From Unstructured Data to Demand Counterfactuals: Theory and Practice*

Timothy Christensen† Giovanni Compiani‡

March 20, 2026

Abstract

Empirical models of demand for differentiated products rely on low-dimensional product representations to capture substitution patterns. These representations are increasingly proxied by applying ML methods to high-dimensional, unstructured data, including product descriptions and images. When proxies fail to capture the true dimensions of differentiation that drive substitution, standard workflows will deliver biased counterfactuals and invalid inference. We develop a practical toolkit that corrects this bias and ensures valid inference for a broad class of counterfactuals. Our approach applies to market-level and/or individual data, requires minimal additional computation, is efficient, delivers simple formulas for standard errors, and accommodates data-dependent proxies, including embeddings from fine-tuned ML models. It can also be used with standard quantitative attributes when mismeasurement is a concern. In addition, we propose diagnostics to assess the adequacy of the proxy construction and dimension. The approach yields meaningful improvements in predicting counterfactual substitution in both simulations and an empirical application.

Keywords: Unstructured data, demand estimation, bias correction, fine tuning, product embeddings, differentiated products.

*We are grateful to Steve Berry, Phil Haile, Francesca Molinari and seminar and conference participants at ASSA 2026, BU, Chicago Booth, Cornell, Duke, EIEF, Pittsburgh, Seoul National University, and Wisconsin. This material is based upon work supported by the National Science Foundation under Award No. 2521471 (Christensen).

†Department of Economics, Yale University.

‡Booth School of Business, University of Chicago.

1 Introduction

Many questions in economics and other social sciences require researchers to estimate demand for differentiated products. A common strategy is to estimate discrete-choice models which specify the utility of a product as a function of the product's price and other observable attributes, often allowing for rich types of consumer heterogeneity (Berry et al. (1995, 2004), henceforth BLP). Among other applications, this approach has been used to study the impact of horizontal mergers (Nevo, 2000), new product launches (Hausman, 1994; Petrin, 2002), trade policy (Goldberg, 1995), school choice (Bayer et al., 2007; Neilson, 2017), two-sided markets (Fan, 2013; Lee, 2013), and the evolution of markups over time (Grieco et al., 2024).

The success of demand models in predicting (counterfactual) quantities of interest (*counterfactuals* hereafter) hinges on their ability to capture substitution patterns. Doing so requires using product attributes as model inputs that correctly reflect the underlying dimensions of differentiation. Choosing the correct attributes to use as model inputs poses a fundamental measurement challenge (Berry and Haile, 2021). First, consumer choices are often driven by hard-to-quantify characteristics, such as visual design, user friendliness, or style. In these cases, a growing literature shows that product images, descriptions, and reviews contain valuable information to capture substitution patterns (Compiani et al., 2025; Han and Lee, 2025; Lee, 2025). Consumer surveys may also provide measures of product differentiation (Magnolfi et al., 2025). To use these high-dimensional, unstructured data in demand models, researchers often transform them into lower-dimensional numerical variables, or *embeddings*, using machine learning (ML) methods. Second, even numeric attributes could be mismeasured (e.g., Nevo, 2001; Allcott and Wozny, 2014), or could be high-dimensional and collinear, requiring dimension-reduction (e.g., Backus et al., 2021). In all of these cases, the variables used as inputs in the demand model are *proxies* for the true attributes that drive consumer choices. It is essential that these proxies adequately capture the true dimensions of differentiation: poor proxies can lead to biased estimates of demand model parameters and, in turn, biased counterfactuals.

In this paper, we propose a simple, post-estimation *bias correction* for counterfactuals. We take the naive estimator that treats the proxies as if they were the true dimensions of differentiation (as is implicitly done and reported in practice) and add a correction term designed to achieve two goals. First, it is chosen to mitigate

the bias in counterfactuals arising from mismeasurement of the true dimensions of differentiation. Second, it is chosen so that the bias-corrected estimator is *efficient*, meaning that it has the lowest possible asymptotic variance among a broad class of estimators. We also provide simple formulas for standard errors, making valid inference easy. In addition, we show how two simple *diagnostics* can be used to help assess the adequacy of different proxies in capturing substitution. Together, these diagnostics help guide the choice of *how many* and *which* proxies/attributes should be included in the model—practical questions that researchers need to answer in any instance. Answering these questions is further complicated by the fact that, unlike standard prediction problems, the counterfactual is not observed in the data.

