

Low-Rank Demand Estimation in High-Dimensional Product Spaces*

Markus Trunschke^{1,2,3}

¹Hoover Institution, Stanford University, USA

²ZEW–Leibniz Centre for European Economic Research Mannheim, Germany

³KU Leuven, Belgium

August 2026

- Preliminary version. Do not cite or distribute -

Abstract

Flexible product-space demand systems like the Almost Ideal Demand System become infeasible to estimate for markets with more than a handful of products. Cross-price parameters grow quadratically in the number of products, requiring more than 200 time periods for identification in a 200-product market. Typical production panels span only 20 to 50 periods and lack product characteristics that characteristic-space models use to circumvent this issue. This paper develops an approximate low-rank approach to estimating the Almost Ideal Demand System that compresses the cross-price parameter block onto a small number of pervasive substitution dimensions, making flexible demand estimation feasible for large differentiated-products markets without observed product characteristics. Estimation of both the number and the structure of the substitution dimensions is entirely data-driven. The parameter count falls from over 20,000 to around 2,300 for a 200-product market with 10 substitution dimensions. Monte Carlo simulations confirm reliable rank recovery and precise structural estimates across market sizes from 20 to 500 products.

Keywords: demand estimation, almost ideal demand system, low-rank approximation, rank selection, differentiated products markets, product space, MPEC

JEL Classification: C13, C38, D12, L11

*I thank Dr. Kenneth L. Judd at the Hoover Institution for helpful discussions and his financial support. Correspondence to: markus.trunschke@stanford.edu (M. Trunschke)

1 Introduction

Estimating demand systems is central to empirical industrial organization and empirical economics as a whole. Recovered substitution patterns determine markups, pass-through, merger simulations, welfare counterfactuals, and the design of competition policy. As firm-level and product-level datasets grow to cover hundreds of products within a single market, the empirical environment that the demand-estimation toolkit has to handle becomes increasingly challenging. Datasets routinely contain a large number of products relative to the time dimension, straining flexible demand systems that model all possible substitution patterns between products against the high parameter growth in the number of products. For the Almost Ideal Demand System and its linear approximation, the number of free cross-price parameters grows quadratically in the number of products. With 200 products, the price coefficient matrix alone accounts for 19,900 of the system’s 20,298 free parameters, requiring at least 201 time periods for the theoretical ability to identify all parameters, while typical annual firm-level panels span only 20 to 50 periods. The same dimensionality problem affects other flexible demand systems in the same lineage (Christensen et al., 1975; Banks et al., 1997; Lewbel and Pendakur, 2009). Therefore, researchers have turned to characteristic-based demand systems, which write substitution patterns as functions of distances in a low-dimensional characteristic space to remain tractable as the number of products grows (Berry, 1994; Berry et al., 1995; Nevo, 2000). However, characteristic-based demand systems require to observe and choose all relevant product characteristics. Firm-level or product-firm-level panels typically only record prices, quantities, input costs, and firm characteristics, but not full product characteristics vectors needed for the estimation of characteristic-based demand systems. Widely used sources such as India’s Prowess and Annual Survey of Industries, or Chile’s Encuesta Nacional Industrial Anual, illustrate the pattern. Therefore, estimating characteristic-space models is infeasible with typical firm production datasets. The estimation approach developed in this paper aims at resolving the infeasibility of estimating large flexible demand systems using production datasets while focusing on the most prominent representative of product-space demand models, the Almost Ideal Demand System (AIDS) and its linear approximation (LA-AIDS) developed by Deaton and Muellbauer (1980). The approach does not rely on observing product characteristics and instead exploits a lower dimensional approximation of the full cross-price parameter matrix to reduce the growth of the free parameter count in the number of products from quadratic to linear while allowing for fully flexible substitution patterns.

The approach developed in this paper bridges two strands of literature that have largely developed in isolation but speak directly to this problem. I combine techniques designed to determine the number of factors in factor-models (Bai and Ng, 2002) to estimate the number of dominant substitution dimensions with the structural demand system estimation approaches that are based on consumer-theory prominent in industrial organization and agricultural economics (Deaton and Muellbauer, 1980; Berry, 1994; Berry et al., 1995; Nevo, 2000; Fally and Ligon, 2026). I develop an approximate spectral decomposition of the AIDS price coefficient matrix and then apply the estimation approach from Bai and Ng (2002) on a residualized market share expression to extract the rank of the substitution patterns together with a basis of the substitution subspace. I then embed the results in an MPEC-GMM estimator to recover the structural parameters. The resulting estimator allows to identify all parameters of the AIDS with hundreds of products without requiring product characteristics or arbitrary choices of the structure of

the cross-price parameter matrix. The closest antecedent is Fally and Ligon (2026), whose K-aggregator framework provides the structural interpretation for the rank restriction but whose estimator requires rich product attribute data. Instead of the attribute-based projection used in Fally and Ligon (2026), the estimation approach developed in the present paper utilizes a data-driven basis extracted from the residualized expenditure share matrix. The estimator thereby makes estimating flexible product-space demand systems such as the AIDS feasible for large markets from production data alone.

The estimator delivers substantial parameter savings relative to the standard AIDS. With 200 products and 4 dominant substitution dimensions, the total parameter count falls from 20,298 to 1,188, a reduction of 94.1%. From the 1,188 structural parameters, 398 are product-specific intercepts and expenditure coefficients. Only 790 parameters come from the price coefficient matrix. At 500 products with 10 dominant substitution dimensions, the standard system carries over 125,000 parameters while the proposed estimation approach estimates only 5,943, of which 4,945 are cross-price parameters, a total reduction of more than 95%. The reduced parameter count brings estimation well within reach of typical production datasets. While the standard AIDS with 200 products requires at least 201 time periods for identification, the proposed estimation approach requires as few as 12 time periods at the same market size with 10 dominant substitution dimensions, well below the 20 to 50 annual periods that characterize most firm-level panels. Importantly, the number of dominant substitution dimensions is not an arbitrary design choice. The approach rather estimates the number of sample-supported dominant substitution dimensions in a data-driven manner directly from the residualized expenditure share matrix.

I assess the performance of the estimation approach on simulated data in a Monte Carlo study with market sizes ranging from 20 to 500 products over 50 time periods. The first two stages of the estimator recover the correct rank in all simulation runs across all market sizes, including at 500 products, where the ratio of parameters to time periods is the most challenging for the estimator. Given the rank estimate and the recovered latent basis from the first two stages, the GMM-MPEC estimator in the third stage converges in all repetitions and recovers all structural parameters with high precision throughout. Product-specific intercepts α and expenditure coefficients α^D are recovered with small absolute errors throughout, and the implied price coefficient matrix H and Marshallian elasticities exhibit median absolute errors that remain small relative to own-price elasticities across all market sizes. Median relative errors for the substitution eigenvalues remain below 5% across all factors even at 500 products.

The remainder of the paper is organized as follows. Section 2 presents the Almost Ideal Demand System and explains the parameter growth problem. Section 3 develops the approximate rank- r reparameterization of the AIDS. Section 4 develops the estimation approach. Section 5 discusses the identification of structural parameters and latent objects. Section 6 evaluates the performance of the estimator with simulated data. Section 7 analyzes the approach’s relationship to the literature. Section 8 discusses implications before concluding.

2 The Almost Ideal Demand System

The basic unit of observation is a product-period combination (j, t) with $j = 1, \dots, J$ products in period $t = 1, \dots, T$. I use the term product to refer to either a firm producing a homogeneous output or a distinct product variety. The derivations below apply equally to both interpretations. Production datasets

typically record, in each period t the output price vector $P_t^Q = (P_{1,t}^Q, \dots, P_{J,T}^Q)$, a vector of output quantities $Q_t = (Q_{1,t}, \dots, Q_{J,T})$, and some additional variable vectors such as input prices or firm characteristics.

Assumption 1 (Maintained demand model). *The observed expenditure shares are generated by the AIDS share expression given by*

$$w_{j,t} = \alpha_j + \left[H \ln P_t^Q \right]_j + \alpha_j^D RE_t + \delta_{j,t}, \quad (1)$$

where the demand shock $\delta_{j,t}$ satisfies $\mathbb{E}[\delta_{j,t}] = 0$.

The AIDS is defined by its expression for the market share $w_{j,t}$ of product j in period t in Equation (1). The α_j parameter is a product-specific intercept, while the $J \times J$ matrix $H := [\eta_{j,i}]$ includes all own-price effects on its diagonal and all cross-price effects on its off-diagonal. α_j^D is the expenditure coefficient and RE_t represents the logarithm of total real expenditures in period t . Real expenditures in the AIDS are defined as

$$RE_t := \ln \left(\frac{X_t}{\mathcal{P}_t} \right), \quad (2)$$

where $X_t := \sum_{j=1}^J P_{j,t}^Q Q_{j,t}$ is the total nominal expenditure and $\mathcal{P}_t$ the price index defined as

$$\ln \mathcal{P}_t := \alpha^0 + \alpha^\top \ln P_t^Q + \frac{1}{2} (\ln P_t^Q)^\top H \ln P_t^Q. \quad (3)$$

The linear approximate AIDS replaces the exact price index in the real expenditure term of the exact AIDS with the Stone price index defined as $\ln \mathcal{P}_t^S := \sum_i w_{i,t} \ln P_{i,t}^Q$, leading to a linear expression for the output shares. The LA-AIDS is a popular alternative to the exact AIDS as it simplifies the estimation of the demand system substantially while retaining the same approximate structure (Deaton and Muellbauer, 1980). Because output and prices are observed for every product in the observed panel, expenditure shares $w_{j,t}$ can be computed directly from observables as

$$w_{j,t} := \frac{P_{j,t}^Q Q_{j,t}}{X_t}. \quad (4)$$

The shares satisfy $\sum_j w_{j,t} = 1$ by construction over the observed product set in each period.

I further impose the three canonical restrictions of the AIDS system. *Homogeneity* of degree zero in prices and total expenditure lead to $\sum_i \eta_{j,i} = 0$ for each j . *Slutsky symmetry* requires $\eta_{j,i} = \eta_{i,j}$ for all pairs (j, i) . *Adding-up* requires $\sum_j \alpha_j = 1$, $\sum_j \eta_{j,i} = 0$ for each i , and $\sum_j \alpha_j^D = 0$; the cross-price adding-up constraint follows automatically from homogeneity and Slutsky symmetry and is therefore not an independent restriction. Taking the J -th share equation to be determined residually by adding-up, the count of free parameters in the baseline AIDS is¹

$$\underbrace{(J-1)}_{\alpha_j} + \underbrace{\frac{(J-1)J}{2}}_{\eta_{j,i}} + \underbrace{(J-1)}_{\alpha_j^D} = \frac{J^2 + 3J - 4}{2}. \quad (5)$$

which grows quadratically in the number of products. For $J = 10$ products number of parameters is 63, but for 200 products it is already 20,298. The minimum number of time periods to theoretically being

¹If α^0 is not fixed a priori as Deaton and Muellbauer (1980) suggest but estimated, the parameter count increases by one.

able to identify all parameters of the system is $T \geq J + 1$, which for $J = 200$ requires a panel of at least 201 periods.² This requirement exceeds the length of virtually any available production panel, which is predominantly annual and typically spans $T = 20$ to 50 periods. The quadratic parameter growth is driven entirely by the price coefficient matrix H as the intercept and expenditure parameter counts grow only linearly in J . Reducing the dimensionality of H is therefore the central modeling challenge which the estimation approach below addresses.

3 Approximate Rank- r Structure of Cross-Price Parameters

The parameter count of the basic AIDS described above grows quadratically with the number of products making it infeasible to estimate in most empirical settings with more than a handful of products. This fast growth arises entirely from the price coefficient matrix H , whose (j, i) entry encodes the effect of the price of every product i on the share of every product j . In markets with many products, however, it is implausible that every product pair carries an entirely independent cross-price relationship. Consumers rather make purchasing decisions on a number of factors that is often substantially smaller than the number of products in the market (Bresnahan et al., 1997; Hausman and Leonard, 2002). These factors can, for example, contain branding (Erdem and Keane, 1996), advertisement (Akerberg, 2001, 2003), quality (Sudhir, 2001), geographic brand effects (Bronnenberg et al., 2007, 2009), or a number of product characteristics such as weight, size, color, or engine size (Berry et al., 1995; Verboven, 1996). This leads the bulk of product substitution to operate through these substitution dimensions rather than through individual product-by-product channels. Prominent approaches in the industrial organization literature implement this argument by incorporating effects of product characteristics directly into the model (McFadden, 1978; Berry, 1994; Berry et al., 1995; Nevo, 2000; Grigolon and Verboven, 2014). Fally and Ligon (2026) subsume these approaches within a single general demand framework based on price aggregators. However, each of these approaches requires observing and picking relevant product characteristics, which production panels often lack. Yet the underlying substitution structure leaves a pattern in co-movements of expenditure shares that price variation generates. The estimator developed here exploits these patterns to recover the substitution dimensions without observing product characteristics.

In the AIDS, the price coefficient matrix H governs how price changes translate into expenditure-share movements. Therefore, the directions generating those patterns are encoded in H . Slutsky symmetry makes H real symmetric, so it admits the general spectral decomposition

$$H = VDV^\top, \quad V^\top V = \mathbf{I}, \quad V^\top \mathbf{1} = \mathbf{0}. \quad (6)$$

If consumers substitute between products along r dimensions, r eigenvalues are non-zero and H has rank r (Fally and Ligon, 2026).³ The eigenvector matrix V is then $J \times r$ and the diagonal eigenvalue matrix

²Even though the panel has JT observations, with $(J-1)T$ of them usable after adding-up, and only $\frac{(J-1)(J+4)}{2}$ parameters, symmetry of H renders $\frac{(T-2)(T-3)}{2}$ of those rows redundant. Identification requires $(J-1)T - \frac{(T-2)(T-3)}{2} \geq \frac{(J-1)(J+4)}{2}$ which first holds at $T = J + 1$.

³Fally and Ligon (2026) put the low-rank structure on the cross-price coefficient matrix only after separating a full-rank diagonal of own-price effects. The AIDS considered here instead applies the low-rank structure to the complete price coefficient matrix. This does not eliminate own-price effects, a negative slope of demand in own price, or negative own-price elasticities. Every product retains a diagonal coefficient, and these coefficients can be strictly negative and generate

D is $r \times r$ with the eigenvalues ordered by $|d_1| \geq \dots \geq |d_r|$. Its columns form an orthonormal basis for the r -dimensional subspace along which substitution operates. The constraint $V^\top \mathbf{1} = \mathbf{0}$ enforces price homogeneity of the demand system because it implies $VDV^\top \mathbf{1} = \mathbf{0}$. The decomposition has fewer free parameters than H whenever $r < J - 1$, which the estimator exploits to reduce the number of parameters to estimate.

However, in empirical settings the price coefficient matrix will almost never be exactly of rank r because of noise or additional idiosyncratic unsystematic substitution patterns. H rather exhibits r pervasive eigenvalues capturing the dominant substitution dimensions and $J - r$ small eigenvalues capturing residual idiosyncratic variation, with a spectral gap between the two groups. I, therefore, rewrite H as the spectral approximation

$$H = VDV^\top + E, \quad V^\top V = \mathbf{I}, \quad V^\top \mathbf{1} = \mathbf{0}, \quad (7)$$

where the symmetric residual matrix E captures the remaining $J - r$ eigenvalues and their eigenvectors from the noise, while V and D keep all their properties. The product share expression in Equation (1) becomes then

$$w_{j,t} = \alpha_j + \left[VDV^\top \ln P_t^Q \right]_j + \alpha_j^D RE_t + \tilde{\delta}_{j,t}, \quad (8)$$

where the composite error $\tilde{\delta}_{j,t} := \delta_{j,t} + \left[E \ln P_t^Q \right]_j$ collects the idiosyncratic structural shock $\delta_{j,t}$ and the price response attributable to the discarded eigenvalue-eigenvector tail E . Note that since $H\mathbf{1} = \mathbf{0}$ by homogeneity, this forces $E\mathbf{1} = \mathbf{0}$ making the small eigenvalue-eigenvector tail also satisfying the homogeneity constraint.

