

Note that, for a given price vector $\tilde{p}$, both the intercept $\alpha_{1g}(\tilde{p})$ and the slope $\alpha_{2g}(\tilde{p})$ are constant parameters that can be estimated using linear regression of $F_{0g}(\tilde{p},c_i)$ on $[1 \quad c_i]$.

The linear Engel curves imply that households' preferences can be represented by the Price Independent Generalized Logarithm (PIGLOG) cost function (Muellbauer 1975, 1976). However, the specifications of price effects in the PIGLOG cost function differ

across models. In the next section, I show the CLSS data rejects the specification of AIDS and propose a new reverse modeling strategy.

4. A Reverse Demand Modeling Strategy

In this section, I show that the specification of price effect in the AIDS is not consistent with the CLSS data. This discovery motivates a new modeling strategy that is opposite to the traditional modeling strategy by which the AIDS (Deaton and Muellbauer 1980) and the TRANSLOG (Christensen, Jorgenson, and Lau 1975) are developed. The AIDS specifies a cost function *a priori* and then derives budget share equations using Shepard's lemma; the TRANSLOG specifies a utility function *a priori* and then derives budget share equations using Roy's Identity. This traditional modeling strategy intends to approximate unobservable functions (i.e., the cost function in the AIDS and the utility function in the TRANSLOG), from which it derives observable budget share equations. [4] An obvious drawback of the traditional modeling strategy is that if the unobservable functions are misspecified in the first place then the derived observable functions must also be misspecified. My reverse modeling strategy is to recover an unobservable function from observable functions, which avoids the misspecification problem. Specifically, I first specify a semi-parametric model for budget share Engel curves and then retrieve the PIGLOG cost function by solving two differential equations.

4.1 The Specification Problem of the AIDS

As a first step, note that the PIGLOG cost function

[4] By "unobservable", I mean the functions contain latent variables on either right or left hand side, and therefore cannot be estimated.

(15) $$c(p_i,u_i) = a(p_i) + b(p_i)\,u_i\,,$$

always gives rise to the budget share functions that satisfy $F_0(p_i,c_i) = A(p_i)H(c_i)$, where c_i, p_i, and u_i denote log total expenditure, log prices, and indirect utility, respectively, both $a(p_i)$ and $b(p_i)$ are functions of log prices, $A(p_i)$ is a $G\times 2$ random matrix, and $H(c_i) = \begin{bmatrix} 1 & c_i \end{bmatrix}'$. This result can be verified using Shepard's lemma and substituting $u_i = \left[c_i - a(p_i)\right]/b(p_i)$, i.e.,

(16) $$\begin{aligned} F_0(p_i,c_i) &= \partial c(p_i,u_i)/\partial p_i \\ &= d\,a(p_i)/dp_i + \left[d\,b(p_i)/dp_i\right]u_i \\ &= \begin{bmatrix} A_1(p_i) & A_2(p_i) \end{bmatrix}\begin{bmatrix} 1 \\ c_i \end{bmatrix} = A(p_i)H(c_i), \end{aligned}$$

where $A(p_i) = \begin{bmatrix} A_1(p_i) & A_2(p_i) \end{bmatrix}$ with

(17) $$A_1(p_i) = \left[d\,a(p_i)/dp_i\right] - a(p_i)A_2(p_i)$$

and

(18) $$A_2(p_i) = d\ln b(p_i)/dp_i\,.$$

Both $A_1(p_i)$ and $A_2(p_i)$ represent a column vector of G unknown functions of log prices, i.e., $A_1(p_i) = \begin{bmatrix} \alpha_{1g=1}(p_i) & \cdots & \alpha_{1g=G}(p_i) \end{bmatrix}'$ and $A_2(p_i) = \begin{bmatrix} \alpha_{2g=1}(p_i) & \cdots & \alpha_{2g=G}(p_i) \end{bmatrix}'$. The g^{th} elements of $A_1(p_i)$ and $A_2(p_i)$ are $\alpha_{1g}(p_i) = \partial \ln a(p_i)/\partial p_{ig} - a(p_i)\alpha_{2g}(p_i)$ and $\alpha_{2g}(p_i) = \partial \ln b(p_i)/\partial p_{ig}$, respectively, where p_{ig} represents the log price of the g^{th} commodity for household i.

Given a price vector $\tilde{p}$, $\alpha_{1g}(p_i) = \alpha_{1g}(\tilde{p})$ and $\alpha_{2g}(p_i) = \alpha_{2g}(\tilde{p})$, the model in (16) becomes the same as the model in (14). Because the model in (14) is consistent with the CLSS data, the model in (16) with appropriately specified $a(p_i)$ and $b(p_i)$ in the

PIGLOG cost function is also consistent with the CLSS data. Demand theory, however, provides little guidance on the functional forms for $a(p_i)$ and $b(p_i)$.

In the demand literature, flexible functional forms have been proposed. For example,

$$a(p_i) = \alpha_0 + \sum_{g=1}^{G} \alpha_g \ln P_{ig} + \frac{1}{2} \sum_{g=1}^{G} \sum_{g'=1}^{G} \gamma_{gg'} \ln P_{ig} \ln P_{ig'} \tag{19}$$

$$b(p_i) = \theta_0 \prod_{g=1}^{G} P_{ig}^{\theta_g} \ , \tag{20}$$

where α, γ, and θ are parameters and P_{ig} represent the price of the g^{th} commodity for household *i* (i.e., $p_{ig} = \ln P_{ig}$). The specifications of (19)-(20) result in the popular AIDS

$$F_{0g}(p_i, c_i) = \alpha_g + \sum_{g'=1}^{G} \gamma_{gg'} \ln P_{ig'} + \theta_g \ln(C_i / P_i^*) , \tag{21}$$

where $\ln P_i^* = \alpha_0 + \sum_{g=1}^{G} \alpha_g \ln P_{ig} + \frac{1}{2} \sum_{g=1}^{G} \sum_{g'=1}^{G} \gamma_{gg'} \ln P_{ig} \ln P_{ig'}$ and C_i represents total expenditure (i.e., $c_i = \ln C_i$).

One justification for the specification of $a(p_i)$ in (19) is based on Taylor series approximation. Deaton and Muellbauer (1980) argue that the resulting cost function possesses enough parameters so that at any single point its derivatives $\partial c_i / \partial p_{ig}$, $\partial c_i / \partial u_i$, $\partial^2 c_i / \partial p_{ig} \partial p_{ig'}$, $\partial^2 c_i / \partial p_{ig} \partial u_i$, and $\partial^2 c_i / \partial u_i^2$ can be set to those of any arbitrary cost function. But Taylor series can only provide good local approximation at some point in the price and utility space rather than the entire support of a cost function. In general, Taylor series approximation has no relation to least squares approximation to an unknown function (White 1980). Therefore, estimated parameters and hypothesis tests based on Taylor series approximation can be misleading (Chalfant and Gallant 1985).

To overcome this problem, Chalfant (1987) proposed the globally flexible AIDS using Fourier series to approximate $a(p_i)$, i.e.,

$$
\begin{aligned}
a(p_i) = \alpha_0 + \delta' \ln P_i + \frac{1}{2} \ln P_i' \Theta \ln P_i \\
+ \sum_{m=1}^{M} \left\{ u_{0m} + 2 \sum_{t=1}^{T} t \left[u_{tm} \cos(t \lambda k_m' \ln P_i) + v_{tm} \sin(t \lambda k_m' \ln P_i) \right] \right\},
\end{aligned} \tag{22}
$$

where k_m denotes a multi-index (a vector of integer numbers), λ is a scaling factor, and M and T determine the number of sine and cosine terms. The parameters are α_0, δ, u_{0m}, u_{tm}, and v_{tm}; Θ is defined by $-\sum_{m=1}^{M} \lambda^2 u_{0m} k_m k_m'$. By increasing M and T, Fourier series can approximate an unknown function arbitrarily close over its entire support. So, Fourier series approximation is global approximation and comparable to least square approximation. Using the specifications of $a(p_i)$ in (22) and $b(p_i)$ in (20), the globally flexible AIDS specifies the budget share equations as

$$
\begin{aligned}
F_{0g}(p_i, c_i) = \delta_g + \theta_g \ln(C_i / P_i^*) \\
- \lambda \sum_{m=1}^{M} \left\{ u_{0mg} \lambda k_m' \ln P_i + 2 \sum_{t=1}^{T} t \left[u_{tmg} \sin(t \lambda k_m' \ln P_i) + v_{tmg} \cos(t \lambda k_m' \ln P_i) \right] \right\} k_m .
\end{aligned} \tag{23}
$$

The globally flexible AIDS can provide better approximation than the locally flexible AIDS in (21). However, both the locally and globally flexible AIDS share the same specification of $b(p_i)$ in (20), which is not consistent with the CLSS data.

To demonstrate the specification problem of $b(p_i)$, I rewrite equation (20) as $\ln b(p_i) = \ln \theta_0 + \sum_{g=1}^{G} \theta_g p_{ig}$, which by equation (18) implies that $A_2(p_i)$ in model (16) is a vector of constants with its g^{th} elements $\alpha_{2g}(p_i) = \partial \ln b(p_i) / \partial p_{ig} = \theta_g$ for all p_i. This implication, however, contradicts the nonparametric estimates of budget share Engel

curves $\hat{F}_0(\tilde{p},c_i)$. As shown in Figure 1, all the budget share Engel curves are approximately linear, but their slopes are visually different across the six price vectors for the first five food items, which suggests that $\alpha_{2g}(p_i)$ should be a function of log prices rather than a constant. In Table 4, I report the estimated intercept and slope parameters, $\hat{\alpha}_{1g}(\tilde{p})$ and $\hat{\alpha}_{2g}(\tilde{p})$, from regressing $\hat{F}_{0g}(\tilde{p},c_i)$ on $[1 \quad c_i]$ for the six price vectors. The different values of $\hat{\alpha}_{2g}(\tilde{p})$ across $\tilde{p}$ further confirm the visual impression that $\alpha_{2g}(p_i)$ should not be a constant θ_g . Consequently, the AIDS is not an ideal model for the CLSS data although it has the right rank (i.e., rank two) and basis functions (i.e., $[1 \quad c_i]$) for the budget share Engel curve function space.