We develop the bias corrections and diagnostics for two widely-used empirical frameworks. The first follows Berry et al. (1995, 2004): prices vary across markets, instrumental variables are used to address price endogeneity, and market-level data may be supplemented with individual choice data for a subset of markets. Many papers in industrial organization (IO) and fields using IO tools fit in this category. The second framework consists of models estimated on individual choice data with product-level fixed effects, which are very common in marketing applications (see Dubé and Rossi (2019) for a review). While we illustrate our approach for workhorse specifications (e.g., mixed logit with normal random coefficients), the method does not rely on specific parametric functional forms.

The bias corrections and diagnostics are computationally light and integrate easily into the standard demand estimation workflow. The bias corrections take as inputs the naive parameter estimates, which treat the proxies as the true dimensions of differentiation, and require neither bootstrapping nor any optimization. Similarly, the diagnostics are Lagrange Multiplier (LM) statistics evaluated at the naive estimates. All bias corrections, standard errors, and diagnostics admit closed-form expressions. These involve first derivatives of choice probabilities and counterfactuals, which are easily computed using automatic differentiation.

The key insight underlying our approach is that mismeasurement of the true dimensions of differentiation using proxies induces a form of model misspecification, as distinct from a measurement error problem.[1] We address this misspecification by

[1]In our setting, the relevant unit of observation is the market and/or individual level, whereas mismeasurement occurs at the product level. By contrast, mismeasurement is at the observation level in a standard measurement error problem. Misspecification and measurement error both cause bias, but do so for different reasons and require different corrections.

reparameterizing the model with a composite parameter that captures how proxies interact with structural parameters to affect utilities. Different proxies correspond to different values of the composite parameter. Framing the model in this way allows us to target counterfactuals using standard two-step estimation methods, where the first-step estimator corresponds to the composite parameter value pinned down by the proxies and naive parameter estimates.

An advantage of this approach is that it allows us to correct bias while remaining agnostic about the form of mismeasurement. This is particularly valuable because proxies are often obtained via black-box ML models, making it difficult to justify specific assumptions on the nature of mismeasurement. Importantly, the approach accommodates proxies that depend on the choice data, such as when they are obtained by *fine tuning* ML models on the same data that is used to estimate the demand model. For instance, researchers may fine tune neural networks or LLMs to obtain embeddings of product descriptions and images that better fit the observed substitution patterns than those produced by off-the-shelf algorithms. Moreover, because we correct how proxies and structural parameters jointly affect utility rather than the proxies themselves, our approach does not require practitioners to take a stand on the units of the proxies and/or true dimensions of differentiation. This is especially important for proxies for hard-to-quantify characteristics like visual design or user friendliness that lack natural units of measurement.

Simulations confirm that the bias correction improves performance for a range of levels of mismeasurement of the true dimensions of differentiation. Specifically, the corrected estimator has lower bias and lower variance than the naive estimator across all levels of mismeasurement. The bias correction leads to slightly higher variance in the knife-edge case in which the proxies perfectly capture differentiation, but the efficiency loss is small.[2] Simulations also confirm that our diagnostics convey useful information for selecting which proxies to use when estimating counterfactuals.

Finally, using the experimental data from Compiani et al. (2025), we show that the bias correction materially improves the model's ability to predict counterfactual choices following product removals. To this end, we leverage the fact that the data features both consumers' first and second choices. We estimate model parameters

[2]The fact that there is an efficiency loss in this knife-edge case is to be expected: the naive approach maintains the assumption that the proxies are measured without error, whereas the corrected estimator does not. What is surprising is that the efficiency loss is relatively small.

on the first choice data alone, then compare how well the naive and bias-corrected estimators predict second choices. The second-choice data provide a ground truth to assess the effectiveness of our approach. The bias correction meaningfully improves the model's ability to predict a product's closest substitute, improving the hit rate from 40% to 70% for our preferred specification. Further, our diagnostics correctly identify the set of proxies that perform best at the counterfactual prediction task, indicating that they can be valuable tools for practitioners. We show that the key mechanism behind the improvement is that the correction helps the model break away from logit-like IIA substitution patterns—where substitution is driven by market shares rather than product similarity—by restoring the role of product proximity in the embedding space.