The rank- r re-parameterization from Equation (8) replaces the $J(J - 1)/2$ free entries of H with imposed symmetry and homogeneity with $Jr - r(r + 1)/2$ free parameters in (V, D) . The decomposed version is strictly more parsimonious whenever $r \leq J - 2$, with growing parsimony gains when the gap between r and J increases. For example, when $J = 200$ and $r = 10$, the standard AIDS estimator carries 19,900 free cross-price entries, while the rank- r parameterization retains only 1,945. What the decomposition does imply is that two products loading identically on the substitution directions have the same cross-price relationship with the other products.

Equation (7) merely partitions a symmetric, homogeneous H into retained and discarded eigendirections and therefore imposes no further assumptions. However, if E is large, H will not be well-approximated by the retained part $VDV^\top$. Furthermore, the contribution of E to expenditure shares must be sufficiently weak for the rank- r spectral approximation to be applicable in an estimator for the model in Equation (8). That contribution enters as the term $\left[E \ln P_t^Q \right]_j$ of the composite error. To characterize the price variation on which E can act, define

$$\mathcal{M} := \mathbf{I} - V(V^\top V)^{-1}V^\top - \mathbf{1}(\mathbf{1}^\top \mathbf{1})^{-1}\mathbf{1}^\top = \mathbf{I} - VV^\top - J^{-1}\mathbf{1}\mathbf{1}^\top \quad (9)$$

as the orthogonal projector that, when applied to a vector, removes its components along the r retained substitution directions and the constant direction. Each observed log-price vector can then be decomposed

economically large own-price elasticities. The maintained assumption is that own- and cross-price effects are organized by the same r systematic substitution dimensions and linked through homogeneity, rather than that own-price effects arise from an additional product-specific block.

as

$$\ln P_t^Q = \underbrace{J^{-1}\mathbf{1}\mathbf{1}^\top \ln P_t^Q}_{\text{common direction}} + \underbrace{VV^\top \ln P_t^Q}_{\text{retained directions}} + \underbrace{\mathcal{M} \ln P_t^Q}_{\text{remainder}}.$$

Premultiplying by E gives the decomposition of the tail's contribution to expenditure shares

$$E \ln P_t^Q = \underbrace{EJ^{-1}\mathbf{1}\mathbf{1}^\top \ln P_t^Q}_0 + \underbrace{EVV^\top \ln P_t^Q}_0 + E\mathcal{M} \ln P_t^Q. \quad (10)$$

The first term vanishes by homogeneity, and the second because the retained and discarded eigendirections are orthogonal. Hence, only the remaining price variation $\mathcal{M} \ln P_t^Q$ can generate a share response through the discarded eigenvalue tail. An estimator of Equation (8) therefore only needs a restriction of the potentially non-zero part in Equation (10) to control its contamination. How much this part contributes to expenditure shares depends on how strongly E transmits the residual variation, and how much residual variation there is to transmit. A small E does not suffice when the residual variation is large, and small residual variation does not suffice when E is large.

The root-mean-square contribution defined as

$$\zeta_{\text{RMS}} := \lambda_{\max} \left(\mathbb{E} \left[\left(E\mathcal{M} \ln P_t^Q \right) \left(E\mathcal{M} \ln P_t^Q \right)^\top \right] \right)^{1/2}, \quad (11)$$

measures exactly the magnitude of the tail's contribution to expenditure shares. Here, $\lambda_{\max}(Y)$ denotes the largest eigenvalue of a matrix Y . The negligible-tail condition below requires ζ_{RMS} to vanish sufficiently fast that the discarded component neither distorts rank selection nor contaminates the estimating moments asymptotically. By contrast, the retained block must generate expenditure-share variation that is pervasive across products, so that its substitution directions can be recovered from the panel. Assumption 2 formalizes the conditions on the retained block and the tail. Define the retained price-response score as $\psi_t := DV^\top \ln P_t^Q$ and to isolate the variation that cannot be absorbed by the intercept and real-expenditure term, let $\psi_t^\perp$ denote the residual from the population linear projection of ψ_t on $(1, RE_t)$.

Assumption 2 (Pervasiveness and negligible tail). *(a) Pervasiveness: $J \text{Var}(\psi_t^\perp) \rightarrow \bar{\Sigma}$ for some positive definite $\bar{\Sigma} \in \mathbb{R}^{r \times r}$ as $J \rightarrow \infty$; the loadings are delocalized, $\max_j \sqrt{J} \|V_{j,\cdot}\| \leq \bar{C}_V$ for a constant $\bar{C}_V$ not depending on J ; and the retained eigenvalues satisfy $0 < \underline{c}_D \leq J|d_r| \leq \dots \leq J|d_1| \leq \bar{C}_D < \infty$ for constants not depending on J .*

(b) Negligible tail: $J\|E\|_2 \rightarrow 0$ and $J^{3/2} \zeta_{\text{RMS}} \rightarrow 0$ as $J \rightarrow \infty$.

Part (a) carries the strong-factor requirement of the approximate factor model literature over to the demand system. The variance condition ensures that every retained price-response score contains variation that cannot be absorbed by real expenditure. Delocalization spreads the corresponding expenditure-share response across a non-vanishing fraction of products, preventing any substitution direction from being driven by a single product. The eigenvalue bounds prevent retained directions from becoming either asymptotically weak or disproportionately strong relative to the average budget share. Together with $J\|E\|_2 \rightarrow 0$ of part (b), the lower bound also separates them from the discarded eigenvalue tail.

Part (b) restricts the tail in two distinct ways. The condition $J\|E\|_2 \rightarrow 0$ makes the discarded eigenvalues negligible relative to the retained eigenvalues because it forces them to vanish faster than the

retained eigenvalues, which ensures that $VDV^\top$ converges to H . The condition $J^{3/2}\zeta_{\text{RMS}} \rightarrow 0$ instead makes the root-mean-square expenditure-share response generated by the tail negligible at the rate required for the estimator.⁴ Neither condition requires E to be sparse or low dimensional. It may remain dense and contain a growing number of nonzero product-specific links.

4 Estimation

The estimation procedure contains three stages. The goal of the first two stages is to recover the dominant substitution subspace from the data, which is the key input for the third stage. The first stage transforms the observed expenditure shares into a form that admits a factor structure with loadings equal to the dominant substitution dimensions. The second stage applies asymptotic principal components, determines the rank via an information criterion, and delivers the loading matrix for the third stage. The third stage formulates the MPEC-GMM problem to estimate the structural parameters: the price coefficient matrix H , the expenditure coefficients α^D , and the intercepts α . Throughout, I maintain the demand model of Assumption 1 and the approximate rank- r structure of Assumption 2. The remaining assumptions are introduced at the stage where they are first needed.

4.1 Stage 1: From Share Equation to Factor Model

First, recognize that the share expression in Equation (1) almost exhibits a factor model structure but includes two additional components that prevent the direct application of methods designed to determine the rank of a factor model, the real expenditure block RE_t and the intercept α_j . I therefore first net out the additional components from the share equation to create a factor model structure of the residualized shares

$$w_{j,t} - \alpha_j - \alpha_j^D RE_t = \left[H \ln P_t^Q \right]_j + \delta_{j,t}. \quad (12)$$

However, as α_j and α_j^D are unknown parameters, whose values need to be estimated first. Regressing product shares on real expenditures and a constant yields such estimates. The first stage estimation equation is therefore

$$w_{j,t} = \alpha_j + \alpha_j^D RE_t^S + \varepsilon_{j,t}, \quad (13)$$

with $\varepsilon_{j,t} := \left[H \ln P_t^Q \right]_j + \delta_{j,t}$. This expression uses the Stone price index to construct real expenditure $RE_t^S := \ln(X_t/P_t^S)$, as it is computable from observables. Under LA-AIDS, RE_t^S coincides with the structural RE_t . Under exact AIDS, RE_t^S serves as an observable proxy for the unobserved $RE_t^{\text{exact}} := \ln\left(\frac{X_t}{P_t}\right)$, which depends on parameters estimated only in Stage 3.

The estimation equation in Equation (13) contains two problems. First, RE_t is an endogenous variable as it is a function of equilibrium prices and quantities, which respond to demand shocks $\delta_{j,t}$. Simply

⁴The two rates reflect two ways the tail enters. The estimating equations see it one direction at a time and require only $J\zeta_{\text{RMS}} \rightarrow 0$. Rank selection instead sees it through its total contribution across products, $\text{tr}(\text{Var}(E\mathcal{M} \ln P_t^Q)) \leq J\zeta_{\text{RMS}}^2$, which inflates the $(r+1)$ -th eigenvalue of the residualized second-moment matrix. That contribution must stay below the per-product noise variance of order $1/J^2$ against which the criterion calibrates its penalty, and $J\zeta_{\text{RMS}}^2 < 1/J^2$ is exactly $J^{3/2}\zeta_{\text{RMS}} \rightarrow 0$. The exponent is therefore set by the asymmetry between the two sides of the comparison, the tail entering summed across products and the noise floor product by product, and the rate it delivers is sufficient for consistent selection rather than necessary.

regressing $w_{j,t}$ on a constant and RE_t by OLS without taking the endogeneity into account and taking residuals based on these estimates would therefore generate spurious cross-sectional dependence in the residualized expenditure shares with the potential of biasing the subsequent rank selection. Second, the regression omits the variable $\left[H \ln P_t^Q\right]_j$ which is correlated with both $w_{j,t}$ and RE_t , introducing an additional bias.

Fortunately, using valid instruments for RE_t circumvents the first issue. Any valid instrument set has to satisfy two distinct conditions, exogeneity with respect to the demand shock $\delta_{j,t}$ and relevance for the regressors it has to move.

Assumption 3 (Instrument exogeneity). *The raw instrument vectors $Z_t^{\text{raw}} = (Z_t^{\text{raw},1}, \dots, Z_t^{\text{raw},M})$, with $M \geq 1$ and each $Z_t^{\text{raw},m} \in \mathbb{R}^J$ collecting a product-level time-varying instrument, satisfy $\mathbb{E}[\delta_{j,t} \mid Z_t^{\text{raw}}] = 0$ for all j and t .*

Assumption 3 states the exogeneity condition every valid instrument set must satisfy.⁵ Firm's input prices are the canonical choice in the literature, making the implicit assumption that individual product demand realizations do not affect input markets. How many raw instruments are needed depends on how instruments in Stage 3 get generated, which Section 4.3.2 discusses. The first stage itself needs only one exogenous instrument for the only endogenous variable RE_t . However, valid first-stage instruments require in addition that these instruments' shifts affect real expenditure. The raw instrument set Z_t^{raw} might not satisfy this condition, because product-specific instruments may not have a sufficient impact on aggregate real expenditure in the market. However, functions of the raw instruments, such as averages over all products for a given period, can shift real expenditure sufficiently to satisfy the relevance condition. Therefore, adding a constant to $L \geq 1$ functions of the raw instruments the instrument vector for each period in the first stage is

$$Z_t^{\text{FS}} := (1, Z_{1,t}^{\text{FS}}, \dots, Z_{L,t}^{\text{FS}})^{\top}, \quad (14)$$

with each $Z_{\ell,t}^{\text{FS}}$ being a measurable function aggregating Z_t^{raw} . Assumption 4 below formally states the relevance condition for the first-stage instruments.

Assumption 4 (First-stage relevance). *The first-stage instruments are relevant for real expenditure. The matrix $\mathbb{E}[(1, RE_t)^{\top} (Z_t^{\text{FS}})^{\top}]$ has rank 2.*

Under Assumptions 3 and 4 I estimate (α_j, α_j^D) by 2SLS, denoting the first-stage estimates as $(\hat{\alpha}_j^{\text{FS}}, \hat{\alpha}_j^{D,\text{FS}})$. However, the second issue, the presence of the omitted variables, stated above biases the resulting estimates. Fortunately, the bias does not affect the rank of the residualized shares defined as

$$\tilde{w}_{j,t} = w_{j,t} - \hat{\alpha}_j^{\text{FS}} - \hat{\alpha}_j^{D,\text{FS}} RE_t^S. \quad (15)$$

To see why, take the population probability limit of the estimated coefficients. Substituting the structural form of Equation (1) into the regression projection of $w_{j,t}$ on $(1, RE_t^S)$ with instruments Z_t^{FS} , and using $\mathbb{E}[\delta_{j,t} \mid Z_t^{\text{FS}}] = 0$ to eliminate the structural-shock term, gives

$$\hat{\alpha}_j^{D,\text{FS}} \xrightarrow{p} \alpha_j^D + V_{j,\cdot}^{\top} \mu + e_j, \quad \hat{\alpha}_j^{\text{FS}} \xrightarrow{p} \alpha_j + V_{j,\cdot}^{\top} \nu + f_j, \quad (16)$$

⁵Weaker orthogonality would suffice if we restricted any instrument generated from the raw instrument set to be a linear function of the raw instruments.

with $\mu := DV^\top \text{Cov}(\ln P_t^Q, \widehat{RE}_t^S) / \text{Var}(\widehat{RE}_t^S) \in \mathbb{R}^r$, $\nu := \mathbb{E}[DV^\top \ln P_t^Q] - \mu \mathbb{E}[RE_t^S] \in \mathbb{R}^r$, and $V_{j,\cdot}$ denoting the j -th row of V . The scalars $e_j := E_{j,\cdot} \text{Cov}(\ln P_t^Q, \widehat{RE}_t^S) / \text{Var}(\widehat{RE}_t^S)$ and $f_j := E_{j,\cdot} \mathbb{E}[\ln P_t^Q] - e_j \mathbb{E}[RE_t^S]$ are the contributions of the discarded eigenvalue tail. They are exactly zero when H has exact rank r , and under the approximate model they vanish asymptotically because of Assumption 2(b). The derivation shows that the dominant bias is not idiosyncratic and that it almost exclusively loads on the product-specific row $V_{j,\cdot}^\top$ of the structural loading matrix. The only part that escapes the substitution subspace is the tail pair (e_j, f_j) , which is uniformly asymptotically vanishing. Substituting Equation (16) into Equation (15) then yields

$$\tilde{w}_{j,t} = V_{j,\cdot}^\top \tilde{\psi}_t + u_{j,t}, \text{ with } \tilde{\psi}_t := \psi_t - \nu - \mu RE_t^S, \quad (17)$$

where $\psi_t = DV^\top \ln P_t^Q$ is the retained price-response score. The combined error term $u_{j,t} := \tilde{\delta}_{j,t} - (f_j + e_j RE_t^S) - (\hat{\alpha}_j^{\text{FS}} - \text{plim } \hat{\alpha}_j^{\text{FS}}) - (\hat{\alpha}_j^{D,\text{FS}} - \text{plim } \hat{\alpha}_j^{D,\text{FS}}) RE_t^S$ collects the idiosyncratic error $\tilde{\delta}_{j,t}$ from Equation (8), the tail pair (e_j, f_j) , and the finite-sample error of the first-stage coefficients which vanishes in probability as $T \rightarrow \infty$. Equation (17) is an approximate factor model with the same loading matrix V as the product share expression in Equation (8) and the shifted price-response score $\tilde{\psi}_t$. The dominant bias factors through V and is absorbed into the shifted score $\tilde{\psi}_t$. The idiosyncratic error $\tilde{\delta}_{j,t}$ passes through the residualization unaltered up to the tail component $-(f_j + e_j RE_t^S)$, which Assumption 2(b) keeps too small to register as an additional pervasive factor. The estimation procedure therefore preserves the approximate rank- r structure.

One might consider estimating Equation (15) by OLS instead of 2SLS arguing that the estimates are biased with either method. However, this can break the structure. The OLS analogue of Equation (16) adds a $\frac{\text{Cov}(\delta_{j,t}, RE_t^S)}{\text{Var}(RE_t^S)}$ term to $\hat{\alpha}_j^{D,\text{OLS}}$, which is nonzero and which, unlike the omitted-variable bias, does not in general factor through $V_{j,\cdot}^\top$. The OLS-residualized shares therefore carry an additional component making the factor model not generally rank r . The 2SLS step neutralizes this extra component by replacing RE_t^S with $\widehat{RE}_t^S$, for which $\text{Cov}(\delta_{j,t}, \widehat{RE}_t^S) = 0$ if the instrument vector is valid.