4.2 Recovering the PIGLOG Cost Function Nonparametrically

To develop an appropriate model for the data, I propose a new modeling strategy. Instead of specifying $a(p_i)$ and $b(p_i)$ *a priori*, I recover them directly from $A_1(p_i)=\left[\alpha_{1g=1}(p_i) \quad \cdots \quad \alpha_{1g=G}(p_i)\right]'$ and $A_2(p_i)=\left[\alpha_{2g=1}(p_i) \quad \cdots \quad \alpha_{2g=G}(p_i)\right]'$ defined in equations (16)-(18). Assuming $A_1(p_i)$ and $A_2(p_i)$ are continuous functions of p_i, I derive

$$b(p_i)=\exp\left(\sum_{g=1}^{G}\int \alpha_{2g}(p_i)dp_{ig}\right) \tag{24}$$

$$a(p_i)=b(p_i)\left[\sum_{g=1}^{G}\int \alpha_{1g}(p_i)/b(p_i)\,dp_{ig}+k\right], \tag{25}$$

where k represents a constant, by solving two differential equations. The specifications in (15) and (24)-(25) constitute a PIGLOG cost function that is consistent with both demand theory and the CLSS data.

First, I prove equation (24). In equation (18), $A_2(p_i) = d\ln b(p_i)/dp_i$, whose g^{th} element is $\alpha_{2g}(p_i) = \partial \ln b(p_i)/\partial p_{ig}$. The total differential of $\ln b(p_i)$ is $d\ln b(p_i) = \sum_{g=1}^{G} \alpha_{2g}(p_i)\, dp_{ig}$. Integrating both sides yields $\int d\ln b(p_i) = \sum_{g=1}^{G} \int \alpha_{2g}(p_i)\, dp_{ig}$, which in turn implies $\ln b(p_i) = \sum_{g=1}^{G} \int \alpha_{2g}(p_i)\, dp_{ig}$. Exponentiating both sides results in $b(p_i) = \exp\left(\sum_{g=1}^{G} \int \alpha_{2g}(p_i) dp_{ig} \right)$.

Second, I prove equation (25). In equation (17), $A_1(p_i) = \left[d\, a(p_i)/dp_i \right] - a(p_i) A_2(p_i)$, whose g^{th} element is $\alpha_{1g}(p_i) = \partial a(p_i)/\partial p_{ig} - a(p_i)\alpha_{2g}(p_i)$. The total differential of $a(p_i)$ is

$$d\, a(p_i) = \sum_{g=1}^{G} \frac{\partial a(p_i)}{\partial p_{ig}} dp_{ig} = \sum_{g=1}^{G} \left[\alpha_{1g}(p_i) + a(p_i)\alpha_{2g}(p_i) \right] dp_{ig}.$$

Let $\mu(p_i)$ be a scalar integrating factor and multiply both sides of the total differential equation by $\mu(p_i)$, I obtain

$$\mu(p_i) d\, a(p_i) = \mu(p_i) \sum_{g=1}^{G} \left[\alpha_{1g}(p_i) + a(p_i)\alpha_{2g}(p_i) \right] dp_{ig}. \tag{26}$$

I assume

$$\mu(p_i)\alpha_{2g}(p_i) = -\frac{\partial \mu(p_i)}{\partial p_{ig}} \tag{27}$$

and substitute equation (27) into equation (28). This substitution results in

$$\begin{aligned}
&\mu(p_i) d\, a(p_i) + a(p_i) \sum_{g=1}^{G} \frac{\partial \mu(p_i)}{\partial p_{ig}} dp_{ig} = \sum_{g=1}^{G} \mu(p_i)\alpha_{1g}(p_i) dp_{ig} \\
&\Rightarrow d\mu(p_i)\, a(p_i) = \sum_{g=1}^{G} \mu(p_i)\alpha_{1g}(p_i) dp_{ig} \Rightarrow \int d\mu(p_i)\, a(p_i) = \sum_{g=1}^{G} \int \mu(p_i)\alpha_{1g}(p_i) dp_{ig} \\
&\Rightarrow \mu(p_i)\, a(p_i) - k_1 = \sum_{g=1}^{G} \int \mu(p_i)\alpha_{1g}(p_i) dp_{ig} \\
&\Rightarrow a(p_i) = \left[\sum_{g=1}^{G} \int \mu(p_i)\alpha_{1g}(p_i) dp_{ig} + k_1 \right] \Big/ \mu(p_i)
\end{aligned} \tag{28}$$

for some constant k_1. Note that equation (27) implies $\partial \ln \mu(p_i)/\partial p_{ig} = -\alpha_{2g}(p_i)$. The total differential of $\ln \mu(p_i)$ is $d\ln \mu(p_i) = -\sum_{g=1}^{G} \alpha_{2g}(p_i) dp_{ig}$. Integrating both sides yields $\ln \mu(p_i) = -\sum_{g=1}^{G} \int \alpha_{2g}(p_i) dp_{ig} + k_2$, which in turn implies that

$$\mu(p_i) = \exp\left[-\sum_{g=1}^{G} \int \alpha_{2g}(p_i) dp_{ig} + k_2 \right] = k_3 \exp\left[-\sum_{g=1}^{G} \int \alpha_{2g}(p_i) dp_{ig} \right] = \frac{k_3}{b(p_i)} \tag{29}$$

for some constants k_2 and $k_3 = exp(k_2)$. Plug equation (29) into equation (28), I obtain $a(p_i) = b(p_i)\left[\sum_{g=1}^{G} \int \alpha_{1g}(p_i)/b(p_i)\, dp_{ig} + k \right]$ for some constant $k = k_1/k_3$.

This new modeling strategy reverses the modeling strategy of the AIDS. The AIDS first specifies $a(p_i)$ and $b(p_i)$ in the PIGLOG cost function and then derives $A_1(p_i)$ and $A_2(p_i)$ in the budget share equations. The new modeling strategy first estimates $A_1(p_i)$ and $A_2(p_i)$ nonparametrically using the semi-parametric model in (14) and then retrieves the unknown functions $a(p_i)$ and $b(p_i)$ in the PIGLOG cost function. Because the model in (14) is consistent with the CLSS data, the recovered PIGLOG cost function is also consistent with the data.

A correctly specified PIGLOG cost function is essential for welfare analyses, especially for calculating compensating variations. Note that the constant *k* in (25) does not affect calculating compensating variations. For household *i*, the compensating variation is $C_i^* - C_i = \exp(c_i^*) - \exp(c_i)$, which represents the additional cost to maintain the same indirect utility u_i when the log prices change from p_i to p_i^*. The indirect utility can be obtained by inverting the PIGLOG cost function in (15), i.e., $u_i = [c_i - a(p_i)]/b(p_i)$,

where c_i represents the log total expenditure at the original log prices p_i. To achieve the indirect utility u_i, the log total expenditure at the new log price p_i^* is

$$\begin{aligned} c_i^* &= a(p_i^*) + b(p_i^*)\, u_i \\ &= b(p_i^*)\left[\sum_{g=1}^{G} \int \alpha_{1g}(p_i^*) \big/ b(p_i^*)\, dp_{ig}^* + k\right] + b(p_i^*)\left[\frac{c_i}{b(p_i)} - \sum_{g=1}^{G} \int \alpha_{1g}(p_i) \big/ b(p_i)\, dp_{ig} - k\right] \\ &= b(p_i^*)\left[\sum_{g=1}^{G} \int \alpha_{1g}(p_i^*) \big/ b(p_i^*)\, dp_{ig}^* - \sum_{g=1}^{G} \int \alpha_{1g}(p_i) \big/ b(p_i)\, dp_{ig} + \frac{c_i}{b(p_i)}\right], \end{aligned}$$

in which k is cancelled out. So, we can simply set k=0 when calculating compensating variations.

Next, I provide some practical suggestions on simulating $a(p_i)$ and $b(p_i)$ defined in (24) and (25) nonparametrically. For the g^{th} commodity, let $p_{g(1)} < \cdots < p_{g(s)} < p_{g(s+1)} < \cdots < p_{g(S)}$ be an additive sequence of prices starting from $p_{g(1)}$, ending by $p_{g(S)}$, with an increment $p_{g(s+1)} - p_{g(s)}$ equal to $\Delta_{gS} = \left[p_{g(S)} - p_{g(1)}\right] \big/ (S-1)$. The first element $p_{g(1)}$ represents the lower integration limit, the last element $p_{g(S)}$ represents the upper integration limit, and $(S-1)$ is the order of the integration. Using these notations, I can approximate $a(p_i)$ and $b(p_i)$ using

$$\hat{b}(p_i) = \exp\left(\sum_{g=1}^{G} \Delta_{gS} \sum_{s=1}^{S-1} \hat{\alpha}_{2g}(p_{i1}, \cdots, p_{g(s)}, \cdots, p_{iG})\right) \tag{30}$$

and

$$\hat{a}(p_i) = \hat{b}(p_i)\left[\sum_{g=1}^{G} \Delta_{gS} \sum_{s=1}^{S-1} \frac{\hat{\alpha}_{1g}(p_{i1}, \cdots, p_{g(s)}, \cdots, p_{iG})}{\hat{b}(p_{i1}, \cdots, p_{g(s)}, \cdots, p_{iG})} + k\right], \tag{31}$$

respectively, where $\hat{\alpha}_{1g}(p_{i1}, \cdots, p_{g(s)}, \cdots, p_{iG})$ and $\hat{\alpha}_{2g}(p_{i1}, \cdots, p_{g(s)}, \cdots, p_{iG})$ are the estimated intercept and slope parameters in model (14) when $\tilde{p} = (p_{i1}, \cdots, p_{g(s)}, \cdots, p_{iG})$.