We emphasize that our approach is also helpful for practitioners using standard numeric attributes. As noted above, mismeasurement may be a concern even in this case, particularly when dimension-reduction methods are used to shrink the attribute set. Our bias corrections provide a practical remedy. Further, the choice of which attributes to include is generally ad hoc even with numeric attributes. Our diagnostics help guide practitioners in making these decisions. Beyond mismeasurement concerns, our approach yields easy-to-compute, efficient estimators of counterfactuals (even when model parameters are estimated inefficiently) across many empirical settings, including combined market-level and microdata (e.g., Petrin, 2002; Berry et al., 2004, and many subsequent works). Our standard-error formulas also allow easy inference without bootstrapping. To the best of our knowledge, these contributions are new.[3]

Our approach is related to double/debiased ML (DML), which has recently been used in single-equation demand estimation with unstructured data (Bach et al., 2024). Both aim to estimate a target parameter in the presence of nuisance parameters. In our setting, the target is the counterfactual and the nuisance are both the latent dimensions of differentiation, which are "estimated" using proxies, and the demand model parameters.[4] Standard DML methods typically require models for the nuisance

[3] Grieco et al. (2025) study efficient estimation of model parameters in mixed logit models with combined market-level and microdata. Our focus is instead on efficient estimation of counterfactuals. For counterfactuals that depend on data moments in addition to model parameters (e.g., average welfare and average price elasticity), efficient estimators of model parameters do not necessarily lead to efficient estimators of counterfactuals (Brown and Newey, 1998; Ai and Chen, 2012).

[4] In contrast, Bach et al. (2024) treats the embeddings as perfect proxies for product attributes. Correspondingly, it uses DML to correct the estimation of nuisance functions as for partially linear regression, not to correct for mismeasurement of product attributes.

parameters and access to the data used to estimate them. In contrast, our approach accommodates proxies that are the outputs of black-box ML models trained on data to which the researcher might have limited to no access. To do so, we reparameterize the model via a composite parameter and rely on standard two-step estimation methods, including an orthogonalization step which shares similarities with DML.

A recent literature recognizes that naively treating ML-generated variables as data leads to measurement-error bias and develops corrections for it. But as noted above, the problem we study is one of model misspecification rather than measurement error, since the unit of observation (markets and/or individuals) is different from the unit of mismeasurement (products). Moreover, almost all strategies in this literature rely on validation data linking ML-generated variables and their ground-truth values.[5] In our setting, however, the true dimensions of differentiation are latent and can at best be only imperfectly proxied via survey data, rendering these methods inapplicable. One exception is Battaglia et al. (2024) who develop analytical bias corrections without validation data, but their approach is specific to linear regression.

The remainder of the paper is structured as follows. Section 2 presents our bias corrections and diagnostics for BLP-type models, while Section 3 does the same for models with individual-level choice data and product fixed effects. Each section first presents the model, develops the bias corrections and diagnostics, and concludes with a practitioner's guide detailing the steps involved and giving practical recommendations. Simulations and the empirical application are presented in Sections 4 and 5, respectively, with additional empirical results deferred to Appendix B. Section 6 presents all theoretical results while all proofs are presented in Appendix A.

2 Case 1: Endogenous Prices

We first consider a setting where prices vary at the market level and identification is achieved through instruments.

[5] See, e.g., Fong and Tyler (2021); Allon et al. (2023); Angelopoulos et al. (2023); Egami et al. (2023); Zhang et al. (2023); Carlson and Dell (2025) and references therein. These works build on an earlier literature on auxiliary data (Chen et al., 2005, 2008).

2.1 Model and Data

Following an established literature (Berry and Haile, 2014; Freyberger, 2015), we assume that the researcher has data from a large number T of markets in which (subsets of) J goods are sold.[6] In addition to the outside option (denoted by 0), each market t features products $\mathcal{J}_t \subseteq \{1, \ldots, J\}$, for which the researcher has access to data on prices $p_t = (p_{jt})_{j \in \mathcal{J}_t}$, exogenous product attributes $x_t = (x_{jt})_{j \in \mathcal{J}_t}$, and market shares $s_t = (s_{jt})_{j \in \mathcal{J}_t}$. Consumer choices are also driven by unobserved quality levels $\xi_t = (\xi_{jt})_{j \in \mathcal{J}_t}$. The model predicts market shares as a function of p_t, x_t, ξ_t, and a parameter vector θ:

$$s_{jt} = \sigma_j(p_t, x_t, \xi_t; \theta), \quad j \in \mathcal{J}_t. \tag{1}$$