4.2 Stage 2: Selecting the Rank and Estimating the Factors

The targets in the second stage are the rank of the price coefficient matrix and an estimate for the latent basis. The procedure delivers a factor-model-based rank estimate from the residualized expenditure share matrix from Equation (15) and the corresponding latent basis from the factor loadings for the estimation of the structural parameters in Stage 3. The estimate for the rank is consistent under the approximate-factor-model conditions Bai and Ng (2002) describe and the latent basis estimate recovers V up to rotation.

Equation (17) takes the form of an approximate factor model in the sense of Chamberlain and Rothschild (1983) as it expresses $\tilde{w}_{j,t}$ as a linear combination of r factors $\tilde{\psi}_t$ with loading matrix V plus an idiosyncratic but cross-sectionally correlated component $u_{j,t}$. Therefore, the information criterion developed in Bai and Ng (2002) estimates the number of factors.⁶ As discussed above, the number of pervasive

⁶The error term of the residualized shares does not have exactly zero mean, as assumed by Bai and Ng (2002), because of the finite sample error from Stage 1's estimation. However, rank selection does not need an exact zero-mean error, because the criterion only evaluates how much each additional factor reduces the unexplained variation in the panel. The Stage-1 estimation error does not contribute enough to that variation for the criterion to register an additional factor and its contribution shrinks with the number of periods. The mean of the tail contamination through $\tilde{\delta}_{j,t}$ gets absorbed in the

eigenvalues of $\tilde{w}_{j,t}$ equals the number of pervasive eigenvalues of H , which is the rank r of the price coefficient matrix H .

To select the best fitting factor, Bai and Ng (2002) propose to estimate the best k -factor approximation to $\tilde{w}_{j,t}$ via asymptotic principal components for each candidate rank k defined as

$$\min_{\Lambda, F_1, \dots, F_T} \sum_{j=1}^J \sum_{t=1}^T \left(\tilde{w}_{j,t} - \Lambda_{j,\cdot}^\top F_t \right)^2, \quad \text{s.t. } \Lambda^\top \Lambda = \mathbf{I}_k, \quad (18)$$

where Λ collects the estimated loadings and F_t the estimated factor scores at time t , neither of which is observed. After computing the asymptotic principal components, Bai and Ng (2002) select the rank $\hat{r}$ as the minimizer of the proposed criterion⁷

$$\text{IC}(k) = \ln(\hat{\sigma}_u^2(k)) + k \left(\frac{J+T}{JT} \right) \ln(\min(J, T)). \quad (19)$$

The quantity $\hat{\sigma}_u^2(k) := (JT)^{-1} \sum_{j=1}^J \sum_{t=1}^T \hat{u}_{j,t}(k)^2$ is the average squared residual from the asymptotic principal components in Equation (18), where $\hat{u}_{j,t}(k)$ denotes the fit residual. The asymptotic principal components estimator in Equation (18) amounts to eigendecomposing the cross-product matrix $T^{-1} \tilde{w} \tilde{w}^\top$, a $J \times J$ matrix whose eigenvalues reflect the factor structure of the expenditure shares. Under Assumption 2, the r eigenvalues of this matrix that correspond to the structural substitution directions and the remaining $J - r$ arising from idiosyncratic demand shocks and the tail diverge by a ratio that grows with J . This widening gap is the approximate factor structure that allows the information criterion to consistently separate the r structural factors from noise.

Under exact AIDS, the residualization based on the Stone price index in Stage 1 introduces one additional factor into the panel $\tilde{w}_{j,t}$ beyond the r structural substitution dimensions. The spurious factor has loading α^D and its score is the component of the exact-versus-Stone deflator discrepancy that survives residualization. When this discrepancy has large enough time-series variation to register as a separate factor in the residualized expenditure shares, the Bai–Ng criterion selects rank $\hat{r} = r + 1$ rather than r . This potential overestimation of the rank by one is benign for Stage 3. Overestimating the rank over-parameterizes the Stage 3 rotated eigenvalue matrix Ξ rather than misspecifies it. Stage 2 recovers all r structural directions, so the pervasive part $VDV^\top$ of the true price coefficient matrix H remains representable as $\hat{\Lambda} \Xi \hat{\Lambda}^\top$ even when the selected rank is $r + 1$. Stage 3 remains consistent for H , α , and α^D , while the extra free parameters inflate the finite-sample variance of $\hat{H}$ without biasing the point estimates.

Stage 2 not only provides an estimate for the rank of the price coefficient matrix, $\hat{r}$, but also the corresponding loadings $\hat{\Lambda}$ and factor scores $\hat{F}_t$ from asymptotic principal components. $\hat{\Lambda}$ is a consistent estimate of V up to an orthogonal rotation, meaning that $\hat{\Lambda} \xrightarrow{P} V\Phi$ for some orthogonal Φ . The rotation arises because the factor model $\tilde{w}_{j,t} = V_{j,\cdot}^\top \tilde{\psi}_t + u_{j,t}$ is observationally equivalent to its rotated version $\tilde{w}_{j,t} = (V\Phi)_{j,\cdot}^\top (\Phi^\top \tilde{\psi}_t) + u_{j,t}$ for any orthogonal Φ , as redefining loadings as $V\Phi$ and factor scores as $\Phi^\top \tilde{\psi}_t$ leaves both the fitted values and the orthonormality constraint $\Lambda^\top \Lambda = \mathbf{I}$ unchanged. However,

construction of $u_{j,t}$.

⁷Bai and Ng (2002) propose a class of appropriate criteria while focusing on three examples. In their simulations, the first and second of the three criteria show the strongest finite sample performance. From those two I choose the second, however, replacing the information criterion with another one of the proposed criteria does not change the general result.

note that the factorization of H in Equation (7) is not unique but rotationally invariant. This means that, even though Stage 2 cannot identify V itself, Stage 3 can use $\hat{\Lambda}$ to construct $VDV^\top$ as $VDV^\top = (V\Phi)(\Phi^\top D\Phi)(V\Phi)^\top = \hat{\Lambda}\Xi\hat{\Lambda}^\top$. Stage 3 then only has to find Ξ that satisfies the structural constraints given $\hat{\Lambda}$ instead of estimating both V and D in Stage 3. Appendix A1 develops a version of the estimator that does not use $\hat{\Lambda}$ from Stage 2 but treats V as an unknown to be estimated in Stage 3. The key advantage of using $\hat{\Lambda}$ in Stage 3 is that the criterion function in the Stage 3 estimator is globally convex and has substantially fewer unknowns reducing the complexity of the MPEC problem while speeding up computation. However, both versions deliver consistent structural estimates.

The shifted scores $\tilde{\psi}_t$ and the demand shocks δ_t are generally correlated, as prices are equilibrium outcomes responding to demand shocks. This correlation is generally harmless for asymptotic principal components unless the shocks of a non-vanishing fraction of products drive the price variation along the same substitution direction. The shock structure then contaminates the recovered basis. Consistency of the estimator therefore requires the restriction on the spread of the correlation across the cross-section of products stated in Assumption 5.

Assumption 5 (Cross-sectional dependence of demand shocks). *For each $k = 1, \dots, r$, the shifted price-response scores satisfy*

$$\sum_{j=1}^J \text{Corr}(\tilde{\psi}_{k,t}, \delta_{j,t})^2 \leq \bar{C}_\rho$$

for a constant $\bar{C}_\rho$ not depending on J .

The assumption holds comfortably when demand shifts are idiosyncratic across products and even coordinated shock patterns of a substantive number of products are permitted. Only shock patterns that aligned with the lines that organize substitution for a non-vanishing number of products violate the assumption.

4.3 Stage 3: Estimating the Structural Parameters

Stages 1 and 2 have already delivered estimates for the rank r and the latent basis Λ . Stage 3 uses these estimates to identify the rest of the unknowns Ξ , α^D , and α in a constrained GMM framework built into a GMM-MPEC estimator. Estimates for Ξ and Λ recover the full price coefficient matrix H up to the negligible eigenvalue tail E .⁸

4.3.1 Moment Conditions

Stage 3 must identify the remaining $2(J-1) + \frac{r(r+1)}{2}$ parameters $\theta = (\alpha, \alpha^D, \Xi)$ and therefore needs at least the same amount of moment conditions. With an appropriate Stage 3 instrument vector $Z_{j,t}^{TS} = (Z_{j,t}^1, \dots, Z_{j,t}^q)^\top$ for each equation j , the Stage 3 moment conditions are

$$\mathbb{E}[Z_{j,t}^{TS} \delta_{j,t}] = \mathbf{0}, \quad j = 1, \dots, J-1, \quad (20)$$

⁸The entries of $\hat{V}$ and $\hat{D}$ can be recovered post estimation from the eigendecomposition of $\hat{\Xi}$, up to the ordering of equal-magnitude entries. Writing $\hat{\Xi} = \hat{Q}\hat{D}\hat{Q}^\top$ with $\hat{Q}$ orthogonal gives $\hat{\Lambda}\hat{\Xi}\hat{\Lambda}^\top = (\hat{\Lambda}\hat{Q})\hat{D}(\hat{\Lambda}\hat{Q})^\top$, so $\hat{D}$ collects the eigenvalues of $\hat{\Xi}$ while $\hat{V} = \hat{\Lambda}\hat{Q}$.

which holds under the exogeneity of the instruments. However, the estimation procedure cannot separate the structural demand shock $\delta_{j,t}$ from the tail term $[E \ln P_t^Q]_j$. The empirical moments therefore have to utilize the rank- r residual,

$$\tilde{\delta}_{j,t}(\theta) = w_{j,t} - \alpha_j - \alpha_j^D RE_t - [\hat{\Lambda} \Xi \hat{\Lambda}^\top \ln P_t^Q]_j. \quad (21)$$

Therefore, even at the true parameters and population basis the residual still contains the tail, $\tilde{\delta}_{j,t}(\theta_0) = \delta_{j,t} + [E \ln P_t^Q]_j$, leaving the population counterpart of the feasible moment not exactly zero,

$$\mathbb{E}[Z_{j,t}^{TS} \tilde{\delta}_{j,t}(\theta_0)] = \underbrace{\mathbb{E}[Z_{j,t}^{TS} \delta_{j,t}]}_{=0} + \mathbb{E}[Z_{j,t}^{TS} [E \ln P_t^Q]_j].$$

The surviving term is the tail's contribution to the moments which Assumption 2(b) forces to vanish asymptotically. Therefore, the feasible moments coincide asymptotically with Equation (20) and consistency of the structural estimator rests on the same condition on the tail-response as Stage 2. Exogeneity of the instruments is hence only required with respect to the structural shock, as in Assumption 3, and not with respect to the tail.

Averaging each equation's moments over periods, and stacking the first $J - 1$ equations while dropping the J -th to enforce adding-up, gives the empirical sample moment vector

$$g(\theta) = (g_1(\theta)^\top, \dots, g_{J-1}(\theta)^\top)^\top, \quad g_j(\theta) := \frac{1}{T} \sum_{t=1}^T Z_{j,t}^{TS} \tilde{\delta}_{j,t}(\theta). \quad (22)$$

Each equation contributes q moment conditions. Therefore, $g(\theta)$ collects $q(J - 1)$ moment conditions, giving at least as many moments as parameters whenever $q \geq 2 + \frac{r(r+1)}{2(J-1)}$.

4.3.2 Instruments

The terms multiplying the parameters, the constant, real expenditure RE_t , and the r components of $\hat{\Lambda}^\top \ln P_t^Q$, carry no product index, so one instrument vector serves all $J - 1$ equations, therefore $Z_{j,t}^{TS} = Z_t^{TS}$. Equation j then provides q moment conditions and carries $r + 2$ parameters, the intercept α_j , the expenditure coefficient α_j^D , and the r components of the row $\hat{\Lambda}_{j,\cdot}^\top \Xi$. Because the same instruments enter every equation, they have to make $r + 2$ of those q conditions linearly independent such that they pin down all $r + 2$ parameters.

The following assumption formalizes the relevance condition that any such instrument vector has to satisfy.

Assumption 6 (Third-stage relevance). *Let Λ be the population loading matrix that Stage 2 estimates. Let $Z_t^0 \in \mathbb{R}^q$ be a Stage-3 instrument vector constructed from the raw instruments and Λ . For some, equivalently every, such Λ , the matrix $\mathbb{E}[Z_t^0 (1, RE_t, (\Lambda^\top \ln P_t^Q)^\top)] \in \mathbb{R}^{q \times (r+2)}$ has full column rank $r + 2$.*

Assumption 6 requires the instruments to generate variation that separates the constant, real expenditure, and all r projected prices from one another. Full column rank requires $q \geq r + 2$, turning the necessary minimum for q from Section 4.3.1 into the sufficient condition for identification.

RE_t and the price block are correlated with the structural error $\delta_{j,t}$ and therefore cannot serve as their own instruments. Both endogenous blocks are aggregates over products, which can leave instruments varying over both j and t individually weak. Real expenditure aggregates every product into one index, while the projected prices aggregate prices into r indices, one for each column of $\hat{\Lambda}$. In large markets, it is unlikely that any product carries a meaningful share of either sum. A raw instrument varying at the level of an individual product therefore barely moves the block it is meant to influence. Stacking the MJ raw product-level instruments hence likely yields many individually weak instruments. Aggregating the raw instruments over products using the same weights that define the regressor each is meant to move can alleviate the problem. Aggregating with equal weights delivers one instrument matched to real expenditure, which is exactly the Stage-1 construction of Equation (13), and aggregating with the r columns of $\hat{\Lambda}$ delivers r instruments matched to the projected prices. Firms' input prices are the canonical raw instruments in this setting, though the argument requires only that they satisfy Assumption 3. Aggregating raw instruments through the latent basis matches them to the endogenous directions but does not by itself make them relevant. For the instruments to be relevant for the cross-price block, their effect must pass through into output prices along each of the r substitution directions. Price variation in any other direction is useless here, because it carries no information about the substitution parameters in Ξ .

Two constructions of Z_t^{TS} supply the required $r+2$ components. The first projects all raw instruments through the estimated basis,

$$Z_t^{TS} = \left(1, \hat{\Lambda}^\top Z_t^{\text{raw},1}, \dots, \hat{\Lambda}^\top Z_t^{\text{raw},M} \right)^\top, \quad (23)$$

giving $q = Mr + 1$ components. However, this means the estimator needs $M \geq 2$, i.e. two or more raw instruments. The second aggregates the instruments for the real-expenditure block in the same way as in Stage 1, and projects the instruments for the cross-price block through the latent basis, giving

$$Z_t^{TS} = \left(1, \hat{\Lambda}^\top Z_t^{\text{raw},1}, Z_{\ell,t}^{\text{FS}} \right)^\top. \quad (24)$$

This construction can be done with a single raw instrument as the two constructions extract different directions of variation in the raw instrument. The instrument construction in Equation (23) leaves $Mr + 1 - (r + 2)$ spare instruments in the rank condition, so some components may be irrelevant without breaking identification, whereas Equation (24) meets the condition exactly and requires every instrument to be relevant. Both leave the moment system over-identified, because each instrument generates $J - 1$ moment conditions while Ξ carries only $r(r + 1)/2$ parameters.

The projection through the latent basis parallels the differentiation IVs of Gandhi and Houde (2019) and the price-weighted attribute aggregates of Fally and Ligon (2026) in spirit, as both compress variation onto a low-dimensional space matched to the substitution structure, but uses the data-driven latent basis $\hat{\Lambda}$ in place of weights provided by observed product characteristics. Both constructions replace Λ by its estimate $\hat{\Lambda}$ while Assumption 6 is stated at the population loading matrix. However, Assumption 6 transfers to the sample instruments Z_t^{TS} with probability approaching one by the Stage 2 consistency of $\hat{\Lambda}$.