Finally, I close this section with five remarks on calculating $\hat{a}(p_i)$ and $\hat{b}(p_i)$.

Remark 1: $\hat{b}(p_{i1},\cdots,p_{g(s)},\cdots,p_{iG})$ in the denominator of equation (31) represents the value of $b(p_i)$ evaluated at $p_{ig(s)}=(p_{i1},\cdots,p_{g(s)},\cdots,p_{iG})$. It can be approximated using the kernel regression

$$\hat{b}(p_{i1},\cdots,p_{g(s)},\cdots,p_{iG})=\sum_{j=1}^{n}K\left(h_b^{-1}(p_j-p_{ig(s)})B_b\right)b(p_jB_b)\Big/\sum_{j=1}^{n}K\left(h_b^{-1}(p_j-p_{ig(s)})B_b\right),$$

where h_b denotes the bandwidth, and p_jB_b denotes the NNIR variates of $b(p_j)$ assuming $b(p_j)=b(p_jB_b)$.

Remark 2: It is better not to set $p_{g(1)}=\min\{p_{ig}\}$ and $p_{g(S)}=\max\{p_{ig}\}$ because these extreme values are likely to be the outliers of the price variable. To avoid this problem, one can set $p_{g(1)}$ and $p_{g(S)}$ to the 5% and 95% quantiles, respectively.

Remark 3: The larger the order of integration *S*, the more precise the integration result will be. But on the other hand, the longer it will take to obtain the result. I suggest experimenting with several values of *S* to ensure that a good approximation is achieved within a reasonable amount of time.

Remark 4: $\hat{\alpha}_{1g}(\tilde{p})$ and $\hat{\alpha}_{2g}(\tilde{p})$ can be estimated by regressing $\hat{F}_{0g}(\tilde{p},c_i)$ on $[1 \quad c_i]$. Although the nonparametric estimate $\hat{F}_{0g}(\tilde{p},c_i)$ has a slower convergence rate than $n^{-1/2}$, both $\hat{\alpha}_{1g}(\tilde{p})$ and $\hat{\alpha}_{2g}(\tilde{p})$ are $\sqrt{n}$-consistent because they are weighted averages of $\hat{F}_{0g}(\tilde{p},c_i)$.

Remark 5: The consistency of $\hat{\alpha}_{1g}(\tilde{p})$ and $\hat{\alpha}_{2g}(\tilde{p})$ implies the consistency of $\hat{a}(p_i)$ and $\hat{b}(p_i)$ as both n and S approach to infinity.

5. Conclusions

In this paper, I specify a semi-parametric demand model that is consistent with both demand theory and the high dimensional CLSS data. The semi-parametric model comprises: (i) The budget share equations $E\left(w_i \middle| p_i, c_i\right) = A_1(p_i) + A_2(p_i)c_i$, in which both $A_1(p_i)$ and $A_2(p_i)$ are estimated nonparametrically; (ii) The PIGLOG cost function $c(p_i, u_i) = a(p_i) + b(p_i)\,u_i$, in which both $a(p_i)$ and $b(p_i)$ are recovered from the nonparametric estimates of $A_1(p_i)$ and $A_2(p_i)$.

To develop the model, I propose a consistent nonparametric local rank test and show that the budget share Engel curves can be well approximated by linear functions of log total expenditure via basis function visualization. But the estimated budget share Engel curves has slopes (i.e., $A_2(p_i)$ in the budget share equations) varying considerably with prices, which rejects the specification of the AIDS with constant slopes. The constant slopes of the AIDS come from its misspecified PIGLOG cost function. To avoid model misspecification, I propose a reverse modeling strategy which recovers the PIGLOG cost function from the semi-parametric budget share equations. The reverse modeling strategy is more effective than the traditional modeling strategy using Shepard's Lemma (or Roy's Identity), as it proceeds from observable budget share equations to an unobservable cost function rather than starting from a hypothetical cost function (or utility function). Together, the three steps, "a nonparametric local rank test – basis function visualization – a reverse demand modeling strategy", provide a unified nonparametric and parametric approach for modeling high dimensional demand systems.

Appendix A

The local rank test requires estimating budget share Engel curves nonparametrically conditional on prices. But in practice, because of the "*curse of dimensionality*", price variables are often discarded when estimating nonparametric Engel curves. To see why discarding price variables yields erroneous results, consider the following budget share equations

$$w_i = F_0(p_i, c_i) + U_i \tag{A-1}$$

where i represents an individual household, w, p, and c, and U denote the budget shares, log prices, log total expenditure, and error term. $E(w_i | p_i, c_i) = F_0(p_i, c_i)$ implies $E(U_i | p_i, c_i) = 0$ with probability one. Given exogenous prices, the budget share Engel curves are the expectation of budget shares conditioning on $p_i = \tilde{p}$ and c_i, i.e.,

$$\begin{aligned} E(w_i | p_i = \tilde{p}, c_i) &= E\left(F_0(p_i, c_i) + U_i | p_i = \tilde{p}, c_i\right) \\ &= E\left(F_0(p_i, c_i) | p_i = \tilde{p}, c_i\right) + E\left(U_i | p_i = \tilde{p}, c_i\right) \\ &= F_0(\tilde{p}, c_i) + E\left[E\left(U_i | p_i, c_i\right) | p_i = \tilde{p}\right] = F_0(\tilde{p}, c_i). \end{aligned} \tag{A-2}$$

If the price variables are ignored, the budget share Engel curves are estimated using

$$\begin{aligned} E(w_i | c_i) &= E\left(F_0(p_i, c_i) + U_i | c_i\right) \\ &= E\left(F_0(p_i, c_i) | c_i\right) + E\left(U_i | c_i\right) \\ &= E_p\left[E\left(F_0(p_i, c_i) | p_i, c_i\right)\right] + E_p\left[E\left(U_i | p_i, c_i\right)\right] = E_p\left[F_0(p_i, c_i)\right], \end{aligned} \tag{A-3}$$

where E_p indicates the expectation is taken with respect to the prices. $E_p[F_0(p_i, c_i)]$ in (A-3) is different from $F_0(\tilde{p}, c_i)$ in (A-2). Note that $F_0(\tilde{p}, c_i)$ represents the budget share functions evaluated at the given price vector $\tilde{p}$; whereas $E_p[F_0(p_i, c_i)]$ represents the budget share functions averaged over the price variables. $F_0(\tilde{p}, c_i)$ can be the same as $E_p[F_0(p_i, c_i)]$ only if $\tilde{p} = E(p_i)$ and F_0 is linear in p. However, there exists little

empirical evidence that F_0 is linear in p. Thus, $E(w_i|p_i=\tilde{p},c_i)\neq E(w_i|c_i)$ in general, and the common practice of ignoring price variables leads to incorrect estimates of nonparametric budget share Engel curves.

Appendix B

Nearest Neighbor Inverse Regression (NNIR) algorithm:

Step 1: Standardize X by an affine transformation to yield $\hat{Z}_i=\hat{\Sigma}_{XX}^{-1/2}(X_i-\overline{X})$ $(i=1,2,\cdots,n)$, where $\hat{\Sigma}_{XX}$ and $\overline{X}$ are the sample covariance and sample mean of X, respectively.

Step 2: For each Y_i, find its nearest neighbor Y_{i^*}. Specifically, let $i^*\in\{1,2,\cdots,n\}-\{i\}$ be the indices for which $d(Y_i,Y_{i^*})=\min_{j\neq i,1\leq j\leq n}d(Y_i,Y_j)$, where $d(\cdot,\cdot)$ is some metric such as Euclidean distance.

Step 3: Calculate the matrix $\hat{M}_{NNIR}=(2n)^{-1}\sum_{i=1}^{n}(\hat{Z}_i\hat{Z}_{i^*}'+\hat{Z}_{i^*}\hat{Z}_i')$, where $\hat{Z}_{i^*}$ is the concomitant of $\hat{Z}_i$ according to the nearest neighbor relationship (Y_i, Y_{i^*}) found in Step 2.

Step 4: Compute the eigenvalues and eigenvectors for $\hat{M}_{NNIR}$.

Step 5: Let $\hat{\eta}=(\hat{\eta}_1\ \cdots\ \hat{\eta}_d)$ be the d largest eigenvectors (column vectors) of $\hat{M}_{NNIR}$. The outputs $X_i\hat{B}$, where $\hat{B}=\Sigma_{XX}^{-1/2}\hat{\eta}$, are the estimates of NNIR variates.

Step 6: Perform cross validation to jointly determine the number of NNIR variates, *d*, and the bandwidth for kernel regression, *h*.