Prices p_t are endogenous and may be correlated with the unobservables ξ_t. To address this, we rely on a vector of instrumental variables $w_t = (w_{jt})_{j \in \mathcal{J}_t}$ that satisfy

$$\mathbb{E}[\xi_{jt} | z_{jt}] = 0, \quad j \in \mathcal{J}_t, \quad t = 1, \ldots, T, \tag{2}$$

where $z_{jt} \equiv (x_{jt}, w_{jt})$. We also accommodate "micro BLP" settings (Berry et al., 2004; Berry and Haile, 2024) where, in addition to the above, individual-level data on choices and demographics are available in a subset of markets. The microdata consists of choice indicators $d_{it} = (d_{ijt})_{j \in \mathcal{J}_t}$ taking the value 1 if i chose j in market t and 0 otherwise, and demographic variables y_{it} that vary at the consumer level, such as income, and/or $\bar{y}_{ijt}$ that vary at the product-consumer level, such as distance between a household's home and a school or a hospital. We assume the microdata is available for a fixed set of markets which, without loss, we label $t = 1, \ldots, \tau$.[7] We treat the microdata within each market t as a repeated cross section of size N_t.

This model subsumes many empirical specifications used in the literature. We provide two simple examples below to fix ideas.

Example 1 (BLP)**.** *The utility that individual i derives from good j in market t is*

$$\begin{aligned} u_{ijt} &= \beta_i' x_{jt} - \alpha_i p_{jt} + \xi_{jt} + \varepsilon_{ijt}, \qquad j \in \mathcal{J}_t, \\ u_{i0t} &= \varepsilon_{i0t}, \end{aligned} \tag{3}$$

[6] For simplicity, we assume that J is fixed. It is straightforward to extend our approach to asymptotic thought experiments where J grows slowly with the number of markets T.

[7] The case where microdata is present for all T markets is simpler and the associated derivations are available upon request.

where the ε_{ijt} are iid type 1 extreme value random variables. Market shares are

$$\sigma_j(p_t, \xi_t, x_t; \theta) = \int \frac{e^{\beta' x_{jt} - \alpha p_{jt} + \xi_{jt}}}{1 + \sum_{k \in \mathcal{J}_t} e^{\beta' x_{kt} - \alpha p_{kt} + \xi_{kt}}} dF(\alpha, \beta; \theta), \quad j \in \mathcal{J}_t, \tag{4}$$

for some parametric distribution F.

Example 2 (Micro BLP)**.** *The utility is specified as:*

$$\begin{aligned} u_{ijt} &= \beta_i' x_{jt} - \alpha_i p_{jt} + (x_{jt}, p_{jt})' \Pi y_{it} + \pi' \bar{y}_{ijt} + \xi_{jt} + \varepsilon_{ijt}, \qquad j \in \mathcal{J}_t, \\ u_{i0t} &= \varepsilon_{i0t}, \end{aligned} \tag{5}$$

One can compute micro-moments using the following expression for individual choice probabilities:

$$\begin{aligned} &Pr(d_{ijt} = 1 | y_{it}, \bar{y}_{it}, p_t, \xi_t, x_t; \theta) \\ &\qquad = \int \frac{e^{\beta' x_{jt} - \alpha p_{jt} + (x_{jt}, p_{jt})' \Pi y_{it} + \pi' \bar{y}_{ijt} + \xi_{jt}}}{1 + \sum_{k \in \mathcal{J}_t} e^{\beta' x_{kt} - \alpha p_{kt} + (x_{kt}, p_{kt})' \Pi y_{it} + \pi' \bar{y}_{ikt} + \xi_{kt}}} dF(\alpha, \beta; \theta). \end{aligned} \tag{6}$$

The expression for market shares is obtained by integrating (6) over the (known) distribution of $(y_{it}, \bar{y}_{it})$ for $\bar{y}_{it} = (\bar{y}_{ijt})_{j \in \mathcal{J}_t}$.

So far the model is standard. Our point of departure is to partition

$$x_{jt} \equiv (\bar{x}_{jt}, e_j),$$

where $\bar{x}_{jt}$ is a vector of conventional observed product attributes, such as product size, and e_j is an r-vector of product attributes, such as visual design or user friendliness, that are difficult to capture using standard numeric data. Accordingly, we treat $e = (e_j')_{j=1}^J$ as known to the consumer but latent to the researcher. What is available to the researcher are proxies $\tilde{e} = (\tilde{e}_j')_{j=1}^J$ for the true underlying e. Note that e does not vary across markets, consistent with the fact that difficult-to-quantify characteristics such as visual design or user friendliness are often fixed product characteristics.