4.3.3 MPEC Formulation

I formulate the GMM problem as a mathematical program with equilibrium constraints (MPEC) (Su and Judd, 2012). The MPEC minimizes the GMM criterion function subject to the demand model, imposed as explicit constraints. The constraints contain definitions of the relevant endogenous variables together with restrictions among the parameters and variables. The resulting constrained problem treats all unknown parameters (α, α^D, Ξ) and endogenous variables $\tilde{\delta} = (\tilde{\delta}_{1,1}, \dots, \tilde{\delta}_{J-1,T})$ as explicit decision variables. Writing the problem as an MPEC exposes structure that a substituted criterion function would hide. Modern constrained solvers exploit this structure to speed up convergence, reading the sparsity of the Jacobian and Hessian from the explicit constraints, eliminating linearly dependent constraints, preconditioning the resulting KKT systems, and automatically detecting the convexity of the criterion.

With $\hat{\Lambda}$ fixed at its Stage 2 value and an appropriate weighting matrix W , the GMM-MPEC estimation problem is then

$$\begin{aligned}
& \arg \min_{\theta, \tilde{\delta}} g(\theta)^\top W g(\theta) \\
& \text{s.t.} \quad \text{(I)} \quad \tilde{\delta}_{j,t} = w_{j,t} - \alpha_j - \alpha_j^D RE_t - \left[\hat{\Lambda} \Xi \hat{\Lambda}^\top \ln P_t^Q \right]_j, \quad 1 \leq j \leq J-1, \forall t, \\
& \quad \text{(II)} \quad \Xi = \Xi^\top, \\
& \quad \text{(III)} \quad \alpha_J = 1 - \sum_{j=1}^{J-1} \alpha_j, \\
& \quad \text{(IV)} \quad \alpha_J^D = - \sum_{j=1}^{J-1} \alpha_j^D.
\end{aligned} \tag{25}$$

Constraint (I) defines the rank- r structural residuals $\tilde{\delta}_{j,t}$ from the AIDS share equation evaluated at the observed shares $w_{j,t}$, the natural logarithm of the product price vector P_t^Q , real expenditure RE_t , and candidate parameter values. The symmetry constraint (II) on Ξ is the only restriction on the cross-price block and is what imposes Slutsky symmetry on H . AIDS requires $H = H^\top$, and with orthonormal $\hat{\Lambda}$ the parameterization $H = \hat{\Lambda} \Xi \hat{\Lambda}^\top$ is symmetric if and only if $\Xi = \Xi^\top$. Therefore, the full $J \times J$ symmetry requirement collapses to symmetry of the small $r \times r$ matrix Ξ . Homogeneity, by contrast, needs no constraint at all, because it follows automatically from $\hat{\Lambda}^\top \mathbf{1} = \mathbf{0}$, which Stage 2 already builds into the loadings. Constraints (III) and (IV) impose adding-up on the intercepts and expenditure coefficients.

Under exact AIDS, the logarithm of real expenditure RE_t is not observed in the data because it is built from the exact price index $\mathcal{P}_t^{\text{exact}}$ which depends on unknowns of the MPEC. Therefore, for exact AIDS, the GMM-MPEC in Expression (25) needs to be appended with the additional constraint for the exact price index $\mathcal{P}_t^{\text{exact}}$ and the logarithm of real expenditure RE_t ,

$$\begin{aligned}
\text{(II)} \quad \ln \mathcal{P}_t^{\text{exact}} &= \alpha^0 + \alpha^\top \ln P_t^Q + \frac{1}{2} (\ln P_t^Q)^\top \hat{\Lambda} \Xi \hat{\Lambda}^\top \ln P_t^Q, \quad \forall t, \\
\text{(II)} \quad RE_t &= \ln X_t - \ln \mathcal{P}_t^{\text{exact}}, \quad \forall t.
\end{aligned} \tag{26}$$

Normalizing $\alpha^0 = 0$ follows Deaton and Muellbauer (1980) to pin the location of the translog intercept, which Deaton and Muellbauer (1980) states is only weakly identified. The deflator constraint remains linear in (α, Ξ) , because Ξ enters only as the matrix of a quadratic form in log prices. Constraint (I), however, now contains the bilinear product $\alpha_j^D \ln \mathcal{P}_t^{\text{exact}}$ of two unknowns, which breaks the global convexity of the criterion of the LA-AIDS MPEC. Starting values close to the true parameters therefore gain importance,

which the solution of the LA-AIDS MPEC in (25) could provide. All other aspects of the MPEC such as the instruments, the moment conditions, and the symmetry and adding-up constraints, carry over unchanged across both specifications.

The substitution block Ξ contains $r(r+1)/2$ free parameters, leading to a total free parameter count in Stage 3 of the estimator of $2(J-1) + r(r+1)/2$, which is constant in J in the cross-price block and grows only linearly in J through the product-specific intercepts and expenditure coefficients. With $J = 500$ and $r = 5$ this gives $998 + 15 = 1,013$ free parameters, of which 998 are product-specific intercepts and expenditure coefficients and only 15 come from the cross-price block. The corresponding free parameter count of the estimator developed in Appendix A1, which treats V as an unknown matrix in Stage 3, is $2(J-1) + Jr - r(r+1)/2$. Even though the number of parameters in the cross-price block grows linearly with J , the total parameter count in Stage 3 is substantially higher. At $J = 500$ and $r = 5$ Stage 3 contains $998 + 2,485 = 3,483$ free parameters putting a substantially higher computational burden on the MPEC solver than the projected estimator.

Favourable starting values for Ξ can be obtained from the sample variances of the factor scores $\hat{F}_{k,t}$ returned by Stage 2. However, they should be sign-corrected to start from a negative semidefinite substitution block, i.e. $\Xi_{k,k}^0 = -\widehat{\text{Var}}(\hat{F}_{k,t})$. Starting values for α and α^D can be the Stage 1 coefficients $(\hat{\alpha}_j^{\text{FS}}, \hat{\alpha}_j^{D,\text{FS}})$.

5 Identification

The proposed estimator identifies every entry of the price coefficient matrix without estimating the individual entries directly. Under symmetry and homogeneity H carries $J(J-1)/2$ distinct cross-price parameters. Treating each of those parameters as a separate unknown leads to the infeasible quadratic parameter growth in J as the data has to supply identifying variation for every one. The approximate rank- r parameterization instead writes each entry as $\eta_{j,i} = V_{j,\cdot}^\top D V_{i,\cdot} + E_{j,i}$, combining the two products' loadings on the r substitution directions through a common $r \times r$ diagonal D . Each cross-price effect is therefore implied by how the two products load on the substitution directions together with the r eigenvalues weighting those directions. The number of directions itself is not assumed but estimated from the data in Stage 2.

5.1 Identification of the Structural Parameters

Identification is sequential, with each stage identifying an object the next stage needs. Stage 1 eliminates the intercept and the expenditure term from of the expenditure share expression, Stage 2 determines the rank and the substitution subspace from what remains, and Stage 3 uses that subspace to identify the structural parameters from the moment conditions in Equation (20).⁹

In AIDS the cross-price effects enter the share expression of Equation (1) linearly through $[H \ln P_t^Q]_j$, which the rank- r structure of Equation (7) reduces to $V_{j,\cdot}^\top D V^\top \ln P_t^Q$ up to the tail.¹⁰ Stage 1 nets out the two additional terms, the intercept α_j and the real-expenditure term $\alpha_j^D RE_t$, through the estimation

⁹Appendix A1 derives a version of the estimator in which the basis V gets directly identified in Stage 3 and $\hat{\Lambda}$ from Stage 2 only supplies starting values.

¹⁰Under exact AIDS the cross-price parameters additionally enter indirectly through the price index in RE_t^{exact} . Stage 1 residualizes with the Stone index in either case and is therefore unchanged.

of Equation (13). Each product's intercept and expenditure coefficient appear in that product's equation only, allowing both to be identified from each product's own time series alone. Therefore, Stage 1 estimates them in J 2SLS time-series regressions.

Two biases arise in each of these regressions, but only one of them needs to be addressed. First, RE_t is correlated with the structural error $\delta_{j,t}$, a bias instruments fulfilling Assumptions 3 and 4 remove. Second, the regression omits the cross-price block $V_{j,\cdot}^\top DV^\top \ln P_t^Q$, which introduces a bias that the instruments cannot remove, because any instrument relevant for RE_t also moves output prices and therefore leaves μ in the probability limit of Equation (16) nonzero. However, the second bias is harmless for the residualization because of the direction it takes. Being a linear function of $V_{j,\cdot}$, it enters the residualized panel as the shift of the scores from ψ_t to $\tilde{\psi}_t$ in Equation (17) and leaves the loadings untouched. Absorbing the bias in the scores is harmless because neither Stage 2 nor Stage 3 use the scores. What survives is the tail pair (e_j, f_j) of Equation (16), which is too small to register as a spurious pervasive direction and vanishes asymptotically. A bias that did not factor through $V_{j,\cdot}$, such as the one eliminated by instruments, would instead introduce a spurious direction, which inflates the rank if it were pervasive. Weak instruments, retaining too little of the time-series variation in RE_t , can likewise bias $\hat{r}$ upward. The biased Stage 1 coefficients are not needed beyond the residualization, apart from supplying starting values for Stage 3.

Stage 2 identifies the rank r and the substitution basis from the residualized panel Stage 1 delivers. Once Stage 1 nets out the intercept and the real-expenditure term, what remains in $\tilde{w}_{j,t}$ apart from an error term loads on the r substitution directions through the shifted score $\tilde{\psi}_t$ of Equation (17), an approximate factor model with $J \times r$ loading matrix V and r factors. Under Assumption 2, the eigenvalues of the population covariance of $\tilde{w}$ therefore split into r pervasive ones and a noise-driven remainder, with the gap between the two widening as the number of products grows. The rank is the number of eigenvalues on the pervasive side of that gap, which the Bai and Ng (2002) information criterion recovers consistently as J and T grow. Instead of only exploiting time-series variation as Stage 1 does, Stage 2 draws on the co-movement of products. Two products that load on the same substitution direction have residualized shares that co-move over time, and the cross-product covariance of $\tilde{w}$ records these patterns. The leading r eigenvectors of that covariance span the column space of V , and asymptotic principal components return a basis $\hat{\Lambda}$ for it that differs from V by an orthogonal $r \times r$ rotation, $\hat{\Lambda} = V\Phi$. The rotation is harmless for Stage 3 because the free symmetric block Ξ absorbs it exactly. Rank selection sets the binding requirement for the tail's contribution to expenditure shares as the criterion evaluates the sum of all discarded eigenvalues, therefore aggregates that contribution over products and evaluates it against the noise floor. Stages 1 and 3 impose a weaker requirement on that contribution, but they are the ones that need the first part of Assumption 2(b) for consistency. The factor scores returned alongside $\hat{\Lambda}$ play no structural role and serve only as starting values for Ξ in Stage 3.

Based on the estimates of r and Λ from Stage 2, Stage 3 point-identifies all structural parameters of Equation (8) under Assumptions 2–6 from the moment conditions of Equation (20). The intercepts and expenditure coefficients are identified separately equation by equation through time-series variation. Ξ , however, is common to all $J - 1$ equations and is therefore jointly identified from time-series and cross-sectional variation. Real expenditure RE_t and the r projected prices $\hat{\Lambda}^\top \ln P_t^Q$ are correlated with the demand shock $\delta_{j,t}$ and therefore cannot instrument themselves in Stage 3 moments. Any instrument set therefore needs external instruments for the two blocks that satisfy exogeneity with respect to the

demand shock and relevance for the terms it has to move. The constant in the Stage 3 instrument set Z_t^{TS} identifies the intercepts α_j , leaving the remaining components to cover the expenditure coefficients α_j^D and the block Ξ . The two Z_t^{TS} constructions from Section 4 divide this work differently. The construction in Equation (23) projects M raw instruments through $\hat{\Lambda}$, dedicating none exclusively to RE_t . Consequently, it needs at least two raw instruments whose projections vary independently within the retained substitution directions and move each of the r projected prices and real expenditure separately. The instrument set in Equation (24) instead uses the Stage-1 instrument for real expenditure and reserves r projected components for the projected prices. Therefore, a single raw instrument suffices, provided it moves real expenditure beyond what the r projected components already do. However, if constructed from one raw instrument alone, it leaves no spare components for over-identifying restriction tests. Under LA-AIDS the Stage 3 moments are linear in θ , making the GMM criterion convex once $\hat{\Lambda}$ is fixed, so the identified θ is its unique minimizer. Under exact AIDS the deflator makes the moment bilinear in α^D and the price index, and convexity no longer holds.

Given estimated θ , the price coefficient matrix can be recovered as $\hat{H} = \hat{\Lambda}\hat{\Xi}\hat{\Lambda}^\top$, a symmetric $J \times J$ object that is consistent for H under Assumption 2 as $J\|E\|_2 \rightarrow 0$. Diagonalizing $\hat{\Xi}$ in the columns of $\hat{\Lambda}$ then returns $\hat{V}$ and $\hat{D}$, identified up to a signed permutation of the eigendecomposition. The tail E from the spectral approximation of H in Equation (7) is not identified from the structural demand shock $\delta_{j,t}$ because both get absorbed into a combined error term that cannot be separated.

5.2 Detectable Rank in Finite Samples

The Bai and Ng (2002) information criterion requires a large enough panel to recover the rank of the price coefficient matrix consistently. Two distinct constraints can drive a wedge between the rank the procedure detects and the structural rank it targets. The sample dimensions may be too small to host all r substitution directions, or an individual factor representing an actual substitution dimension may be too weak to separate from the idiosyncratic noise. Whenever either binds, the estimated rank falls short of the structural rank. Selection can also err in the opposite direction. Under exact AIDS the Stone-index residualization leaves a single additional factor in the panel, so the criterion may return $r + 1$, which Section 4.2 shows over-parameterizes Ξ without misspecifying it.

The first constraint is geometric. The residualized panel $\tilde{w}$ has at most $\min(J, T)$ non-zero singular values. Therefore, any estimator can at most recover $\min(J, T)$ factors from $\tilde{w}$ regardless of signal strength. The rank- r reparameterization assumes $r < J$, which is the premise that motivates the estimator. Under this assumption the cross-sectional dimension does not bind and the only constraint is the time series. The estimator would lose its parsimony advantage over the original estimator in Deaton and Muellbauer (1980) if the low-rank property did not hold and $r \rightarrow J$, which is in principle possible but becomes increasingly unlikely with large J . A process with $r > T$ leaves no trace of more than T dimensions in the sample even when $r < J$, making it impossible to recover r .

The second constraint is a signal-to-noise floor that can bind even when $r \leq \min(J, T)$. A signal eigenvalue can only be distinguished from an eigenvalue of the tail if it is sufficiently large compared to the eigenvalues of the noise (Marchenko and Pastur, 1967; Onatski, 2012). A sufficiently weak factor is therefore not just estimated imprecisely but invisible to the procedure, discarded by the penalty term of the Bai and Ng (2002) information criterion. The detectability threshold is not a fixed feature of the

model but a property of the available data. It rises with the noise level and with the number of products relative to the number of time periods. A factor below the threshold is noise-equivalent at a given sample size. Therefore, the same factor could be detectable in one sample but be missed in another.

6 Monte Carlo Simulations

This section evaluates the performance of the estimation approach on simulated data. The Monte Carlo study has two main objectives. The first is to assess if the Bai–Ng IC_2 criterion is capable of correctly determining the rank r of the price coefficient matrix from the IV-residualized expenditure share matrix across a range of market sizes. The second is to evaluate how accurately the three-stage projected MPEC-GMM estimator recovers the structural parameters (α, α^D, H) and the eigenvalue matrix D . Furthermore, I calculate the Marshallian demand elasticity ε^M implied by the estimates to analyze how well the low-rank approximation recovers the true demand elasticity while neglecting the eigenvalue tail. I therefore simulate 100 independent samples whose data generating process adheres to the model assumptions and apply the estimation procedure to each one.