Appendix C

In this appendix, I prove the consistency of $\tilde{\Gamma}$ under assumptions A1-A7. First, I decompose

$$\tilde{\Gamma}=\frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}\left(F_1(X_jB)+U_j\right)\left(F_1(X_kB)+U_k\right)'=\Xi_1+\Xi_2+\Xi_3,$$

where

$$\Xi_1 = \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{ik} F_1(X_j B) F_1(X_k B)' ,$$

$$\Xi_2 = \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \neq j}^{n} \left[\tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{ik} F_1(X_j B) U_k' + \tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{ik} U_j F_1(X_k B)' \right], \text{ and}$$

$$\Xi_3 = \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{ik} U_j U_k' .$$

I then show that (a) $\Xi_1 \xrightarrow{p} \Gamma$, (b) $\Xi_2 = O_p(n^{-1/2})$, and (c) $\Xi_3 = O_p(n^{-1} h^{-d/2})$. The result (a) holds because

$$\begin{aligned}
\Xi_1 &= \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{ik} F_1(X_j B) F_1(X_k B)' \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{n} \sum_{j=1}^{n} \tilde{\Upsilon}_{ij} \cdot \frac{1}{n-1} \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ik} \right] \left[\sum_{j=1}^{n} \tilde{\Upsilon}_{ij} F_1(X_j B) \Big/ \sum_{j=1}^{n} \tilde{\Upsilon}_{ij} \right] \left[\sum_{k \neq j}^{n} \tilde{\Upsilon}_{ik} F_1(X_k B)' \Big/ \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ik} \right] \\
&\xrightarrow{p} \frac{1}{n} \sum_{i=1}^{n} p(\tilde{X}_i B)^2 F_1(\tilde{X}_i B) F_1(\tilde{X}_i B)' \xrightarrow{p} E\left[p(\tilde{X}_i B)^2 F_1(\tilde{X}_i B) F_1(\tilde{X}_i B)' \right] = \Gamma.
\end{aligned}$$

The first $\xrightarrow{p}$ is due to the standard results for kernel estimates

$$\frac{1}{n} \sum_{j=1}^{n} \tilde{\Upsilon}_{ij} \xrightarrow{p} p(\tilde{X}_i B), \ \frac{1}{n-1} \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ik} \xrightarrow{p} p(\tilde{X}_i B),$$

$$\sum_{j=1}^{n} \tilde{\Upsilon}_{ij} F_1(X_j B) \Big/ \sum_{j=1}^{n} \tilde{\Upsilon}_{ij} \xrightarrow{p} F_1(\tilde{X}_i B), \ \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ik} F_1(X_k B)' \Big/ \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ik} \xrightarrow{p} F_1(\tilde{X}_i B)'$$

if $nh^d \to \infty$ and $h \to 0$ as $n \to \infty$ and by the Slutsky Theorem. The second $\xrightarrow{p}$ holds by the Law of Large Numbers. The results (b) and (c) follow from

$$\frac{1}{n^{3/2}(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{ik} F_0(X_j B) U_k' = O_p(1) \quad \text{and} \quad \frac{h^{d/2}}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{ik} U_j U_k' = O_p(1) ,$$

respectively (see Appendix D for proof). These results imply that $\tilde{\Gamma} \xrightarrow{p} \Gamma$ if $nh^d \to \infty$ as $n \to \infty$ because $\tilde{\Gamma} = \Xi_1 + \Xi_2 + \Xi_3 = \left(\Gamma + o_p(1) \right) + O_p(n^{-1/2}) + O_p(n^{-1} h^{-d/2}) = \Gamma + o_p(1).$

Next, I consider the eigenvalues of $\tilde{\Gamma}\Sigma^{-1}$ and their asymptotic properties under H_0: $\tilde{L}_0 = \tilde{L}$. The eigenvalues are invariant to the transformation of $\Psi'\tilde{\Gamma}\Psi$ and $\Psi'\Sigma\Psi$, where Ψ is a nonsingular $G \times G$ matrix chosen so that $\Psi'\Xi_1\Psi$ is diagonal and contains the eigenvalues of $\Xi_1\Sigma^{-1}$ and $\Psi'\Sigma\Psi = I_G$. Partition $\Psi = \begin{pmatrix}\Psi_1 & \Psi_2\end{pmatrix}$ such that $\Psi_2'\Xi_1\Psi_2 = \underline{0}$, where Ψ_1 is $G \times \tilde{L}$ and Ψ_2 is $G \times (G - \tilde{L})$. Using Lemma 1 of Fujikoshi (1977), the smallest $G - \tilde{L}$ eigenvalues of $\Psi'\tilde{\Gamma}\Psi$ are equal to the eigenvalues of

$$0 \cdot I_{G-L} + \Psi_2'\Xi_2\Psi_2 + \Psi_2'\Xi_3\Psi_2 + \Psi_2'\Xi_2\Psi_1\left(\Psi_1'\Xi_1\Psi_1\right)^{-1}\Psi_1'\Xi_2\Psi_2.$$

Note that $\Psi_2'\Xi_1\Psi_2 = \underline{0}$ implies $\Psi_2'\Xi_2\Psi_2 = \underline{0}$ and $\Psi_2'\Xi_2\Psi_1 = \underline{0}$, which in turn implies that $0 \cdot I_{G-L} + \Psi_2'\Xi_2\Psi_2 + \Psi_2'\Xi_3\Psi_2 + \Psi_2'\Xi_2\Psi_1\left(\Psi_1'\Xi_1\Psi_1\right)^{-1}\Psi_1'\Xi_2\Psi_2 = \Psi_2'\Xi_3\Psi_2$. So, the sum of the smallest $G - \tilde{L}$ eigenvalues of $\tilde{\Gamma}\Sigma^{-1}$ is

$$\sum_{g=1}^{G-\tilde{L}} \lambda_g\left(\tilde{\Gamma}\Sigma^{-1}\right) = \sum_{g=1}^{G-\tilde{L}} \lambda_g\left(\Psi'\tilde{\Gamma}\Psi \cdot (\Psi'\Sigma\Psi)^{-1}\right) = \sum_{g=1}^{G-\tilde{L}} \lambda_g\left(\Psi'\tilde{\Gamma}\Psi\right) = \sum_{g=1}^{G-\tilde{L}} \lambda_g\left(\Psi_2'\Xi_3\Psi_2\right).$$

The sum of the smallest $G - \tilde{L}$ eigenvalues of $\Psi_2'\Xi_3\Psi_2$ is a second-order *U*-statistic with kernel $g\left(X_j, U_j, X_k, U_k\right) = \frac{1}{2}\left[\Phi_{jk}U_k'\Psi_2\Psi_2'U_j + \Phi_{kj}U_j'\Psi_2\Psi_2'U_k\right]$, where $\Phi_{jk} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}$, because

$$\begin{aligned}
&\sum_{g=1}^{G-\tilde{L}} \lambda_g\left(\Psi_2'\Xi_3\Psi_2\right) = tr\left(\Psi_2'\Xi_3\Psi_2\right) = \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n} tr(\Psi_2'\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}U_jU_k'\Psi_2) \\
&= \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}U_k'\Psi_2\Psi_2'U_j = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}\right)U_k'\Psi_2\Psi_2'U_j \\
&= \frac{2}{n(n-1)}\sum_{j=1}^{n}\sum_{k<j}^{n}\frac{1}{2}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}\right)U_k'\Psi_2\Psi_2'U_j + \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}\right)U_j'\Psi_2\Psi_2'U_k\right].
\end{aligned}$$

By the Central Limit Theorem for *U*-statistics, $nh^{d/2}\sum_{j=1}^{G-L}\lambda_j\left(\Psi_2'\Xi_3\Psi_2\right)$ has a normal distribution with mean zero (because $E\left(U_jU_k' \mid X_j, X_k\right) = 0 \;\forall\; j \neq k$) and variance

$\sigma^2 = 2(G-\tilde{L}) \| K \|_4^3 \, E\left(p(X_i B)^3\right)$. The second-order *U*-statistic has the first order degeneracy because $E\left[g\left(X_j, U_j, X_k, U_k\right) | (X_j, U_j)\right] = 0$. Therefore, the variance of the second-order *U*-statistic is determined by its second order term $E\left[g\left(X_j, U_j, X_k, U_k\right) | (X_j, U_j), (X_k, U_k)\right]$, which is simply $g\left(X_j, U_j, X_k, U_k\right)$. To derive the variance σ^2, I calculate the following variance and covariance terms

$$
\begin{aligned}
&Var\left(\Phi_{jk} U_k' \Psi_2 \Psi_2' U_j\right) = E\left(\Phi_{jk} U_k' \Psi_2 \Psi_2' U_j \cdot U_j' \Psi_2 \Psi_2' U_k \Phi_{kj}\right) \\
&= E\left[tr\left(\Phi_{jk}^2 \cdot \Psi_2' U_j U_j' \Psi_2 \cdot \Psi_2' U_k U_k' \Psi_2\right)\right] = tr\left[E\left(\Phi_{jk}^2 \cdot \Psi_2' U_j U_j' \Psi_2 \cdot \Psi_2' U_k U_k' \Psi_2\right)\right] \\
&= tr\left\{E\left[E\left(\Phi_{jk}^2 \cdot \Psi_2' U_j U_j' \Psi_2 \cdot \Psi_2' U_k U_k' \Psi_2 \middle| X_j, X_k\right)\right]\right\} \\
&= tr\left\{E\left[\Phi_{jk}^2 \cdot \Psi_2' E\left(U_j U_j' \middle| X_j\right) \Psi_2 \cdot \Psi_2' E\left(U_k U_k' \middle| X_k\right)\right) \Psi_2\right]\right\} \\
&= tr\left[E\left(\Phi_{jk}^2\right) \cdot \Psi_2' \Sigma \Psi_2 \cdot \Psi_2' \Sigma \Psi_2\right] = E\left(\Phi_{jk}^2\right) \cdot tr\left(I_{G-\tilde{L}}\right) = (G-\tilde{L}) E\left(\Phi_{jk}^2\right)
\end{aligned}
$$

$$
Cov\left(\Phi_{jk} U_k' \Psi_2 \Psi_2' U_j, \Phi_{kj} U_j' \Psi_2 \Psi_2' U_k\right) = E\left(\Phi_{jk} U_k' \Psi_2 \Psi_2' U_j \cdot U_j' \Psi_2 \Psi_2' U_k \Phi_{kj}\right) = (G-\tilde{L}) E\left(\Phi_{jk}^2\right).
$$

These results suggest that the variance of the second-order *U*-statistic is $\frac{2(G-\tilde{L})}{n(n-1)} E\left(\Phi_{jk}^2\right)$.