In the leading case we study, the econometrician observes unstructured data U_j and computes a low-dimensional representation $\tilde{e}_j$ of U_j, often referred to as an embedding, via ML methods. In this scenario, the embeddings $\tilde{e}_j$ act as proxies for the true latent e_j. To maximize generality, we stay agnostic on the form that U_j takes. It

could be text (product descriptions and reviews), images, audio/video components, or a combination thereof (Compiani et al., 2025; Han and Lee, 2025), or consumer preferences inferred from surveys (Magnolfi et al., 2025). Similarly, we are agnostic on the ML method used to compute $\tilde{e}_j$.

Unlike prior work, we wish to account for the fact that proxies $\tilde{e}_j$ are *not* the ground truth but rather are approximations to the true latent e_j. Different ML methods correspond to different approximations and produce different biases in downstream estimates of counterfactuals. Our first main goal is to develop estimators of model parameters and counterfactuals that are immune to this bias. Another goal is to shed light on what a "good" proxy might look like from the perspective of estimation and inference on counterfactuals. This objective is fundamentally different from the standard problem of choosing proxies for a prediction problem, since in our case the counterfactual is not observed in the data.

Remark 1. *While we focus on embeddings computed from unstructured data, our approach may be used more generally to correct bias from mismeasurement of any product attributes that do not vary across markets (e.g., the "mushiness" of cereal hand-coded by Nevo (2001)). In these scenarios, $\bar{x}_{jt}$ represents the attributes that are not mismeasured and $\tilde{e}_j$ represents the proxies for the true latent attributes, e_j.*

2.2 Bias-Corrected Counterfactuals

We consider a broad class of counterfactuals that can be written as

$$\kappa = \mathbb{E}[k(p_t, \xi_t, \bar{x}_t, e; \theta)], \tag{7}$$

where the expectation is over the distribution of $(p_t, \xi_t, \bar{x}_t)$ across markets t. For instance, κ might represent an average price elasticity, average equilibrium price or consumer welfare measure (possibly after a counterfactual change on the supply side). Expression (7) also subsumes counterfactuals for a specific market, such as the price elasticity at a given $(p, \xi, \bar{x})$. In this case, k is a deterministic function of θ and the expectation becomes redundant. As we discuss below, κ could also represent certain elements of θ, such as the average price coefficient.

Given the observed aggregate data and a candidate set of proxies $\tilde{e}$, estimation

typically proceeds using GMM based on the moment[8]

$$\frac{1}{T}\sum_{t=1}^{T} Z_t \hat{\xi}_t(s_t, p_t, \bar{x}_t, \tilde{e}; \theta),$$

where $\hat{\xi}_t(s_t, p_t, \bar{x}_t, e; \theta) = (\hat{\xi}_j(s_t, p_t, \bar{x}_t, e; \theta))_{j \in \mathcal{J}_t}$ is defined implicitly via

$$s_{jt} = \sigma_j(p_t, \hat{\xi}_t, \bar{x}_t, e; \theta), \quad j \in \mathcal{J}_t, \tag{8}$$

and Z_t is a $\dim(z) \times |\mathcal{J}_t|$ matrix whose columns contain z_{jt} for $j \in \mathcal{J}_t$. Here z_{jt} may be the original instruments in (2) as well as transformations thereof, such as when sieves are used to approximate optimal instruments. When the researcher also has microdata, the GMM criterion can be combined with a minimum-distance criterion based on the micro moments (as in, e.g., Conlon and Gortmaker, 2025). Let $\bar{m}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} m_{it}$, where $m_{it} = m(y_{it}, \bar{y}_{it}, d_{it})$ is a known function of demographic and choice data for individual i in market t. Let $m(p_t, \xi_t, \bar{x}_t, e; \theta)$ denote the model-implied expectation of m_{it} conditional on the market-level data:

$$m(p_t, \xi_t, \bar{x}_t, e; \theta) = \mathbb{E}[m_{it} | p_t, \xi_t, \bar{x}_t, e].$$

Thus

$$\bar{m}_t - m(p_t, \hat{\xi}_t(s_t, p_t, \bar{x}_t, \tilde{e}; \theta), \bar{x}_t, \tilde{e}; \theta), \quad t = 1, \ldots, \tau,$$

give an additional set of micro-moments to match when estimating θ.