Generating Simulated Data

I simulate data with $T = 50$ periods and market sizes of $J = 20, 50, 100, 200, 500$ products. For each market size, the price coefficient matrix is approximately, but not exactly, rank $r = 4$. Where the price coefficient matrix $H = VDV^\top + E$, with the exact rank- r part $VDV^\top$ carrying eigenvalues $D = \text{diag}(-0.60, -0.36, -0.18, -0.12)$ and a small idiosyncratic tail E . Based on the values for D , I generate an appropriate eigenvector matrix via a discrete Fourier transform and construct E to be symmetric and orthogonal to V . The data generating process draws the idiosyncratic demand shocks $\delta_{j,t}$ independently from a symmetrically truncated normal distribution to exclude extreme outliers. Input prices follow an exogenous firm-specific AR(1) processes augmented by a common aggregate shock. The common component ensures the cross-sectional mean has a non-degenerate variance allowing it to serve as a strong instrument for real expenditure in Stage 1, while the idiosyncratic component permits the construction of valid instruments for the cross-price parameter matrix in the GMM-MPEC of Stage 3. The complete data generating process specification, including exact parameter choices is described in Appendix A2.

Results: Rank Selection

Stages 1 and 2 aim at recovering the rank of the price coefficient matrix and the latent basis of its projection. Their ability to provide precise estimates for both is vital for Stage 3 to deliver consistent estimates of the structural parameters. Table 1 reports the fraction of repetitions in which the criterion in Stage 2 selects the true rank from the data generating process for each simulated market size.

The Bai–Ng IC_2 criterion recovers the true rank $r = 4$ in all 100 repetitions across all market sizes demonstrating that both stages are jointly capable of reliably extracting the rank of the price coefficient matrix. This result furthermore shows that the instrumental variable regression in Stage 1 reliably purges

Table 1: Rank Estimation Accuracy

	$J = 20$	$J = 50$	$J = 100$	$J = 200$	$J = 500$
True rank recovered	100%	100%	100%	100%	100%

Entries are percentages of Stages 1 and 2 selecting the true rank of the price-parameter matrix over all 100 runs for each market size. Detailed description of the data generating process is in Appendix A2.

the residualized share matrix from the additional factor introduced by endogenous real expenditures. If Stage 1’s instrumental variable strategy was to fail, the information criterion in the second stage would consistently overestimate the rank.

Results: Parameter Estimates

Intercepts, Price Effects, and Cross-Price Coefficient Matrix: Stage 3 estimates all structural parameters based on the rank and the loadings matrix provided by in Stages 1 and 2. Tables 2–6 report, for each market size, the mean, median, 95th and 99th percentiles, and maximum of the absolute estimation error, pooled element-wise across all repetitions and parameter entries and the corresponding empirical cumulative distribution functions (CDF) are in Figures 1–5 in the appendix. The MPEC-GMM converged reliably in all 100 repetitions for every $J \in \{20, 50, 100, 200, 500\}$.¹¹

The estimator recovers all product-specific intercepts α well for every market size. The median absolute estimation error is 1.9×10^{-3} at $J = 20$ and roughly 3.2×10^{-3} at $J = 50$. It falls by roughly a factor of ten to for markets of 100 and 200 products and reaches 8.8×10^{-5} at $J = 500$. The corresponding CDFs in part (a) of Figures 1–5 show the full error distributions. At every market size the empirical CDF rises steeply from the origin, showing that the errors concentrate near zero with little mass away from the median. The expenditure coefficients α^D follow the same absolute pattern, with median errors of 3.5×10^{-3} at $J = 20$ and 1.9×10^{-4} at $J = 500$. Their absolute-error CDFs (α^D panels of the same figures) display the same tight concentration near zero with short upper tails. The largest absolute error of the estimates for α across all repetitions and estimated elements is 2.4×10^{-2} at $J = 20$ and 2.8×10^{-2} at $J = 50$, and falls to 2.0×10^{-3} for $J \geq 100$. The errors for α^D estimates behave similarly, with maxima of 4.9×10^{-2} at $J = 50$ and 3.5×10^{-3} at $J = 500$. In every cell the maximum stays within a factor of three of the corresponding 99th percentile.

The entry-wise absolute errors of the recovered cross-price parameter matrix H are uniformly small. The median is 9.9×10^{-4} at $J = 20$ and 1.7×10^{-3} at $J = 50$, then drops to 1.6×10^{-4} at $J = 100$, 1.8×10^{-4} at $J = 200$, and 7.7×10^{-5} at $J = 500$. The neglected idiosyncratic tail E adds an approximation error to H , but its sub-threshold magnitude leaves the recovered entries close to the rank- r signal at every J . Because $|E|$ is itself relatively small, recovering the rank- r signal $VDV^\top$ also recovers the full price coefficient matrix H closely. The H panels of Figures 1–5 back this result as the absolute-error CDFs are steep and concentrated near zero at every J . However, the economically relevant objects are the implied

¹¹The MPEC-GMM estimator is implemented in JuMP (Lubin et al., 2023), Julia’s (Bezanson et al., 2017) mathematical programming language and optimized with Artelys Knitro interior point constrained optimizer (Byrd et al., 2006). However, the GMM criterion function is convex and the constraints exhibit a relatively well behaved form. Therefore, other non-commercial optimizers should also be able to converge quickly to the constrained global optimum.

demand elasticities rather than individual entries of H .

Substitution eigenvalues: The estimation procedure recovers the four eigenvalues $D_1, \dots, D_4$ with high precision. Median absolute errors are largest for the dominant factor $D_1 = -0.60$, running from 7.8×10^{-3} at $J = 20$ to about 3.0×10^{-2} at $J = 500$ and peaking at 6.8×10^{-2} at $J = 50$. The estimator recovers the weakest factor $D_4 = -0.12$ most tightly, with a median absolute error at or below 3.9×10^{-3} at every market size. The middle factors $D_2 = -0.36$ and $D_3 = -0.18$ fall between these two.

Marshallian elasticities: The Marshallian price elasticities ε^M are formed from the estimated price coefficient matrix $\hat{H}$ and the price effect parameter $\hat{\alpha}^D$. I evaluate at each repetition's observed sample-mean budget shares using the true data generating parameters and their estimated counterparts. Median absolute errors remain small relative to the range of price elasticities in the AIDS at every market size with 0.020 at $J = 20$, 0.084 at $J = 50$, 0.016 at $J = 100$, 0.035 at $J = 200$, and 0.039 at $J = 500$. The errors are well controlled throughout the distribution. The largest single absolute error across all repetitions and elasticity entries is about 2.2 (at $J = 500$) and at or below 1.1 for every other market size. The ε^M panels of Figures 1–5 show the same picture for the elasticities. The error mass sits close to zero and the CDF flattens well within the plotted range. The large maxima therefore reflect only a handful of entries rather than the bulk.

The projected MPEC-GMM estimator converges in every run for every simulated market size and recovers all structural parameters with high precision, including the largest market size of 500 products, where the dominant factor D_1 retains a median absolute error of about 3.0×10^{-2} and the weakest factor D_4 about 1.4×10^{-3} . Accuracy is not monotone in J . The cell with the highest errors is $J = 50$, where the noise-to-signal ratio peaks at 15 per cent of the mean share and the median eigenvalue errors are correspondingly largest, while the most favourable cell is $J = 100$. Elasticity errors stay within economically interpretable bounds throughout, and the small idiosyncratic tail E leaves both rank selection and structural recovery intact, confirming that the estimator handles the approximate low-rank setting well. Together, these results show the estimator can reliably recover all estimates in markets that are far out of reach for the original AIDS estimator, with a relatively modest number of time periods.

Table 2: Monte Carlo estimation errors: $J = 20$

Parameter	Absolute error $ \hat{\theta} - \theta_0 $				
	Mean	Median	95th pct	99th pct	Max
$\alpha_1, \dots, \alpha_{J-1}$	0.002516	0.001913	0.006798	0.0106	0.02385
$\alpha_1^D, \dots, \alpha_{J-1}^D$	0.004568	0.003498	0.01201	0.01718	0.04405
η	0.001215	0.0009858	0.003146	0.004573	0.01006
D_1	0.01159	0.007779	0.03475	0.04733	0.06331
D_2	0.004095	0.002704	0.009807	0.01333	0.01669
D_3	0.002042	0.001627	0.005276	0.006672	0.006774
D_4	0.001648	0.001339	0.00392	0.00456	0.004563
ε^M	0.025	0.01996	0.06528	0.09545	0.2066

Notes: Entries are the absolute estimation error $|\hat{\theta} - \theta_0|$ pooled element-wise across the converged repetitions and parameter entries (the D_k rows pool one error per repetition). The α and α^D rows cover the $J - 1$ estimated elements only: α_J and α_J^D are recovered from the adding-up restrictions rather than estimated, so their error is the accumulated error of the remaining elements and is not reported. Marshallian elasticities are evaluated at each repetition's observed sample-mean shares and log prices. The data generating process is described in Appendix A2.

Table 3: Monte Carlo estimation errors: $J = 50$

Parameter	Absolute error $ \hat{\theta} - \theta_0 $				
	Mean	Median	95th pct	99th pct	Max
$\alpha_1, \dots, \alpha_{J-1}$	0.00424	0.003241	0.0118	0.01748	0.02791
$\alpha_1^D, \dots, \alpha_{J-1}^D$	0.006685	0.005163	0.01802	0.0272	0.04857
η	0.002111	0.001667	0.005661	0.008113	0.01585
D_1	0.0803	0.06834	0.1863	0.2256	0.2328
D_2	0.02441	0.02065	0.06195	0.0698	0.08802
D_3	0.00701	0.004953	0.01678	0.0299	0.03413
D_4	0.004456	0.003889	0.01097	0.01634	0.01671
ε^M	0.107	0.08384	0.2888	0.4186	1.111

Notes: Entries are the absolute estimation error $|\hat{\theta} - \theta_0|$ pooled element-wise across the converged repetitions and parameter entries (the D_k rows pool one error per repetition). The α and α^D rows cover the $J - 1$ estimated elements only: α_J and α_J^D are recovered from the adding-up restrictions rather than estimated, so their error is the accumulated error of the remaining elements and is not reported. Marshallian elasticities are evaluated at each repetition's observed sample-mean shares and log prices. The data generating process is described in Appendix A2.

Table 4: Monte Carlo estimation errors: $J = 100$

Parameter	Absolute error $ \hat{\theta} - \theta_0 $				
	Mean	Median	95th pct	99th pct	Max
$\alpha_1, \dots, \alpha_{J-1}$	0.0002639	0.0002107	0.0007009	0.0009963	0.0016
$\alpha_1^D, \dots, \alpha_{J-1}^D$	0.0004487	0.0003559	0.001186	0.001677	0.002823
η	0.0001958	0.000155	0.0005194	0.0007675	0.001664
D_1	0.01224	0.01005	0.03601	0.04623	0.04773
D_2	0.003811	0.003166	0.01048	0.01361	0.01518
D_3	0.00154	0.001408	0.003566	0.0052	0.005928
D_4	0.0008507	0.000645	0.002204	0.003036	0.003837
ε^M	0.01989	0.01552	0.05351	0.08037	0.1881

Notes: Entries are the absolute estimation error $|\hat{\theta} - \theta_0|$ pooled element-wise across the converged repetitions and parameter entries (the D_k rows pool one error per repetition). The α and α^D rows cover the $J - 1$ estimated elements only: α_J and α_J^D are recovered from the adding-up restrictions rather than estimated, so their error is the accumulated error of the remaining elements and is not reported. Marshallian elasticities are evaluated at each repetition's observed sample-mean shares and log prices. The data generating process is described in Appendix A2.

Table 5: Monte Carlo estimation errors: $J = 200$

Parameter	Absolute error $ \hat{\theta} - \theta_0 $				
	Mean	Median	95th pct	99th pct	Max
$\alpha_1, \dots, \alpha_{J-1}$	0.0002789	0.0002012	0.0007642	0.001027	0.001988
$\alpha_1^D, \dots, \alpha_{J-1}^D$	0.0005223	0.0004061	0.001383	0.001807	0.003147
η	0.0002317	0.0001756	0.0006342	0.0009998	0.003309
D_1	0.03561	0.02678	0.09037	0.1188	0.2096
D_2	0.006906	0.005571	0.01755	0.02719	0.03134
D_3	0.004401	0.002604	0.01128	0.01675	0.02004
D_4	0.001662	0.001379	0.004318	0.005684	0.006096
ε^M	0.04706	0.03523	0.1301	0.2079	0.7388

Notes: Entries are the absolute estimation error $|\hat{\theta} - \theta_0|$ pooled element-wise across the converged repetitions and parameter entries (the D_k rows pool one error per repetition). The α and α^D rows cover the $J - 1$ estimated elements only: α_J and α_J^D are recovered from the adding-up restrictions rather than estimated, so their error is the accumulated error of the remaining elements and is not reported. Marshallian elasticities are evaluated at each repetition's observed sample-mean shares and log prices. The data generating process is described in Appendix A2.

Table 6: Monte Carlo estimation errors: $J = 500$

Parameter	Absolute error $ \hat{\theta} - \theta_0 $				
	Mean	Median	95th pct	99th pct	Max
$\alpha_1, \dots, \alpha_{J-1}$	0.0002423	8.811e-05	0.001071	0.00156	0.001919
$\alpha_1^D, \dots, \alpha_{J-1}^D$	0.0005056	0.0001868	0.002306	0.002941	0.003507
η	0.0001169	7.671e-05	0.0003769	0.0006412	0.001496
D_1	0.03965	0.02969	0.1097	0.2125	0.2486
D_2	0.01035	0.006999	0.03229	0.04796	0.08323
D_3	0.008028	0.004042	0.03696	0.05643	0.05686
D_4	0.002039	0.001383	0.004408	0.005778	0.03112
ε^M	0.05931	0.0385	0.1911	0.3264	2.191

Notes: Entries are the absolute estimation error $|\hat{\theta} - \theta_0|$ pooled element-wise across the converged repetitions and parameter entries (the D_k rows pool one error per repetition). The α and α^D rows cover the $J - 1$ estimated elements only: α_J and α_J^D are recovered from the adding-up restrictions rather than estimated, so their error is the accumulated error of the remaining elements and is not reported. Marshallian elasticities are evaluated at each repetition’s observed sample-mean shares and log prices. The data generating process is described in Appendix A2.

7 Relationship to the Literature

This paper relates to two literatures, the structural demand systems literature that supplies consumer-theory-based demand models and the approximate-factor-model literature that supplies machinery for determining the low-rank structure, and to a growing body of work applying dimension reduction to demand systems.

Demand Estimation Foundations

The AIDS family of Deaton and Muellbauer (1980), Banks et al. (1997), and Lewbel and Pendakur (2009) delivers a flexible specification linear in log prices and log real expenditure under the canonical restrictions of utility maximization. Its limitation in large markets is the quadratic growth of the free cross-price block. The characteristic-space alternatives such as logit or nested logit models (Berry, 1994; Grigolon and Verboven, 2014), and the BLP demand model, the workhorse model of modern industrial organization (Berry et al., 1995; Nevo, 2000) sidestep this prohibitive parameter growth by writing substitution patterns as functions of distances in a low-dimensional characteristics space. However, their estimation requires the researcher to take a stance on which product characteristics are relevant and to observe all these characteristics, a requirement that production datasets generally do not meet. The estimator developed here retains the AIDS structure and consumer-theoretic foundations while handling the prohibitive quadratic parameter growth without requiring to observe product characteristics.