Consequently, the variance of $nh^{d/2} \sum_{j=1}^{G-L} \lambda_j \left(\Psi_2' \Xi_3 \Psi_2\right)$ is

$$
\begin{aligned}
\sigma^2 &= (nh^{d/2})^2 \, \frac{2(G-\tilde{L})}{n(n-1)} E\left(\Phi_{jk}^2\right) = 2(G-\tilde{L}) h^d E\left(\Phi_{jk}^2\right) + o(1) \\
&= 2(G-\tilde{L}) \| K \|_4^3 \, E\left(p(X_i B)^3\right) + o(1)
\end{aligned}
$$

given that

$$
\begin{aligned}
&E\left(\Phi_{jk}^2\right) = \int\int \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{\Upsilon}_{ik} \tilde{\Upsilon}_{ij}\right)^2 p(X_k B) p(X_j B) dX_k B dX_j B \\
&= \int\int \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \tilde{\Upsilon}_{ik} \tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{i'k} \tilde{\Upsilon}_{i'j}\right) p(X_k B) p(X_j B) dX_k B dX_j B \\
&\xrightarrow{p} h^{-4d} \iint\iint K\left(h^{-1}(X_k - \tilde{X}_i) B\right) K\left(h^{-1}(X_j - \tilde{X}_i) B\right) K\left(h^{-1}(X_k - \tilde{X}_{i'}) B\right) K\left(h^{-1}(X_j - \tilde{X}_{i'}) B\right) \\
&\qquad p(X_k B) p(X_j B) p(\tilde{X}_i B) p(\tilde{X}_{i'} B) dX_k B dX_j B d\tilde{X}_i B d\tilde{X}_{i'} B
\end{aligned}
$$

$$= h^{-d} \iint\iint K(\varphi_1) K(\varphi_2) K(\varphi_1 - \varphi_3) K(\varphi_2 - \varphi_3)$$
$$p(\tilde{X}_i B + h\varphi_1) p(\tilde{X}_i B + h\varphi_2) p(\tilde{X}_i B) p(\tilde{X}_i B - h\varphi_3) d\varphi_1 d\varphi_2 d\varphi_3 d\tilde{X}_i B$$
$$= h^{-d} \iiint K(\varphi_1) K(\varphi_2) K(\varphi_1 - \varphi_3) K(\varphi_2 - \varphi_3) d\varphi_1 d\varphi_2 d\varphi_3$$
$$\int p(\tilde{X}_i B + h\varphi_1) p(\tilde{X}_i B + h\varphi_2) p(\tilde{X}_i B - h\varphi_3) p(\tilde{X}_i B) d\tilde{X}_i B$$
$$= h^{-d} \| K \|_4^3 \, E\left(p(\tilde{X}_i B)^3 \right) + o(1),$$

where $\| K \|_4^3 = \iiint K(\varphi_1) K(\varphi_2) K(\varphi_1 - \varphi_3) K(\varphi_2 - \varphi_3) d\varphi_1 d\varphi_2 d\varphi_3$. The third equality is obtained using the change of variables, from $X_k B$ to $\varphi_1 = h^{-1}(X_k - \tilde{X}_i)B$, from $X_j B$ to $\varphi_2 = h^{-1}(X_j - \tilde{X}_i)B$, and from $\tilde{X}_i B$ to $\varphi_3 = h^{-1}(\tilde{X}_i - \tilde{X}_{i'})B$, with the Jacobians h^{-d}, h^{-d}, and h^{-d}, respectively. The variance σ^2 suggests that the rescaling factor $V = \left[2(G - \tilde{L}) \| K \|_4^3 \, E\left(p(\tilde{X}_i B)^3 \right) \right]^{-1/2}$ so that the test statistic has asymptotic variance one.

Appendix D

First, I prove $\frac{1}{n^{3/2}(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{ik} F_1(X_j B) U_k' = O_p(1)$ under assumptions A1-A7.

The (t, s) element of $\frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{ik} F_1(X_j B) U_k'$, denoted by $\tilde{U}_n$, is a second-order U-statistic with kernel $g(X_j, U_j, X_k, U_k) = \frac{1}{2}\left[\Phi_{jk} F_{1t}(X_j B) U_{ks} + \Phi_{kj} F_{1t}(X_k B) U_{js} \right]$, because

$$\tilde{U}_n = \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \neq j}^{n} \tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{ik} F_{1t}(X_j B) U_{ks}$$
$$= \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{k \neq j}^{n} \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{\Upsilon}_{ij} \tilde{\Upsilon}_{ik} \right) F_{1t}(X_j B) U_{ks} = \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{k \neq j}^{n} \Phi_{jk} F_{1t}(X_j B) U_{ks}$$
$$= \frac{2}{n(n-1)} \sum_{j=1}^{n} \sum_{k<j}^{n} \frac{1}{2}\left[\Phi_{jk} F_{1t}(X_j B) U_{ks} + \Phi_{kj} F_{1t}(X_k B) U_{js} \right].$$

To establish the asymptotic properties of $\tilde{U}_n$, I need to verify the following three conditions: (i) $E\left[g(X_j,U_j,X_k,U_k)\right]=0$, (ii) $E\left\{E\left[g(X_j,U_j,X_k,U_k)\,|\,(X_j,U_j)\right]^2\right\}=C<\infty$, and (iii) $E\left[g(X_j,U_j,X_k,U_k)^2\right]=o(n)$. The first condition states $E\left(\tilde{U}_n\right)=0$ and the proof is trivial because $E(U_{ks})=0\;\forall\,k,s$. The second condition implies that the projection of $g(X_j,U_j,X_k,U_k)$ onto (X_j,U_j) has a finite variance. Therefore, the projection of $\tilde{U}_n$ onto (X_j,U_j), denoted by $\vec{U}_n$, has a variance of order $O(n^{-1})$, which in turn implies that $\sqrt{n}\,\vec{U}_n=O_p(1)$. The proof for condition (ii) is as follows:

$$
\begin{aligned}
&E\left[g(X_j,U_j,X_k,U_k)\,|\,(X_j,U_j)\right]\\
&=\frac{1}{2}E\left[\Phi_{jk}F_{1t}(X_jB)U_{ks}+\Phi_{kj}F_{1t}(X_kB)U_{js}\,|\,(X_j,U_j)\right]\\
&=\frac{1}{2}E\left[\Phi_{jk}\,|\,X_j\right]F_{1t}(X_jB)E(U_{ks})+\frac{1}{2}E\left[\Phi_{kj}F_{1t}(X_kB)\,|\,X_j\right]U_{js}\\
&=\frac{1}{2}E\left[\Phi_{kj}F_{1t}(X_kB)\,|\,X_j\right]U_{js}
\end{aligned}
$$

$$
\begin{aligned}
&E\left[\Phi_{kj}F_{1t}(X_kB)\,|\,X_j\right]=\int\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\Upsilon}_{ik}\tilde{\Upsilon}_{ij}\right)F_{1t}(X_kB)p(X_kB\,|\,X_jB)dX_kB\\
&\xrightarrow{\;p\;}\int\left[\int\tilde{\Upsilon}_{ik}\tilde{\Upsilon}_{ij}d\tilde{X}_iB\right]F_{1t}(X_kB)p(X_kB)dX_kB\\
&=h^{-2d}\int\int K\left(h^{-1}(X_k-\tilde{X}_i)B\right)K\left(h^{-1}(X_j-\tilde{X}_i)B\right)F_{1t}(X_kB)p(X_kB)dX_kBd\tilde{X}_iB\\
&=h^{-d}\int\int K(\varphi_1)K\left(h^{-1}(X_j-\tilde{X}_i)B\right)F_{1t}(\tilde{X}_iB+h\varphi_1)p(\tilde{X}_iB+h\varphi_1)d\varphi_1d\tilde{X}_iB\\
&=h^{-d}\int K\left(h^{-1}(X_j-\tilde{X}_i)B\right)F_{1t}(\tilde{X}_iB)p(\tilde{X}_iB)d\tilde{X}_iB+o(1)\\
&=\int K(\varphi_2)F_{1t}(X_jB-h\varphi_2)p(X_jB-h\varphi_2)d\varphi_2+o(1)\\
&=F_{1t}(X_jB)p(X_jB)+o(1)
\end{aligned}
$$

$$
\begin{aligned}
&E\left\{E\left[g(X_j,U_j,X_k,U_k)\,|\,(X_j,U_j)\right]^2\right\}\\
&=\frac{1}{4}E\left\{E^2\left[\Phi_{kj}F_{1t}(X_kB)\,|\,X_j\right]U_{js}^2\right\}=\frac{1}{4}E\left[F_{1t}(X_jB)^2\,p(X_jB)^2\right]E\left(U_{js}^2\right)=C<\infty
\end{aligned}
$$

because both $E\left[F_{1t}(X_jB)^2\,p(X_jB)^2\right]$ and $E\left(U_{js}^2\right)$ are finite constants under assumptions A3-A5. The condition (iii) is essential for applying Lemma 3.1 of Powell, Stock, and Stocker (1989), which proves the asymptotic equivalence of $\tilde{U}_n$ and its projection $\vec{U}_n$. The proof for condition (iii) is as follows:

$$
\begin{aligned}
&E\left[g(X_j,U_j,X_k,U_k)^2\right]=\frac{1}{4}E\left[\left(\Phi_{jk}F_{1t}(X_jB)U_{ks}+\Phi_{kj}F_{1t}(X_kB)U_{js}\right)^2\right]\\
&=\frac{1}{4}E\left[\left(\Phi_{jk}F_{1t}(X_jB)U_{ks}\right)^2+\left(\Phi_{kj}F_{1t}(X_kB)U_{js}\right)^2+2\Phi_{jk}\Phi_{kj}F_{1t}(X_jB)F_{1t}(X_kB)U_{ks}U_{js}\right]\\
&=\frac{1}{2}E\left[\left(\Phi_{jk}F_{1t}(X_jB)U_{ks}\right)^2\right]=\frac{1}{2}E\left[\left(\Phi_{jk}F_{1t}(X_jB)\right)^2\right]E\left(U_{ks}^2\right)=o(n)
\end{aligned}
$$

because $E\left(U_{ks}^2\right)=O(1)$ and $E\left[\left(\Phi_{jk}F_{1t}(X_jB)\right)^2\right]=o(n)$. $E\left[\left(\Phi_{jk}F_{1t}(X_jB)\right)^2\right]=o(n)$ if $nh^d\to\infty$ as $n\to\infty$ and the proof is as follows:

$$
\begin{aligned}
&E\left[\left(\Phi_{jk}F_{1t}(X_jB)\right)^2\right]=E_{X_j}\left\{E\left[\left(\Phi_{jk}F_{1t}(X_jB)\right)^2\middle|X_j\right]\right\}\\
&=E_{X_j}\left\{E\left[{\Phi_{jk}}^2\middle|X_j\right]F_{1t}(X_jB)^2\right\}=h^{-d}\parallel K\parallel_4^3\,E_{X_j}\left[p(X_jB)^3F_{1t}(X_jB)^2\right]+o(1)\\
&=O(h^{-d})=o(n)
\end{aligned}
$$

where

$$
\begin{aligned}
&E\left[{\Phi_{jk}}^2\middle|X_j\right]=\int\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\Upsilon}_{ik}\tilde{\Upsilon}_{ij}\right)^2p(X_kB\mid X_jB)dX_kB=\int\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\tilde{\Upsilon}_{ik}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{i'k}\tilde{\Upsilon}_{i'j}\right)p(X_kB)dX_kB\\
&\xrightarrow{\;p\;}\int\left(\iint\tilde{\Upsilon}_{ik}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{i'k}\tilde{\Upsilon}_{i'j}p(\tilde{X}_iB)p(\tilde{X}_{i'}B)d\tilde{X}_iBd\tilde{X}_{i'}B\right)p(X_kB)dX_kB\\
&=h^{-4d}\iiint K\left(h^{-1}(X_k-\tilde{X}_i)B\right)K\left(h^{-1}(X_j-\tilde{X}_i)B\right)K\left(h^{-1}(X_k-\tilde{X}_{i'})B\right)K\left(h^{-1}(X_j-\tilde{X}_{i'})B\right)\\
&\qquad\qquad p(\tilde{X}_iB)p(\tilde{X}_{i'}B)p(X_kB)d\tilde{X}_iBd\tilde{X}_{i'}BdX_kB\\
&=h^{-d}\iiint K(\varphi_1-\varphi_3)K(\varphi_1)K(\varphi_2-\varphi_3)K(\varphi_2)\,p(X_jB-h\varphi_1)p(X_jB-h\varphi_2)p(X_jB-h\varphi_3)d\varphi_1d\varphi_2d\varphi_3\\
&=h^{-d}\iiint K(\varphi_1)K(\varphi_2)K(\varphi_1-\varphi_3)K(\varphi_2-\varphi_3)d\varphi_1d\varphi_2d\varphi_3\cdot p(X_jB)^3+o(1)\\
&=h^{-d}\parallel K\parallel_4^3\;p(X_jB)^3+o(1).
\end{aligned}
$$

The fourth equality is obtained using the change of variables, from $\tilde{X}_iB$ to $\varphi_1=h^{-1}(X_j-\tilde{X}_i)B$, from $\tilde{X}_{i'}B$ to $\varphi_2=h^{-1}(X_j-\tilde{X}_{i'})B$, and from X_kB to

$\varphi_3 = h^{-1}(X_j - X_k)B$, with the Jacobians h^{-d}, h^{-d}, and h^{-d}, respectively. Consequently, $E\left[g(X_j,U_j,X_k,U_k)^2\right] = o(n)$ holds. Therefore, by Lemma 3.1 of Powell, Stock, and Stocker (1989), $\sqrt{n}\left(\tilde{U}_n - \vec{U}_n\right) = o_p(1)$ holds. This result, along with $\sqrt{n}\,\vec{U}_n = O_p(1)$ implied by condition (ii), proves that $\sqrt{n}\,\tilde{U}_n = O_p(1)$, i.e., $\frac{1}{n^{3/2}(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}F_1(X_jB)U_k' = O_p(1)$.

Second, I prove $\frac{h^{d/2}}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}U_jU_k' = O_p(1)$ under assumptions A1-A7. The (t, s) element of $\frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}U_jU_k'$, denoted by $\tilde{U}_n$, is a second-order U-statistic with kernel $g(X_j,u_j,X_k,U_k) = \frac{1}{2}\left(\Phi_{jk}U_{jt}U_{ks} + \Phi_{kj}U_{kt}U_{js}\right)$ because

$$\begin{aligned}
\tilde{U}_n &= \frac{1}{n^2(n-1)}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\sum_{i=1}^{n}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}U_{jt}U_{ks} \\
&= \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}\right)U_{jt}U_{ks} = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\Phi_{jk}U_{jt}U_{ks} \\
&= \frac{2}{n(n-1)}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\frac{1}{2}\left(\Phi_{jk}U_{jt}U_{ks} + \Phi_{kj}U_{kt}U_{js}\right).
\end{aligned}$$

To establish the asymptotic properties of $\tilde{U}_n$, I need to verify the following three conditions: (i) $E\left[g(X_j,U_j,X_k,U_k)\right] = 0$, (ii) $E\left[g(X_j,U_j,X_k,U_k)\,|\,(X_j,U_j)\right] = 0$, and (iii) $E\left[g(X_j,U_j,X_k,U_k)^2\right] = O(h^{-d})$. The first condition states $E\left(\tilde{U}_n\right) = 0$ and the proof is trivial because $E(U_{jt}U_{ks}) = 0\ \forall\ j \neq k$. The second condition implies that $\tilde{U}_n$ has first order degeneracy due to $E(U_{kt}) = 0\ \forall\ k,t$. Thus, the convergence rate of $\tilde{U}_n$ is determined by the second order term, which is $g(X_j,U_j,X_k,U_k)$. The third condition proves that the second order term has a variance of order $O(h^{-d})$, and therefore the variance of $\tilde{U}_n$ is of

order $O(n^{-2}h^{-d})$, which in turn implies that $nh^{d/2}\tilde{U}_n = \frac{h^{d/2}}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\tilde{\Upsilon}_{ij}\tilde{\Upsilon}_{ik}U_jU_k' = O_p(1)$.

The proof for condition (iii) is as follows:

$$E\left[g(X_j,U_j,X_k,U_k)^2\right] = \frac{1}{4}E\left[\left(\Phi_{jk}U_{jt}U_{ks} + \Phi_{kj}U_{kt}U_{js}\right)^2\right]$$
$$= \frac{1}{4}E\left[\left(\Phi_{jk}U_{jt}U_{ks}\right)^2 + \left(\Phi_{kj}U_{kt}U_{js}\right)^2 + 2\Phi_{jk}\Phi_{kj}U_{jt}U_{ks}U_{kt}U_{js}\right]$$
$$= \frac{1}{2}E\left(\Phi_{jk}^2\right)\left[E\left(U_{jt}^2\right)E\left(U_{ks}^2\right) + E\left(U_{jt}U_{js}\right)E\left(U_{kt}U_{ks}\right)\right] = O(h^{-d})$$

because $E\left(U_{jt}^2\right)$ and $E\left(U_{jt}U_{js}\right)$ are constants given assumption A5 and $E\left(\Phi_{jk}^2\right) = h^{-d}\,\|K\|_4^3\,E\left(p(\tilde{X}_iB)^3\right) + o(1)$ as shown in Appendix C.

Table 1: Summary Statistics of the Ten Food Items

		Budget Shares		Prices				
No.	Food Item	Mean	Std. dev.	Mean	Std. dev.	Sample Median	Hebei Median	Liaoning Median
1	Vegetables	0.219	0.158	0.792	0.518	0.646	0.500	0.925
2	Milled rice	0.197	0.164	1.240	0.240	1.200	1.375	1.200
3	Flour	0.193	0.167	1.025	0.314	1.000	1.000	1.000
4	Pork	0.146	0.110	5.249	1.003	5.000	5.000	5.000
5	Vegetable oil	0.061	0.057	4.605	0.923	4.500	5.000	4.500
6	Eggs	0.044	0.047	2.732	0.423	2.750	2.800	2.667
7	Corn	0.043	0.077	1.092	0.796	0.900	0.950	0.850
8	Lard	0.034	0.042	4.275	1.199	4.500	4.833	3.750
9	Potatoes	0.032	0.035	0.419	0.151	0.475	0.333	0.500
10	Fruit	0.031	0.038	1.165	0.596	1.000	1.000	1.000