Given an estimate $\hat{\theta}$ of θ, the counterfactual κ is usually estimated as

$$\hat{\kappa} = \frac{1}{T}\sum_{t=1}^{T} k(p_t, \hat{\xi}_t(s_t, p_t, \bar{x}_t, \tilde{e}; \hat{\theta}), \bar{x}_t, \tilde{e}; \hat{\theta}). \tag{9}$$

We call this the *naive estimator* of κ since it does not account for the fact that the proxies $\tilde{e}$ might differ from the true latent attributes e. This mismeasurement has the potential to affect the estimator via two channels: (i) directly, since $\tilde{e}$ is an argument of k, and (ii) indirectly through both $\hat{\theta}$ and $\hat{\xi}_t$.

We now introduce a bias-correction procedure that is designed to mitigate the

[8] With slight abuse of notation, we now write σ_j as functions of $\bar{x}_t$ and e, with the understanding that σ_j depends on e only through $(e_j)_{j \in \mathcal{J}_t}$, and similarly for $\hat{\xi}_t$.

bias from using $\tilde{e}$ in place of the true e. We begin by restricting attention to models in which the attributes e and model parameters θ enter choice probabilities (4) via a lower-dimensional *composite parameter*

$$\gamma \equiv \gamma(\theta, e).$$

Many models feature this property. We illustrate it in the BLP example.

Example 1 (continued). *We partition $\beta_i = (\beta_{\bar{x},i}, \beta_{e,i})$ and write the utilities as:*

$$u_{ijt} = \beta'_{\bar{x},i}\bar{x}_{jt} + \beta'_{e,i}e_j - \alpha_i p_{jt} + \xi_{jt} + \varepsilon_{ijt}, \quad j \in \mathcal{J}_t.$$

Suppose $\alpha_i \sim N(\bar{\alpha}, \sigma_\alpha^2)$, $\beta_{\bar{x},i} \sim N(\bar{\beta}_{\bar{x}}, \Sigma_{\bar{x}})$, $\beta_{e,i} \sim N(\bar{\beta}_e, \Sigma_e)$, and α_i, $\beta_{\bar{x},i}$ and $\beta_{e,i}$ are independent. Then,

$$\theta = \left(\bar{\alpha}, \sigma_\alpha, \bar{\beta}_{\bar{x}}, \bar{\beta}_e, l(\Sigma_{\bar{x}}), l(\Sigma_e)\right),$$

where $l(\Sigma)$ stacks the lower-triangular entries of the Cholesky factor of Σ into a vector.[9] *Note that e_j only enter via $\beta'_{e,i}e_j$. Collecting $\beta'_{e,i}e_j$ across products, we have $e\beta_{e,i} \sim N(e\bar{\beta}_e, e\Sigma_e e')$, where $e\Sigma_e e'$ has rank $r \leq J$ because e is $J \times r$. Hence,*

$$\gamma(\theta, e) = \left(\bar{\alpha}, \sigma_\alpha, \bar{\beta}_{\bar{x}}, e\bar{\beta}_e, l(\Sigma_{\bar{x}}), l_r(e\Sigma_e e')\right),$$

where l_r stacks the lower-triangular entries of the rank-r Cholesky factor of $e\Sigma_e e'$.[10] *For instance, when both $\bar{x}_j$ and e_j are scalars and $J = 2$, we have*

$$\gamma(\theta, e) = \left(\bar{\alpha}, \sigma_\alpha, \bar{\beta}_{\bar{x}}, e_1\bar{\beta}_e, e_2\bar{\beta}_e, \sigma_{\bar{x}}, \sigma_e|e_1|, \sigma_e e_2 \text{sign}(e_1)\right),$$

where $\sigma_{\bar{x}}$ and σ_e are the standard deviations of the random coefficients on $\bar{x}_j$ and e_j.

As can be seen from this example, parameters that do not interact with e are left unchanged, as is the case for the average price coefficient $\bar{\alpha}$, for instance. For the remaining components that interact with e, we expand the parameter space to capture the effect of joint shifts in e (as, for instance, when $\tilde{e}$ is used in place of e) and/or θ. A similar reparameterization for Example 2 is provided in Appendix C.

[9] If $\Sigma_{\bar{x}}$ and/or Σ_e are diagonal, then we replace $l(\Sigma_{\bar{x}})$ and/or $l(\Sigma_e)$ with vectors containing their diagonal entries.