Factor Model Literature

Approximate factor models in the sense of Chamberlain and Rothschild (1983); Bai and Ng (2002) split the covariance matrix into a pervasive component whose leading r eigenvalues diverge with the number of units and a remainder that may be correlated across units and over time but whose eigenvalues stays

bounded. Imposing the rank- r decomposition on the price coefficient matrix imprints the same pattern on how the residualized shares move together across products. The shares contain few pervasive substitution directions, plus idiosyncratic noise made up of the discarded eigenvalue tail and the structural demand shocks. Stock and Watson (2002) establish consistency of asymptotic principal components estimators in approximate factor models. Bai and Ng (2002) develop information criteria for selecting the number of pervasive factors, which Stage 2 of the proposed estimator uses to determine the rank of the price coefficient matrix. Onatski (2010) provides eigenvalue-ratio alternatives. These techniques have been developed primarily for and applied to macroeconomic forecasting and asset pricing, with little connection to structural demand models developed and utilized in empirical microeconomics and especially Industrial Organization. However, these methods are a well-suited tool to reduce the dimensionality of product-space models like the AIDS. This paper supplies that connection by putting methods to estimate approximate factor models to work inside a structural demand estimation approach to achieve what none of the two literatures can do on its own.

Related Dimension-Reduction Work

A growing literature applies statistical dimension reduction to demand systems with many goods. Hoderlein and Lewbel (2012) study regressor dimension reduction in large demand systems under utility-maximization restrictions, showing how principal components and related methods can reduce the dimensionality of price regressors. Barigozzi and Moneta (2016) represent a system of budget shares as an approximate factor model, use principal components to estimate the low-dimensional factor space, and interpret the recovered factors as Engel curves that summarize the main directions of consumption variation. Chernozhukov et al. (2019) address the “many prices” problem with Lasso and ridge-type regularization to estimate average demand and consumer-surplus bounds when the number of price covariates is large. These papers establish that statistical dimension reduction can be used in demand estimation. However, they do not import the approximate-factor-model toolkit to identify the substitution subspace in combination with a structural demand system like the proposed approach does. Outside product demand, factor-structured demand systems have appeared prominently in finance. Koijen and Yogo (2019) develop a characteristic-based asset-demand system in which expected returns and factor loadings depend on asset attributes.

Fally and Ligon (2026) establish a functional-form equivalence between a demand system depending on K scalar functions of prices, which they name *aggregators*, and its cross-price coefficient matrix having rank K . Under utility-maximization, the resulting K -aggregator class nests virtually every flexible demand system in use, from logit and CES through nested logit, BLP-style mixed logit, the IPDL of Fosgerau et al. (2024), and the linearized AIDS. Estimating the K -aggregator specification directly, rather than through a random-utility micro-foundation, lets Fally and Ligon (2026) represent the complementarity between goods that random-utility models rule out by construction. Fally and Ligon (2026) apply their approach on NielsenIQ scanner data, where their K -aggregator model supports richer substitution patterns than BLP allows.

The AIDS discussed in this paper can be seen as a special case of Fally and Ligon (2026)’s K -aggregator class. The two papers therefore meet at the same structural object, and the present paper takes the K -aggregator equivalence theorem as the demand-theoretic derivation of the low-rank structure it estimates.

The leading eigenvectors of the price coefficient matrix span the directions along which Fally and Ligon (2026)’s structural aggregators would load. The approach proposed in this paper does not however rely on Fally and Ligon (2026)’s main theorem or estimation approach. Importantly, Fally and Ligon (2026) impose exact rank on the price coefficient matrix, which their equivalence between K aggregators and rank K requires, while the present paper imposes only an approximate rank structure under a spectral-gap condition, the regime in which the price coefficient matrix is dominated by, but not equal to, a few pervasive substitution directions. Furthermore, Fally and Ligon (2026) recover the substitution directions by projecting onto observed product characteristics, which the researcher must specify and observe. This paper instead extracts the latent basis from the residualized expenditure share matrix by asymptotic principal components, requiring no observed characteristics.

8 Discussion and Conclusion

Estimating standard flexible product-space demand systems such as the AIDS is infeasible in markets with many differentiated products, because the number of cross-price parameters grows quadratically in the number of products. Characteristic-based alternatives that sidestep this prohibitive parameter growth rely on observing product characteristics that production panels typically lack. This paper develops an estimator for the AIDS, in both its exact and linear-approximate forms, by relying on an approximate rank- r cross price parameter matrix. The estimator recovers flexible substitution patterns without observing product characteristics and is therefore well-suited for settings in which firm- or product-level production data are the primary source.

The estimation strategy proceeds in three stages, each delivering the input the next one needs. The first stage residualizes the expenditure share matrix by regressing expenditure shares on real expenditure and an intercept. The resulting estimates are biased, because the regression omits the cross price block, but the dominant bias loads on the substitution directions themselves. Therefore, residualization leaves an approximate factor model whose pervasive dimensions are exactly the substitution dimensions of the cross-price parameter matrix. The second stage determines the number of the substitution dimensions off the residualized expenditure share matrix with the Bai–Ng information criterion and estimates a basis for the substitution subspace by asymptotic principal components. The third stage estimates all structural parameters in a convex MPEC-GMM problem with instruments that are projected onto the same basis, matching the low-dimensional endogeneity structure. The second stage recovers the latent basis only up to rotation, but a free symmetric substitution block absorbs the rotation exactly in the third stage.

The rank- r structure this procedure exploits is purely statistically motivated. The spectral decomposition of Equation (7) is available for any Slutsky-symmetric price coefficient matrix, the pervasiveness and spectral-gap conditions of Assumption 2 are statements about the eigenstructure of the data, and both the number and the structure of the substitution dimensions are estimated, not imposed. Fally and Ligon (2026) prove that imposing exact rank r on the price coefficient matrix is equivalent to consumers evaluating the full price vector through K scalar aggregators, with the AIDS as one special case of their K -aggregator class. The construction here does not rest on that equivalence, but it gives the recovered directions a preference interpretation. Fally and Ligon (2026)’s analysis, however, remains within the exact rank- r case, and their estimator requires rich product-attribute data to project prices onto observed

characteristics, whereas the present estimator projects onto the data-driven latent basis $\hat{\Lambda}$. The substitution dimensions it recovers are therefore latent, and without product characteristics the data cannot supply their economic meaning. The estimator identifies all pervasive directions, but not the names that an attribute-based estimator takes from observed characteristics.

The estimated demand system supports counterfactuals that hold preferences fixed and ask how changes in quantities like prices, expenditure, or product ownership affect endogenous quantities conditional on the cross price parameter matrix is policy-invariant. When simulating the removal of a product the surviving system must be renormalized to restore the demand restrictions and reallocate the discontinued product’s expenditure. However, simulating the introduction of a new product variety is generally infeasible. Because the estimated directions carry no characteristic labels, the system cannot place a variety it has never seen within the substitution subspace, whereas a characteristic-based model such as BLP can. This is the price of not requiring observable characteristics in the first place.

Some conditions limit the settings in which the estimator operates as intended. The cross-price structure must be dominated by a few pervasive substitution directions, so that the price coefficient matrix is approximately low-rank with a spectral gap separating the retained dimensions from the idiosyncratic tail. When the number of pervasive substitution directions approaches the number of products, the advantage of the proposed estimator vanishes, and the cross-price parameter count reverts to the quadratic growth in the number of products that the low-rank restriction was designed to avoid. Furthermore, each retained dimension has to move expenditure shares across many products rather than a handful, as the estimator picks up broad, category-wide substitution more readily than a direction confined to a small cluster of close substitutes. Moreover, the time dimension must be large enough to host the substitution structure, since no more than $\min(J, T)$ directions can be recovered. Pervasive substitution directions need to leave strong enough patterns in expenditure shares, otherwise, they are for the estimator indistinguishable from noise. When these conditions hold, the selected rank is the rank the sample can support, when not, $\hat{r}$ is best read as a lower bound on the structural rank.

Several extensions of the baseline estimator are worth noting. While the linear approximation of the AIDS is the main specification, the estimator already covers the exact AIDS as well. A more substantive direction is to carry the same low-rank reduction to other flexible product-space systems whose cross-price parameters grow quadratically in the number of products such as the translog of Christensen et al. (1975), the quadratic AIDS of Banks et al. (1997), and the EASI system of Lewbel and Pendakur (2009). Each of these contains a quadratically growing price coefficient matrix whose free parameters the low-rank approximation can substantially reduce. Furthermore, the derivation in this paper assumes that the rank and the loading matrix are time-invariant, which is a maintained assumption of the applied approximate factor model estimator. Time-varying loadings can in principle be accommodated through time-varying factor models (Mikkelsen et al., 2019; Forni et al., 2000; Su and Wang, 2017).

Taken together, this paper shows that the AIDS can be estimated in markets far larger than its parameter count would ordinarily permit, using only the prices, quantities, and instruments that production panels record. By decomposing the cross-price block into a small number of pervasive substitution directions whose count is read from the data rather than imposed, the estimator brings markets with well above 100 products well inside its operating range, where the unrestricted AIDS would require infeasibly many periods for identification, and other demand models and their estimators would require a rich set

of observed product characteristics. The Monte Carlo evidence confirms that the estimator determines the rank, all structural parameters, and the implied elasticities precisely across market sizes from 20 to 500 products with only 50 time periods. By connecting the approximate factor model literature to the structural demand estimation approaches of empirical industrial organization, the estimator reduces the parameter growth of the cross-price block and offers a route to doing the same for the other flexible product-space systems.

References

- ACKERBERG, D. A. (2001): “Empirically Distinguishing Informative and Prestige Effects of Advertising,” *The RAND Journal of Economics*, 32, 316–333.
- (2003): “Advertising, Learning, and Consumer Choice in Experience Good Markets: An Empirical Examination,” *International Economic Review*, 44, 1007–1040.
- BAI, J. AND S. NG (2002): “Determining the Number of Factors in Approximate Factor Models,” *Econometrica*, 70, 191–221.
- BANKS, J., R. BLUNDELL, AND A. LEWBEL (1997): “Quadratic Engel Curves and Consumer Demand,” *Review of Economics and Statistics*, 79, 527–539.
- BARIGOZZI, M. AND A. MONETA (2016): “Identifying the Independent Sources of Consumption Variation,” *Journal of Applied Econometrics*, 31, 420–449.
- BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): “Automobile Prices in Market Equilibrium,” *Econometrica*, 63, 841–890.
- BERRY, S. T. (1994): “Estimating Discrete-Choice Models of Product Differentiation,” *RAND Journal of Economics*, 25, 242–262.
- BEZANSON, J., A. EDELMAN, S. KARPINSKI, AND V. B. SHAH (2017): “Julia: A fresh approach to numerical computing,” *SIAM Review*, 59, 65–98.
- BRESNAHAN, T. F., S. STERN, AND M. TRAJTENBERG (1997): “Market Segmentation and the Sources of Rents from Innovation: Personal Computers in the Late 1980s,” *The RAND Journal of Economics*, 28, S17–44.
- BRONNENBERG, B. J., S. K. DHAR, AND J.-P. DUBÉ (2007): “Consumer packaged goods in the United States: National brands, local branding,” *Journal of Marketing Research*, 44, 4–13.
- BRONNENBERG, B. J., S. K. DHAR, AND J. H. DUBÉ (2009): “Brand History, Geography, and the Persistence of Brand Shares,” *Journal of Political Economy*, 117, 87–115.
- BYRD, R. H., J. NOCEDAL, AND R. A. WALTZ (2006): “KNITRO: An integrated package for nonlinear optimization,” in *Large-scale nonlinear optimization*, Springer, 35–59.
- CHAMBERLAIN, G. AND M. ROTHSCILD (1983): “Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets,” *Econometrica*, 51, 1281–1304.
- CHERNOZHUKOV, V., J. HAUSMAN, AND W. NEWEY (2019): “Demand Analysis with Many Prices,” NBER Working Paper 26424, National Bureau of Economic Research.
- CHRISTENSEN, L. R., D. W. JORGENSON, AND L. J. LAU (1975): “Transcendental logarithmic utility functions,” *The American Economic Review*, 65, 367–383.

- DEATON, A. AND J. MUELLBAUER (1980): “An Almost Ideal Demand System,” *American Economic Review*, 70, 312–326.
- ERDEM, T. AND M. P. KEANE (1996): “Decision-Making under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets,” *Marketing Science*, 15, 1–20.
- FALLY, T. AND E. LIGON (2026): “Consumer Demand with Price Aggregators and Low-Rank Cross-Price Effects,” Working paper, Department of Agricultural and Resource Economics, UC Berkeley.
- FORNI, M., M. HALLIN, M. LIPPI, AND L. REICHLIN (2000): “The Generalized Dynamic-Factor Model: Identification and Estimation,” *The Review of Economics and Statistics*, 82, 540–554.
- FOSGERAU, M., J. MONARDO, AND A. DE PALMA (2024): “The Inverse Product Differentiation Logit Model,” *American Economic Journal: Microeconomics*, 16, 329–370.
- GANDHI, A. AND J.-F. HOUE (2019): “Measuring Substitution Patterns in Differentiated-Products Industries,” NBER Working Paper 26375, National Bureau of Economic Research.
- GRIGOLON, L. AND F. VERBOVEN (2014): “Nested Logit or Random Coefficients Logit? A Comparison of Alternative Discrete Choice Models of Product Differentiation,” *The Review of Economics and Statistics*, 96, 916–935.
- HAUSMAN, J. A. AND G. K. LEONARD (2002): “The competitive effects of a new product introduction: A case study,” *The Journal of Industrial Economics*, 50, 237–263.
- 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.
- KOIJEN, R. S. J. AND M. YOGO (2019): “A Demand System Approach to Asset Pricing,” *Journal of Political Economy*, 127, 1475–1515.
- LEWBEL, A. AND K. PENDAKUR (2009): “Tricks with Hicks: The EASI Demand System,” *American Economic Review*, 99, 827–863.
- LUBIN, M., O. DOWSON, J. DIAS GARCIA, J. HUCHETTE, B. LEGAT, AND J. P. VIELMA (2023): “JuMP 1.0: Recent improvements to a modeling language for mathematical optimization,” *Mathematical Programming Computation*.
- MARCHENKO, V. A. AND L. A. PASTUR (1967): “Distribution of Eigenvalues for Some Sets of Random Matrices,” *Mathematics of the USSR-Sbornik*, 1, 457–483.
- McFADDEN, D. (1978): “Modeling the Choice of Residential Location,” in *Spatial Interaction Theory and Planning Models*, ed. by A. Karlqvist, L. Lundqvist, F. Snickars, and J. W. Weibull, Amsterdam: North-Holland, 75–96.
- MIKKELSEN, J. G., E. HILLEBRAND, AND G. URGAS (2019): “Consistent estimation of time-varying loadings in high-dimensional factor models,” *Journal of Econometrics*, 208, 535–562.

- NEVO, A. (2000): “A practitioner’s guide to estimation of random-coefficients logit models of demand,” *Journal of economics & management strategy*, 9, 513–548.
- ONATSKI, A. (2010): “Determining the Number of Factors from Empirical Distribution of Eigenvalues,” *Review of Economics and Statistics*, 92, 1004–1016.
- (2012): “Asymptotics of the Principal Components Estimator of Large Factor Models with Weakly Influential Factors,” *Journal of Econometrics*, 168, 244–258.
- STOCK, J. H. AND M. W. WATSON (2002): “Forecasting Using Principal Components From a Large Number of Predictors,” *Journal of the American Statistical Association*, 97, 1167–1179.
- SU, C.-L. AND K. L. JUDD (2012): “Constrained optimization approaches to estimation of structural models,” *Econometrica*, 80, 2213–2230.
- SU, L. AND X. WANG (2017): “On time-varying factor models: Estimation and testing,” *Journal of Econometrics*, 198, 84–101.
- SUDHIR, K. (2001): “Competitive Pricing Behavior in the Auto Market: A Structural Analysis,” *Marketing Science*, 20, 42–60.
- VERBOVEN, F. (1996): “International Price Discrimination in the European Car Market,” *The RAND Journal of Economics*, 27, 240–268.

A1 The Full Estimator with Free Loadings

This appendix develops the full estimator that treats the loading matrix V as an unknown and estimates (V, D) jointly with (α, α^D) inside the structural third stage, rather than fixing V at the second-stage basis $\hat{\Lambda}$ and parameterizing the cross-price block as $H = \hat{\Lambda}\Xi\hat{\Lambda}^\top$. The full estimator relaxes the maintained assumption of the projected estimator that the column space of $\hat{\Lambda}$ coincides with the true rank- r substitution subspace, at the cost of a parameter count that grows linearly in J and a non-convex objective, leading to a more challenging problem for the constrained numerical optimizer.