Table 2: Price Vectors for Local Rank Test

No.	Food Item	Vector 1 (i = 40)	Vector 2 (i = 242)	Vector 3 (i = 362)	Vector 4 (i = 219)	Vector 5 (i = 304)	Vector 6 (i = 751)
1	Vegetables	1.000	0.150	0.750	0.333	0.646	2.000
2	Milled rice	1.500	0.700	1.600	1.500	1.250	0.667
3	Flour	1.000	1.500	1.200	1.200	0.700	0.900
4	Pork	6.000	6.000	5.000	5.000	5.000	5.000
5	Vegetable oil	5.000	5.000	5.000	4.000	3.500	4.000
6	Eggs	3.000	2.642	3.000	2.250	2.750	2.800
7	Corn	0.950	1.000	0.750	1.000	0.750	0.850
8	Lard	4.000	4.000	5.000	4.833	6.000	3.000
9	Potatoes	0.500	0.500	0.200	0.500	0.150	0. 500
10	Fruit	1.417	1.250	0.667	0.817	1.667	0.750

Table 3: *P*-values of the Local Rank Tests (*d*=3, *h*=1.6)

	Rank			
Price Vectors	*L*=0	*L*=1	*L*=2	*L*=3
1	0.0000	0.0000	0.9998	0.9998
2	0.0000	0.0004	1.0000	1.0000
3	0.0000	0.0088	1.0000	1.0000
4	0.0000	0.1627	1.0000	1.0000
5	0.0000	0.3638	1.0000	1.0000
6	0.0000	0.9036	1.0000	1.0000

Table 4: Estimated Parameters of the Linear Budget Share Engel Curves

Panel (a): Estimated Intercept $A_1(\tilde{p})$

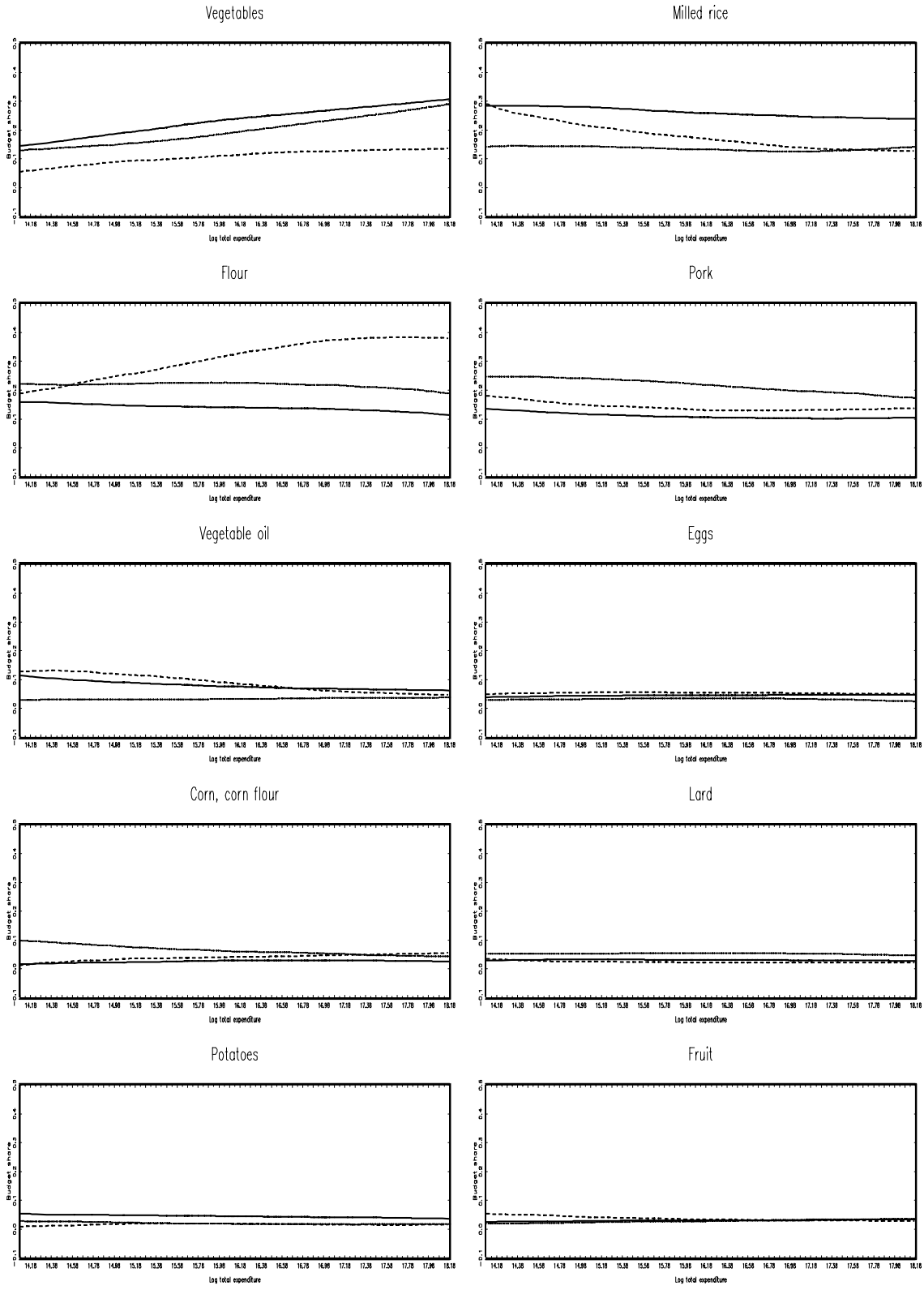
No.	Food Item	Vector 1 (i = 40)	Vector 2 (i = 242)	Vector 3 (i = 362)	Vector 4 (i = 219)	Vector 5 (i = 304)	Vector 6 (i = 751)
1	Vegetables	-0.4088	-0.1662	-0.4271	-0.0441	-0.3681	-0.3312
2	Milled rice	0.4040	0.5542	0.1799	0.4842	0.1063	0.3445
3	Flour	0.2768	-0.5806	0.3244	-0.0694	0.3790	0.2835
4	Pork	0.2423	0.2843	0.5080	0.1594	0.5467	0.1909
5	Vegetable oil	0.3211	0.5136	0.0057	0.3259	0.0334	0.1686
6	Eggs	0.0198	0.0433	0.0430	0.0727	0.0360	0.0918
7	Corn	-0.0126	-0.0969	0.2626	-0.1040	0.1567	0.0139
8	Lard	0.0399	0.0816	0.0787	0.0448	0.1209	0.0895
9	Potatoes	0.1240	0.1425	0.0661	0.0717	0.0560	0.1246
10	Fruit	-0.0066	0.2241	-0.0413	0.0589	-0.0669	0.0238

Panel (b): Estimated Slope $A_2(\tilde{p})$

No.	Food Item	Vector 1 (i = 40)	Vector 2 (i = 242)	Vector 3 (i = 362)	Vector 4 (i = 219)	Vector 5 (i = 304)	Vector 6 (i = 751)
1	Vegetables	0.0398	0.0171	0.0387	0.0116	0.0370	0.0406
2	Milled rice	-0.0090	-0.0238	-0.0029	-0.0194	-0.0005	-0.0079
3	Flour	-0.0085	0.0555	-0.0065	0.0238	-0.0106	-0.0121
4	Pork	-0.0082	-0.0090	-0.0181	-0.0017	-0.0193	-0.0052
5	Vegetable oil	-0.0149	-0.0264	0.0017	-0.0161	-0.0009	-0.0058
6	Eggs	0.0016	0.0006	-0.0006	-0.0014	-0.0012	-0.0020
7	Corn	0.0024	0.0085	-0.0123	0.0093	-0.0041	0.0002
8	Lard	-0.0005	-0.0035	-0.0015	-0.0013	-0.0040	-0.0039
9	Potatoes	-0.0049	-0.0075	-0.0029	-0.0030	-0.0024	-0.0047
10	Fruit	0.0022	-0.0115	0.0043	-0.0018	0.0059	0.0008

Figure 1: Nonparametric Budget Share Engel Curves

Panel (a): Price Vectors 1-3 (Rank Two)



Panel (b): Price Vectors 4-6 (Rank One)