[10] As $e\Sigma_e e'$ is $J \times J$ with rank r, its rank-r Cholesky factor is the unique $J \times r$ matrix L whose above-diagonal entries are all zeros and whose diagonal entries are all positive, such that $LL' = e\Sigma_e e'$.

This reparameterization allows us to simplify notation as follows. First, we note that the right-hand side of (8) depends on (θ, e) only via $\gamma(\theta, e)$. Thus we write $\hat{\xi}_{jt}(\gamma(\theta, e)) = \hat{\xi}_j(s_t, p_t, \bar{x}_t, e; \theta)$ for $j \in \mathcal{J}_t$ (suppressing dependence on $s_t, p_t, \bar{x}_t$) and let $\hat{\xi}_t(\gamma(\theta, e)) = (\hat{\xi}_{jt}(\gamma(\theta, e)))_{j \in \mathcal{J}_t}$. We similarly restrict attention to counterfactuals that depend on (θ, e) only via $\gamma(\theta, e)$ and write $k_t(\gamma(\theta, e)) = k(p_t, \hat{\xi}_t(\gamma(\theta, e)), \bar{x}_t, e; \theta)$. This includes many counterfactuals of interest, such as elasticities with respect to prices or $\bar{x}$, equilibrium prices, and welfare changes associated with changes in prices or $\bar{x}$. It precludes quantifying the effect of changes in the latent attributes e or measuring heterogeneity in preferences for e, but these are of little meaning when e has no natural scale or interpretation. When microdata are also available, we note that the choice probabilities in (6) depend on (θ, e) only via $\gamma(\theta, e)$ and write $m_t(\gamma(\theta, e)) = m(p_t, \hat{\xi}_t(\gamma(\theta, e)), \bar{x}_t, e; \theta)$. Finally, we let $\hat{\gamma} = \gamma(\hat{\theta}, \tilde{e})$ denote the value of γ at the estimated structural parameters $\hat{\theta}$ using the candidate proxies $\tilde{e}$.

With this notation, we can define the *bias-corrected estimator*

$$\hat{\kappa}_{bc} = \frac{1}{T} \sum_{t=1}^{T} \big(k_t(\hat{\gamma}) - \hat{c}' Z_t \hat{\xi}_t(\hat{\gamma})\big) + \sum_{t=1}^{\tau} \hat{d}_t' \big(\bar{m}_t - m_t(\hat{\gamma})\big), \tag{10}$$

where $\hat{c}$ is a $\dim(z) \times 1$ vector of weights for the aggregate moments and $\hat{d}_1, \ldots, \hat{d}_\tau$ are $\dim(m) \times 1$ vectors of weights for the micro moments. Without microdata, the bias-corrected estimator is simply

$$\hat{\kappa}_{bc} = \frac{1}{T} \sum_{t=1}^{T} \big(k_t(\hat{\gamma}) - \hat{c}' Z_t \hat{\xi}_t(\hat{\gamma})\big). \tag{11}$$

We give closed-form expressions for $\hat{c}$ and $\hat{d}_1, \ldots, \hat{d}_\tau$ below. The bias corrections are easy to implement: they simply take the naive estimator $\frac{1}{T} \sum_{t=1}^{T} k_t(\hat{\gamma})$ and add a weighted average of the estimation moments. As such, they require minimal computation beyond what is needed to estimate model parameters θ in the first place.

The idea behind (10) is to choose the weights $\hat{c}$ and $\hat{d}_1, \ldots, \hat{d}_\tau$ so that $\hat{\kappa}_{bc}$ does not depend on $\hat{\gamma}$ to first order.[11] This means that, to first order, $\hat{\kappa}_{bc}$ behaves like the right-hand side of (10) with $\hat{\gamma}$ replaced by the true value $\gamma_0 = \gamma(\theta_0, e_0)$, where θ_0 are the true structural parameters and e_0 are the true latent attributes. In doing so, this

[11] There is a long tradition of using corrections such as these in two-step estimation. See, e.g., Andrews (1994a) and Newey (1994). Of course, similar debiasing ideas underlie the DML literature.

purges the first-order effect of proxying e with $\tilde{e}$. In Section 6.1, we show that, as a result, the asymptotic distribution of $\hat{\kappa}_{bc}$ is centered around the true counterfactual κ_0 and does not depend on $\hat{\gamma}$. This has two important implications: first, κ_{bc} is immune to any bias arising from proxying e with $\tilde{e}$, and second, standard errors do not need to be corrected when $\tilde{e}$ is chosen in a data-dependent way, e.g., by fine tuning an ML model on choice data.