A1.1 Parameterization and Constraints

The full estimator parameterizes $H = VDV^\top$ with $V \in \mathbb{R}^{J \times r}$ a loading matrix and $D = \text{diag}(d_1, \dots, d_r)$ a diagonal matrix of substitution eigenvalues. The structural parameter vector is $\theta^{\text{full}} = (\alpha, \alpha^D, V, D)$ and the constraint set on the cross-price block is then

$$V^\top V = \mathbf{I}_r, \quad (\text{orthonormality}) \quad (\text{A1.1})$$

$$V^\top \mathbf{1}_J = \mathbf{0}_r, \quad (\text{homogeneity}) \quad (\text{A1.2})$$

$$D_{kl} = 0 \quad \text{for } k \neq l, \quad (\text{diagonal eigenvalue block}) \quad (\text{A1.3})$$

together with the adding-up restrictions on (α, α^D) already imposed in the projected estimator. Slutsky symmetry holds automatically from $H = VDV^\top$ with diagonal D . The constraints on V leave $Jr - r(r+1)/2 - r$ free entries in the loading block. Adding the r entries of the diagonal D block gives a total cross-price free parameter count of $Jr - r(r+1)/2$ that grows linearly in J .

The implied price coefficient matrix at any feasible (V, D) satisfies $H\mathbf{1}_J = VDV^\top\mathbf{1}_J = \mathbf{0}$ by (A1.2), so price homogeneity is automatic. Symmetry follows from diagonality of D together with Constraint (A1.1). The diagonal- D fixes the rotational indeterminacy $(V, D) \sim (V\Phi, \Phi^\top D\Phi)$ for $\Phi \in O(r)$.

A1.2 MPEC Formulation

The structural moment conditions and the GMM criterion are identical to those of the projected estimator. The same instruments $Z_t^{TS} = (1, \hat{\Lambda}^\top Z_t^{\text{raw},1}, \dots, \hat{\Lambda}^\top Z_t^{\text{raw},M})^\top$ from Equation (23) deliver the moment vector

$$g(\theta^{\text{full}}) = \frac{1}{T} \sum_{t=1}^T Z_t^{TS} \otimes \delta_t(\theta^{\text{full}}), \quad (\text{A1.4})$$

where $\delta_{j,t}(\theta^{\text{full}}) = w_{j,t} - \alpha_j - \alpha_j^D RE_t - [VDV^\top \ln P_t^Q]_j$ for $j \leq J-1$. The full MPEC reads

$$\begin{aligned}
& \arg \min_{\alpha, \alpha^D, V, D, \delta} \quad g(\theta^{\text{full}})^\top W g(\theta^{\text{full}}) \\
& \text{s.t.} \quad \delta_{j,t} = w_{j,t} - \alpha_j - \alpha_j^D RE_t - [VDV^\top \ln P_t^Q]_j, \quad j \leq J-1, \quad \forall t, \\
& \quad g(\theta^{\text{full}}) = \frac{1}{T} \sum_{t=1}^T Z_t^{TS} \otimes \delta_t, \\
& \quad V^\top V = \mathbf{I}_r, \\
& \quad V^\top \mathbf{1}_J = \mathbf{0}_r, \\
& \quad D_{kl} = 0, \quad k \neq l, \\
& \quad \alpha_J = 1 - \sum_{j=1}^{J-1} \alpha_j, \quad \alpha_J^D = -\sum_{j=1}^{J-1} \alpha_j^D.
\end{aligned} \tag{A1.5}$$

Starting values for V can be spectrally rotated asymptotic principal components loadings $\hat{\Lambda}$ from the rank-selection step. Starting values for D can be the leading eigenvalues of the IV-residualized share covariance, sign-corrected so that the diagonal entries are negative under downward-sloping demand. Starting values for (α, α^D) can be the 2SLS coefficients from the Stage 1 regression in Equation (13). Multi-start over V is advisable to address local optima.

A2 Monte Carlo Data-Generating Process

This appendix gives the complete specification of the data-generating process used in the Monte Carlo study of Section 6.

A2.1 Cross-Price Parameter Matrix

The loading matrix $V \in \mathbb{R}^{J \times r}$ is built from the leading r Fourier columns of the J -dimensional DFT basis. For J even and $r \leq J-1$, the k -th column alternates cosine and sine harmonics at successive frequencies $f = 1, 2, \dots$:

$$V_{j, 2f-1} = \sqrt{\frac{2}{J}} \cos\left(\frac{2\pi j f}{J}\right), \quad V_{j, 2f} = \sqrt{\frac{2}{J}} \sin\left(\frac{2\pi j f}{J}\right), \tag{A2.1}$$

normalized so that $\|V_k\|^2 = 1$ and $V^\top \mathbf{1}_J = \mathbf{0}_r$ by the orthogonality of the Fourier basis. The Fourier construction ensures that each dominant substitution direction is well-spread across all J products at every market size, so no individual product carries a disproportionate fraction of the structural signal.

The substitution eigenvalues are fixed at

$$D = \text{diag}(-0.60, -0.36, -0.18, -0.12), \tag{A2.2}$$

with eigenvalues ordered by decreasing absolute magnitude. The true price coefficient matrix is $H = VDV^\top + E$, where the signal $VDV^\top$ has exact algebraic rank $r = 4$ and E is a small symmetric idiosyncratic tail. I draw E as a symmetric Gaussian matrix projected onto the orthogonal complement of V and $\mathbf{1}_J$, $E = \mathcal{M}\tilde{E}\mathcal{M}$ with $\mathcal{M} = I_J - VV^\top - J^{-1}\mathbf{1}_J\mathbf{1}_J^\top$, then rescale it so its spectral norm equals a fixed fraction of the weakest structural eigenvalue, $\|E\|_2 = \rho_E |D_r|$ with $\rho_E = 0.01$ for $J \leq 200$ and $\rho_E = 0.005$ at $J = 500$. The orthogonality $V^\top E = 0$ keeps the rank- r signal $VDV^\top$ exactly intact, so E enters only as

the residual tail of the spectral approximation (7); because $\|E\|_2$ stays below the detectability wall (A2.3) at every (J, T) , the design is approximately—not exactly—low rank without disturbing rank recovery, and the structural-parameter results therefore speak to the approximate case the theory targets rather than to an exactly low-rank ideal.

A2.2 Intercepts and Expenditure Coefficients

The true budget shares α_j are initialized as $1/J$ plus small centred normal perturbations of scale $0.02/J$, then renormalized to sum to one and clamped above $0.9/J$ to ensure strict positivity. The expenditure coefficients α_j^D are drawn independently from $\mathcal{N}(0, (0.001 s_J)^2)$ with $s_J = \sqrt{J_0/J}$ and centred to satisfy the adding-up restriction $\sum_j \alpha_j^D = 0$. The s_J scaling keeps the product $\alpha_j^D \cdot RE_t$ of the same order as $\alpha_j \approx 1/J$ as the market size grows.

A2.3 Demand Shocks and Prices

The idiosyncratic demand shock $\delta_{j,t}$ is drawn independently across (j, t) from a truncated normal distribution $\mathcal{N}(0, \sigma_\delta^2)$ truncated to $[-2\sigma_\delta, 2\sigma_\delta]$. The noise level σ_δ is chosen to keep all $r = 4$ signal eigenvalues above the Baik–Ben Arous–Péché detectability wall

$$\lambda_k(\text{signal}) > \sigma_\delta^2 (1 + \sqrt{J/T})^2. \quad (\text{A2.3})$$

The level is set in three tiers, $\sigma_\delta = 3 \times 10^{-3}$ for $J \in \{20, 50\}$, 3×10^{-4} for $J \in \{100, 200\}$, and 1×10^{-4} at $J = 500$, each chosen to remain a small fraction—between 3 and 15 per cent—of the mean budget share $1/J$ while keeping all four signal eigenvalues above the wall. The wall tightens as $c = J/T$ grows: at $J = 500$ its height with $T = 50$ is $\sigma_\delta^2 (1 + \sqrt{10})^2 \approx 17\sigma_\delta^2$, which would push the weakest signal eigenvalue $D_4^2 \cdot \text{Var}(V_4^\top \ln P^Q)$ below the wall at the looser noise levels, so σ_δ is tightened to 1×10^{-4} at $J = 500$.

Equilibrium prices are generated by a cost-plus pricing equation with firm-specific and aggregate components from input prices. Labour and material input prices follow firm-specific AR(1) processes augmented by a common shock:

$$\ln P_{j,t}^L = \rho^{PL} \ln P_{j,t-1}^L + \log \varepsilon_{j,t}^L + \sigma_{\text{com}} u_t^L, \quad (\text{A2.4})$$

where $\varepsilon_{j,t}^L$ is a firm-specific innovation drawn from a truncated normal with centre 0.4 and standard deviation $\sigma_{\text{idio}} = 0.03 s_J \sigma_{\text{mult}}$, $u_t^L \sim \mathcal{N}(0, 1)$ is an aggregate shock common to all J firms in period t , $\sigma_{\text{com}} = 0.25$, $\rho^{PL} = 0.5$, and $\sigma_{\text{mult}} = 10$. The same structure governs material prices $\ln P_{j,t}^M$.

This mixed design serves two roles. First, the cross-sectional mean $z_t^L = J^{-1} \sum_j \ln P_{j,t}^L$ has non-degenerate variance $\text{Var}(z_t^L) \approx \sigma_{\text{com}}^2 = 0.0625$ at every J , because the idiosyncratic terms average out by the law of large numbers. The instrument z_t^L therefore provides a strong first stage for real expenditure RE_t in the IV pre-step of Stage 1 regardless of market size. Second, the projection $V_0^\top \ln P_t^L$ onto the leading Fourier columns provides the variation needed for the structural GMM instruments of Stage 2. The common shift cancels from this projection because $V_0^\top \mathbf{1}_J = \mathbf{0}$ by Fourier orthogonality; the GMM instruments therefore carry only idiosyncratic price variation, which is orthogonal to any aggregate demand shock by construction.

All idiosyncratic price-variation parameters scale with $s_J = \sqrt{J_0/J}$, keeping the effective signal-to-noise ratio of the share panel comparable across market sizes.

A2.4 Expenditure and Simulation Procedure

Log total nominal expenditure $\ln X_t$ follows an AR(1) with mean 9.034, persistence $\rho_X = 0.7$, and innovation standard deviation 0.1. The mean is set so that $E[RE_t] = E[\ln X_t] - E[\ln P_t^S] \approx 0.5$, which keeps $\alpha_j^D \cdot RE_t$ of comparable magnitude to α_j and avoids near-collinearity between the intercept and expenditure-coefficient terms in the share equation. Each simulation run discards 10 burn-in periods before recording the $T = 50$ sample observations. The MPEC is solved with KNITRO using an interior-point algorithm with feasibility and optimality tolerances of 10^{-6} and a step-size tolerance of 10^{-8} .

A2.5 Error CDFs

Figures 1–5 display the empirical cumulative distribution functions (CDFs) of the pooled element-wise absolute errors for α , α^D , H , and the Marshallian elasticity ϵ^M . Each figure corresponds to one market size J . The α and α^D panels cover the estimated elements $\alpha_1, \dots, \alpha_{J-1}$ and $\alpha_1^D, \dots, \alpha_{J-1}^D$; the J -th element of each is recovered from adding-up rather than estimated and is therefore excluded.

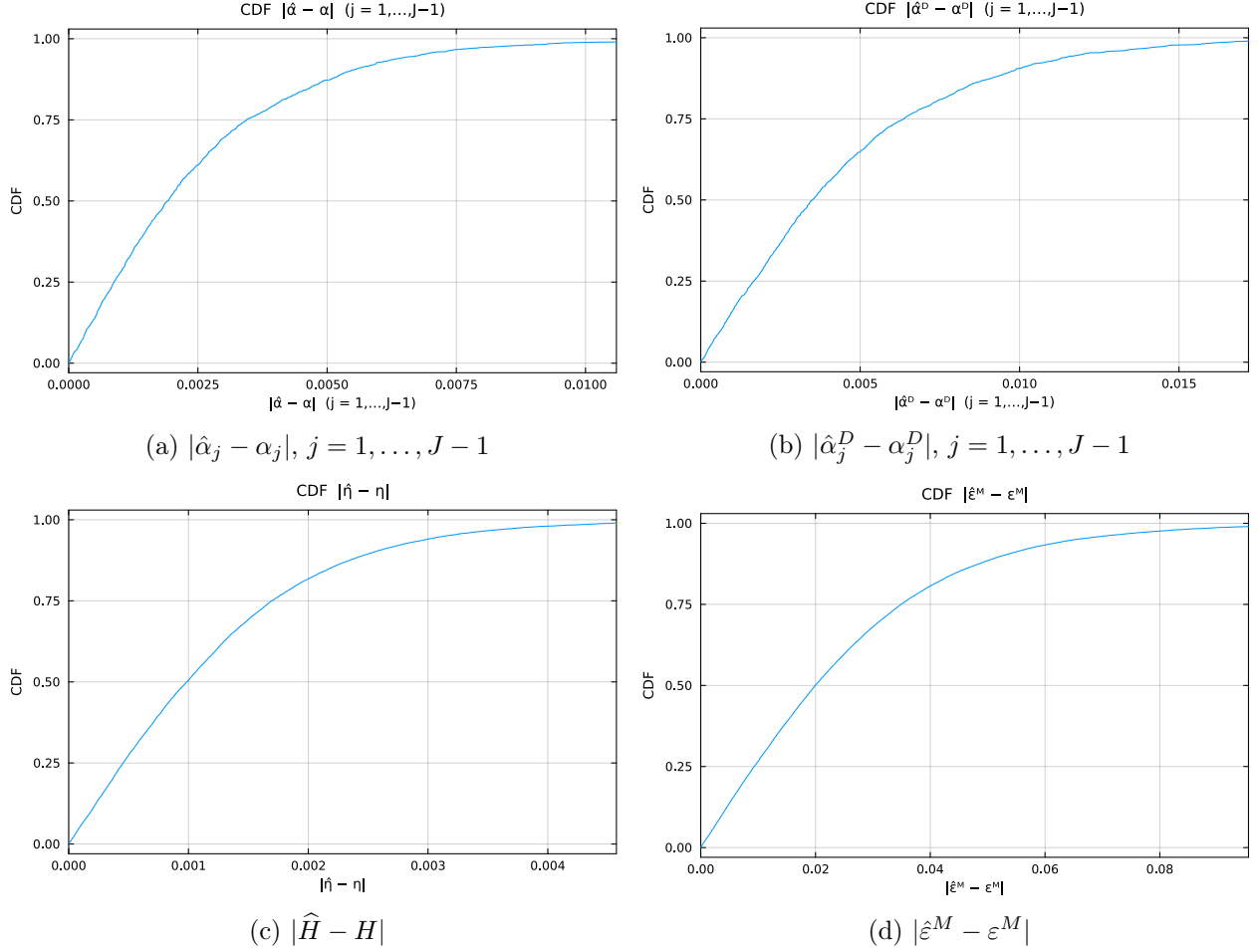


Figure 1: Absolute error CDFs, $J = 20$, $T = 50$, $r = 4$, 100 repetitions. Panels (a) and (b) pool the $J - 1$ estimated elements only: α_J and α_J^D are recovered from the adding-up restrictions rather than estimated, so their error is the accumulated error of the remaining elements and is excluded.