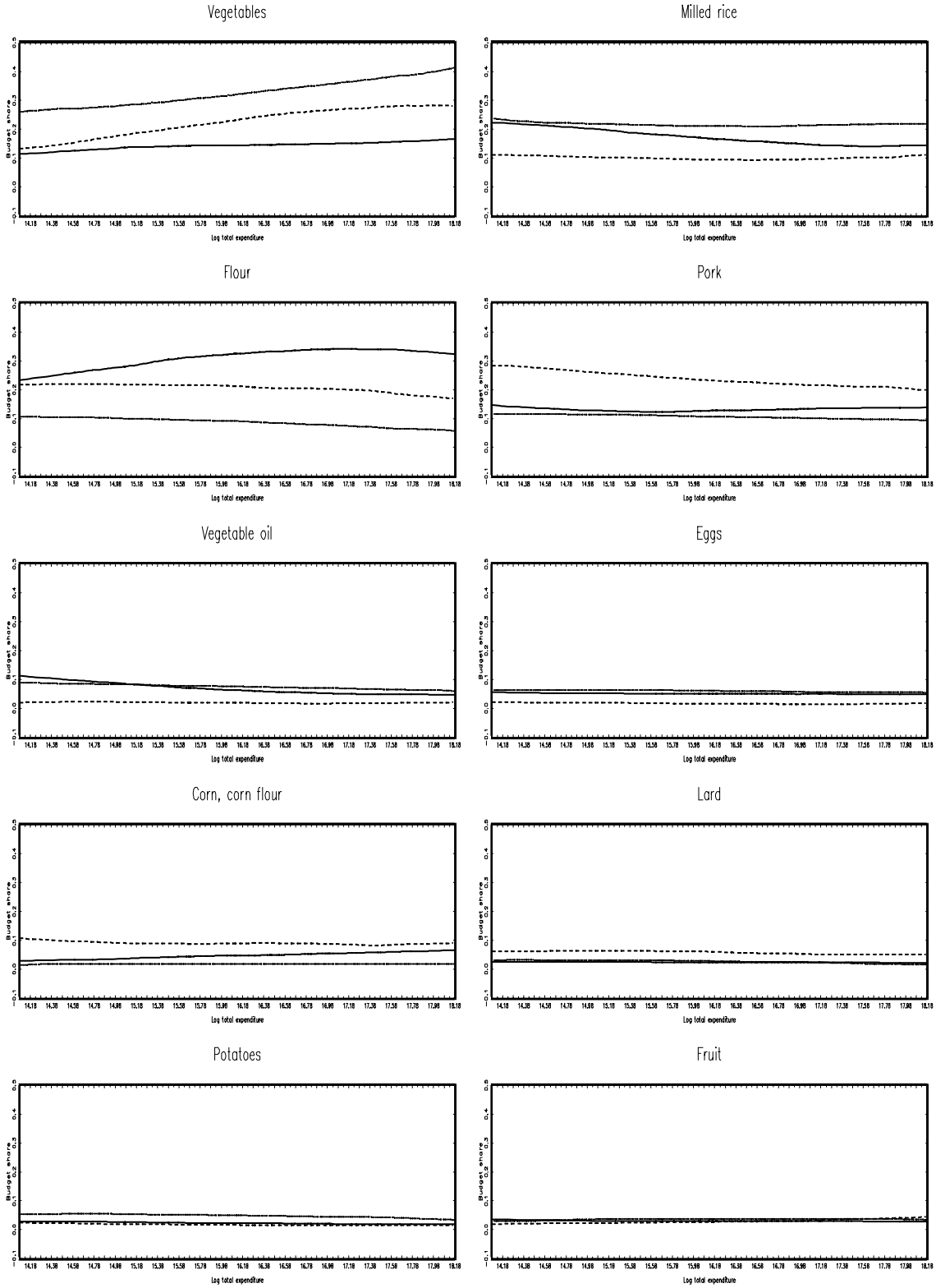
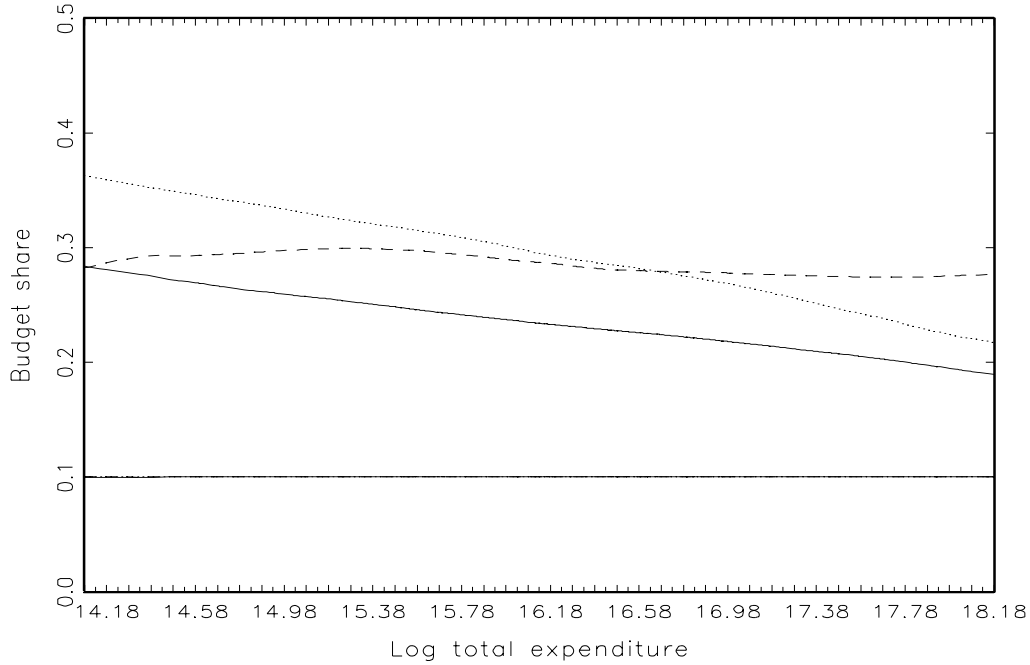
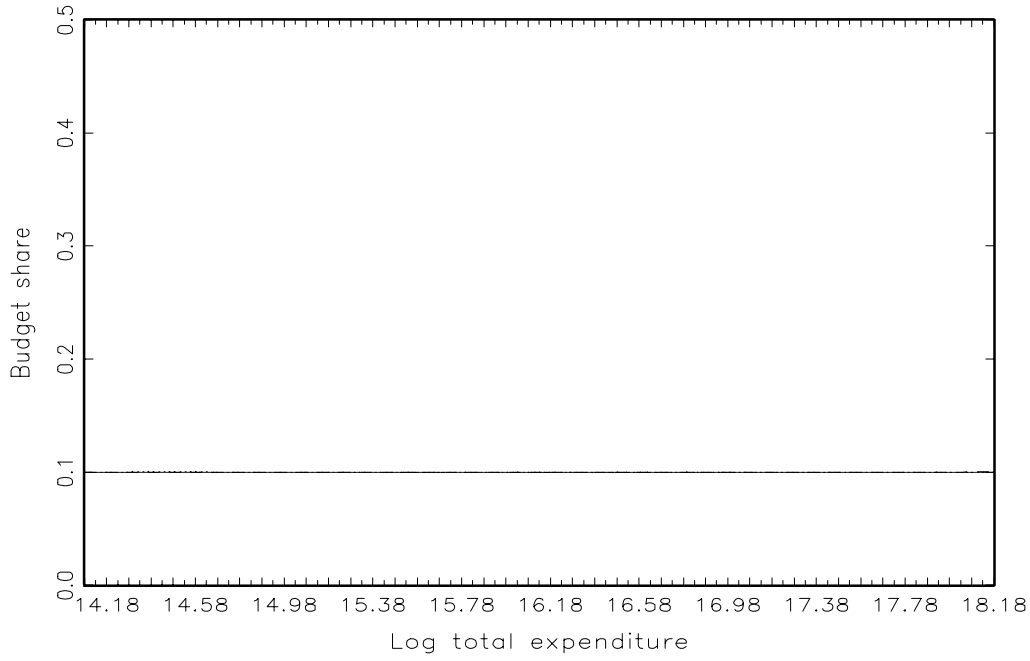


Figure 2: Estimated Basis Functions

Panel (a): Price Vectors 1-3 (Rank Two)



Panel (b): Price Vectors 4-6 (Rank One)



References

J. Banks, R. Blundell, A. Lewbel (1997), "Quadratic Engel Curves and Consumer Demand", The Review of Economics and Statistics, Vol. 79, pp. 527-539.

J. Chalfant, A. Gallant (1985), "Estimating Substitution Elasticities with the Fourier Cost Function: Some Monte Carlo results," Journal of Econometrics, Vol. 28(2), pp. 205-222.

J. Chalfant (1987), "A Globally Flexible, Almost Ideal Demand System", Journal of Business & Economic Statistics, Vol. 5, No. 2, pp. 233-242.

P. Chen and A. Smith (2007), "Dimension Reduction Using Inverse Regression and Nonparametric Factors", working paper.

L. Christensen, D. Jorgenson, L. Lau (1975), "Transcendental Logarithmic Utility Functions", The American Economic Review, Vol. 65, pp 367-383.

A. Deaton, J. Muellbauer (1980), "An Almost Ideal Demand System," The American Economic Review, Vol. 70, pp. 312-326.

S. Donald (1997), "Inference Concerning the Number of Factors in a Multivariate Nonparametric Relationship", Econometrica, Vol. 65, pp. 103-131.

Y. Fujikoshi (1977), "Asymptotic Expansions for the Distributions of Some Multivariate Tests", Multivariate Analysis, Vol. IV, edited by P. R. Krishnaiah. Amsterdam: North Holland, pp. 55-71.

W. Gorman (1981), "Some Engel Curves", in The Theory and Measurement of Consumer Behavior, Angus Deaton (ed.), Cambridge University Press.

T. Hsing (1999), "Nearest Neighbor Inverse Regression", Annals of Statistics, Vol. 27, pp. 697-731.

A. Lewbel (1989a), "A Demand System Rank Theorem", Econometrica, Vol. 57, pp. 701-705.

A. Lewbel (1991), "The Rank of Demand Systems: Theory and Nonparametric Estimation", Econometrica, Vol. 59, pp. 711-730.

A. Lewbel (2003), "A Rational Rank Four Demand System," Journal of Applied Econometrics, Vol. 18, pp. 127-135.

K.C. Li (1991), "Sliced Inverse Regression for Dimension Reduction", Journal of the American Statistical Association, Vol. 86, pp. 316-333.

K.C. Li (2000), "High Dimensional Data Analysis via the SIR/PHD Approach", Mimeo, UCLA.

J. Muellbauer (1975), "Aggregation, Income Distribution and Consumer Demand**",** The Review of Economic Studies, Vol. 42, pp. 525-543.

J. Muellbauer (1976), "Community Preferences and the Representative Consumer", Econometrica, Vol. 44, pp. 979-999.

Nicholas E. Piggott (2003), "The Nested PIGLOG Model: An Application to U.S. Food Demand", American Journal of Agricultural Economics, Vol. 85 (1), 1–15.

J. Powell, J. Stock, and T. Stocker (1989), "Semi-parametric Estimation of Index Coefficients", Econometrica, Vol. 55, pp. 875-891.

H. White (1980), "Using Least Squares to Approximate Unknown Regression Functions", International Economic Review, Vol. 21, pp. 149-170.

H. White (2001), Asymptotic Theory for Econometricians, 2nd edition, Academic Press.