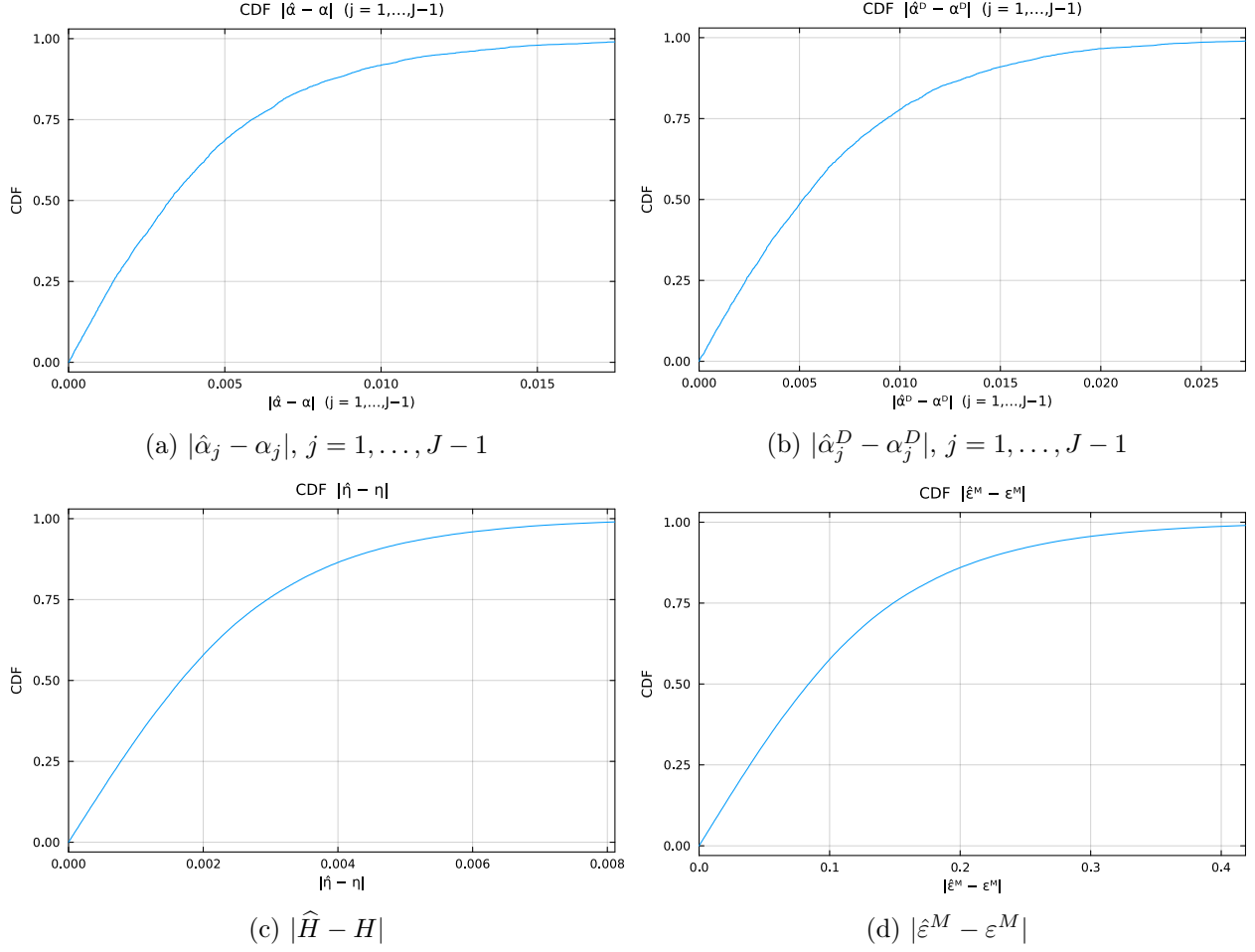
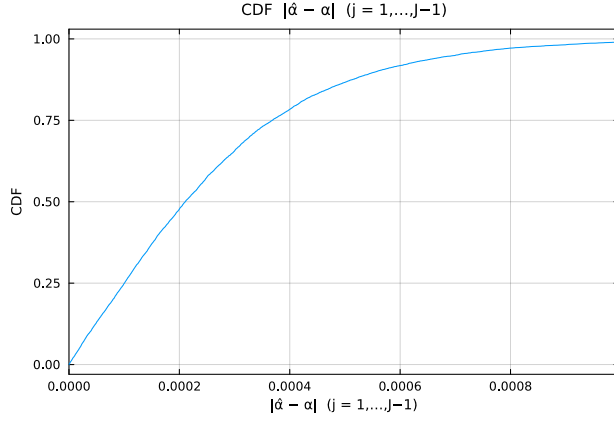
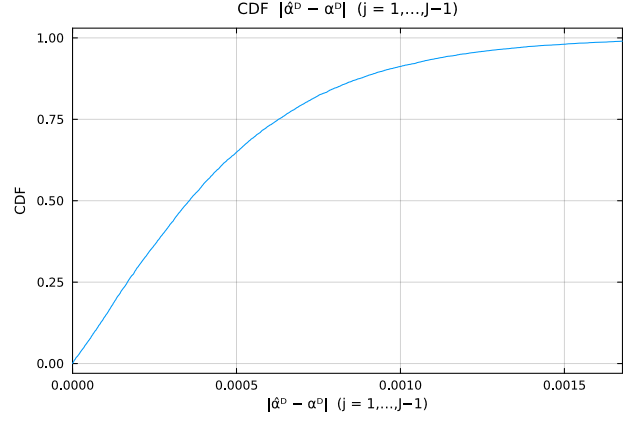


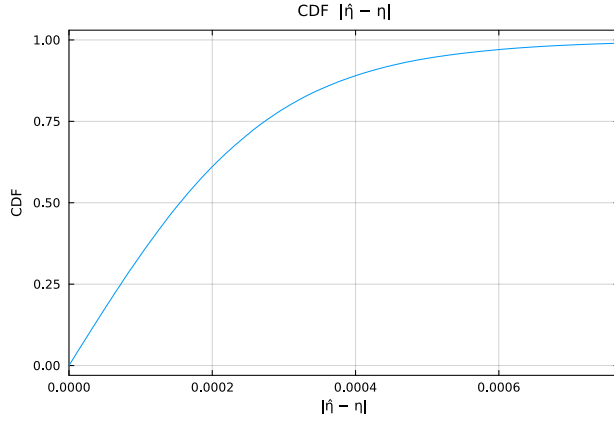
Figure 2: Absolute error CDFs, $J = 50$, $T = 50$, $r = 4$, 100 repetitions. Panels (a) and (b) pool the $J - 1$ estimated elements only: α_J and α_J^D are recovered from the adding-up restrictions rather than estimated, so their error is the accumulated error of the remaining elements and is excluded.



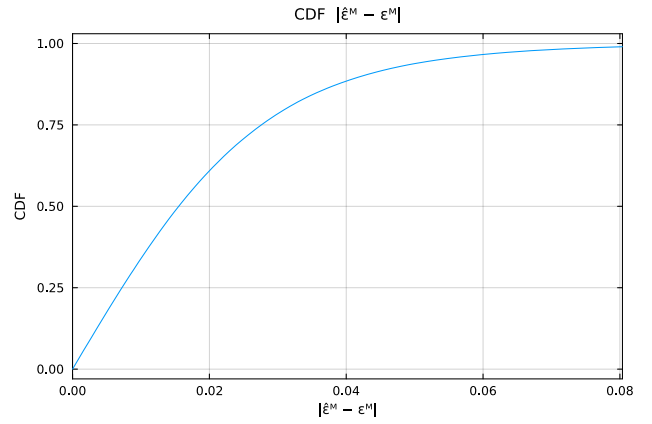
(a) $|\hat{\alpha}_j - \alpha_j|, j = 1, \dots, J-1$



(b) $|\hat{\alpha}_j^D - \alpha_j^D|, j = 1, \dots, J-1$



(c) $|\hat{H} - H|$



(d) $|\hat{\epsilon}^M - \epsilon^M|$

Figure 3: Absolute error CDFs, $J = 100$, $T = 50$, $r = 4$, 100 repetitions. Panels (a) and (b) pool the $J-1$ estimated elements only: α_J and α_J^D are recovered from the adding-up restrictions rather than estimated, so their error is the accumulated error of the remaining elements and is excluded.

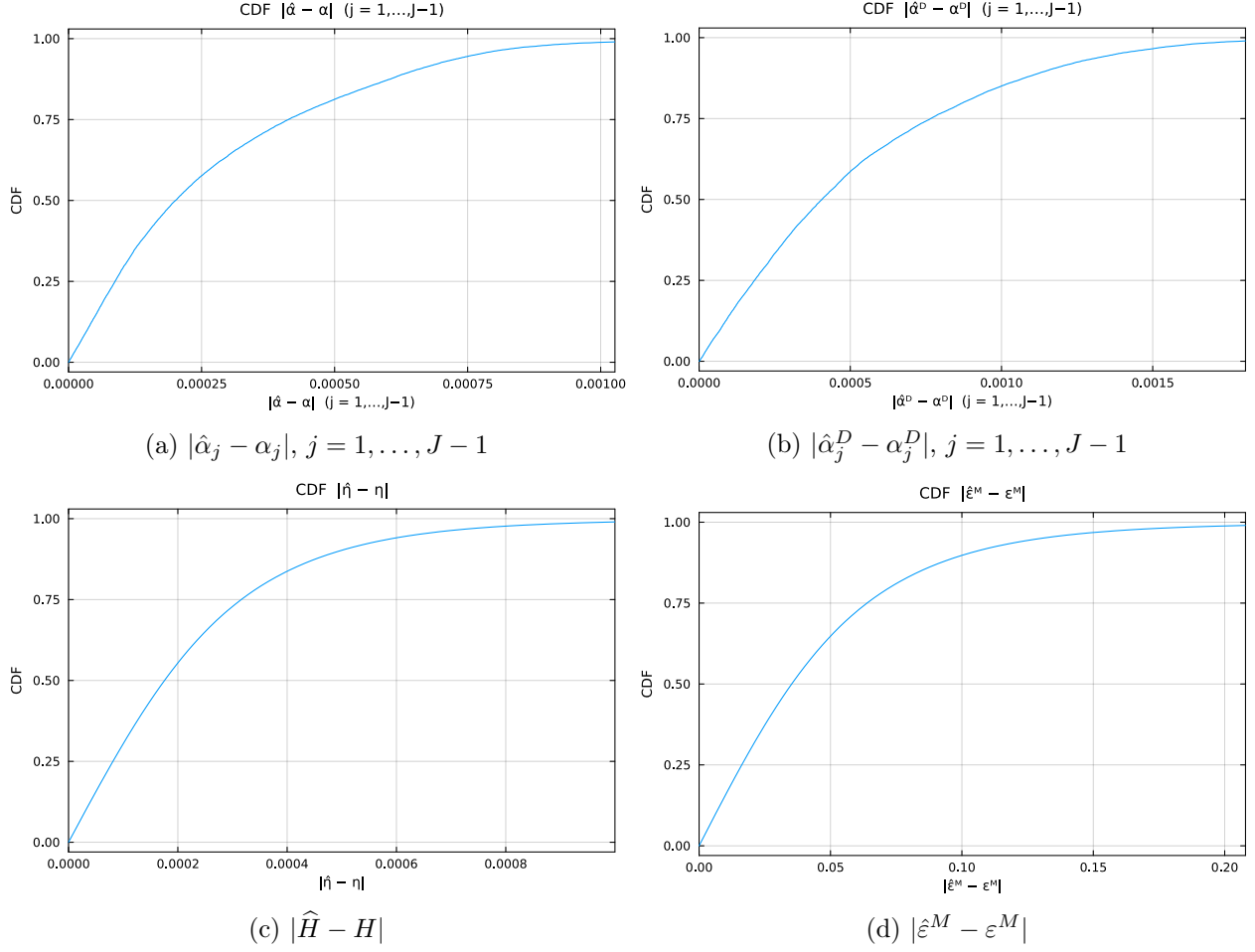


Figure 4: Absolute error CDFs, $J = 200$, $T = 50$, $r = 4$, 100 repetitions. Panels (a) and (b) pool the $J - 1$ estimated elements only: α_J and α_J^D are recovered from the adding-up restrictions rather than estimated, so their error is the accumulated error of the remaining elements and is excluded.

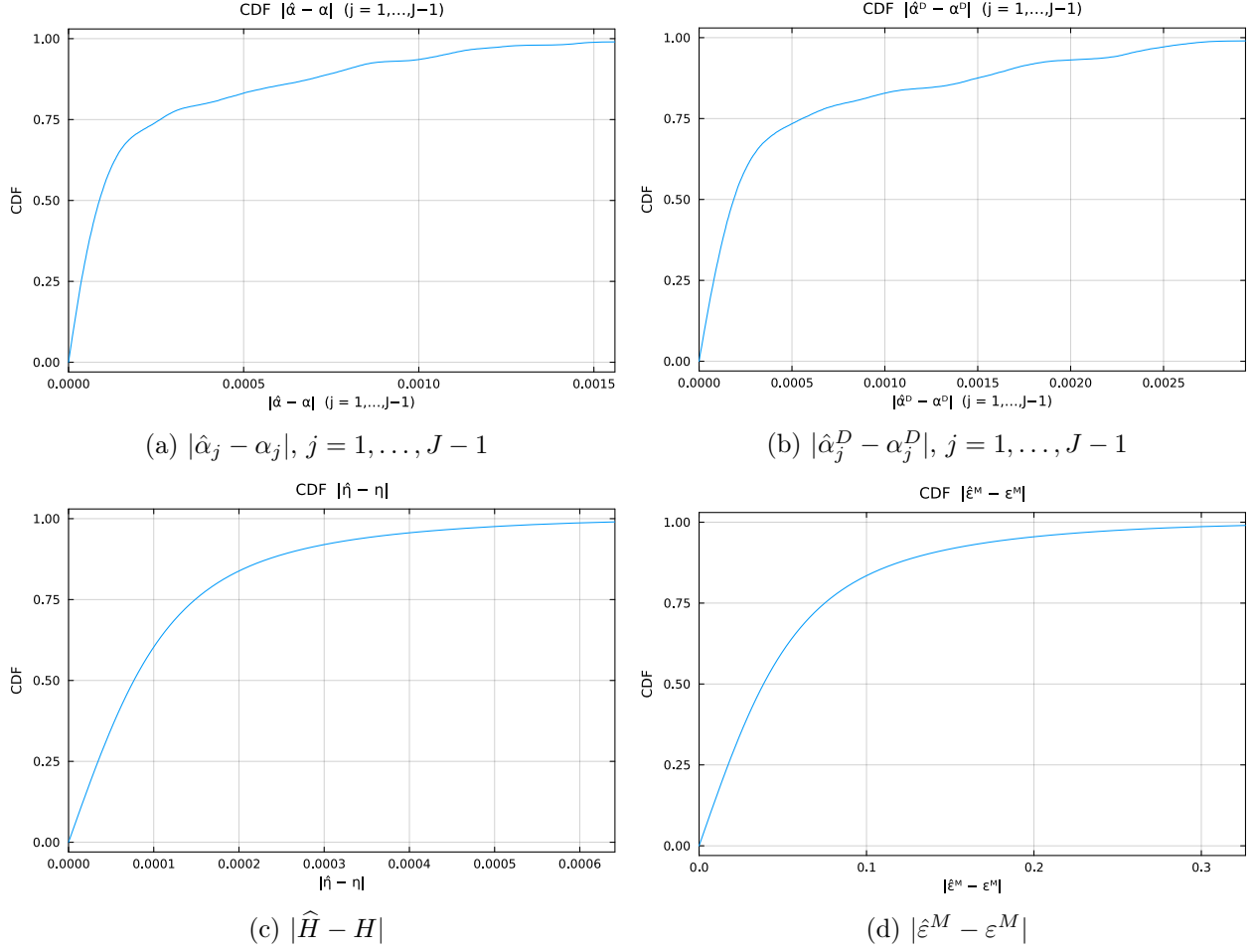


Figure 5: Absolute error CDFs, $J = 500$, $T = 50$, $r = 4$, 100 repetitions. Panels (a) and (b) pool the $J - 1$ estimated elements only: α_J and α_J^D are recovered from the adding-up restrictions rather than estimated, so their error is the accumulated error of the remaining elements and is excluded.