For the intuition, consider the case without microdata. A Taylor expansion of the naive estimator yields

$$\hat{\kappa} \approx \frac{1}{T}\sum_{t=1}^{T}\left(k_t(\gamma_0) + \frac{\partial k_t(\hat{\gamma})}{\partial \gamma'}(\hat{\gamma}-\gamma_0)\right).$$

We wish to eliminate the second problematic term depending on $\hat{\gamma}-\gamma_0$. To do so, we replicate the dependence of $\hat{\kappa}$ on $\hat{\gamma}-\gamma_0$ using the estimation moments. This means that the vector of weights $\hat{c}$ will be chosen so that

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial k_t(\hat{\gamma})}{\partial \gamma'} = \hat{c}'\left(\frac{1}{T}\sum_{t=1}^{T} Z_t \frac{\partial \hat{\xi}_t(\hat{\gamma})}{\partial \gamma'}\right).$$

There are many different weights $\hat{c}$ with this property; the weights introduced below are designed to minimize the asymptotic variance of $\hat{\kappa}_{bc}$. Substituting in the previous display and "undoing" the Taylor expansion, we get

$$\begin{aligned}\hat{\kappa} &\approx \frac{1}{T}\sum_{t=1}^{T}\left(k_t(\gamma_0) + \hat{c}' Z_t \frac{\partial \hat{\xi}_t(\hat{\gamma})}{\partial \gamma'}(\hat{\gamma}-\gamma_0)\right)\\ &= \frac{1}{T}\sum_{t=1}^{T}\left(k_t(\gamma_0) + \hat{c}' Z_t \hat{\xi}_t(\hat{\gamma}) - \hat{c}' Z_t \hat{\xi}_t(\gamma_0)\right).\end{aligned}$$

This suggests that to correct bias we want to adjust the naive estimator by subtracting $\frac{1}{T}\sum_{t=1}^{T}(\hat{c}' Z_t \hat{\xi}_t(\hat{\gamma}) - \hat{c}' Z_t \hat{\xi}_t(\gamma_0))$. The final (infeasible) term depending on γ_0 has mean zero by virtue of (2), so we drop it, leading to the corrected estimator (11). This correction therefore ensures that, to first order, $\hat{\kappa}_{bc}$ depends only on γ_0.

What assumptions are needed for this result? Besides standard regularity conditions, we require that the discrepancy between $\hat{\gamma}$ and the true value γ_0 not be too large relative to sampling error (see Section 6.1 for a discussion). Figure 2 shows that, in simulations, $\hat{\kappa}_{bc}$ has negligible bias up to moderate amounts of mismeasurement

(and thus moderate deviations of $\hat{\gamma}$ from γ_0), while for high amounts of mismeasurement (and thus large deviations of $\hat{\gamma}$ from γ_0), the bias of $\hat{\kappa}_{bc}$ is still well below that of the naive estimator. We also implicitly require that the e_j and $\tilde{e}_j$ have the same dimension r. In the next subsection, we provide two diagnostics to help researchers choose among proxies so that both these conditions are plausibly satisfied.

To introduce the expressions for the weights $\hat{c}$ and $\hat{d}_1, \ldots, \hat{d}_\tau$, let

$$\hat{V} = \frac{1}{T}\sum_{t=1}^{T}(Z_t\hat{\xi}_t(\hat{\gamma}))(Z_t\hat{\xi}_t(\hat{\gamma}))' - \bar{g}\bar{g}',$$
$$\hat{V}_t = \frac{T}{N_t}\left(\frac{1}{N_t}\sum_{i=1}^{N_t} m_{it}m_{it}' - \bar{m}_t\bar{m}_t'\right), \quad t = 1, \ldots, \tau,$$

denote the sample variance of the estimation moments, where $\bar{g} = \frac{1}{T}\sum_{t=1}^{T} Z_t\hat{\xi}_t(\hat{\gamma})$. Define

$$\hat{h} = \hat{k} - \hat{G}'\hat{V}^{-1}\hat{K}, \qquad \hat{H} = \hat{G}'\hat{V}^{-1}\hat{G} + \sum_{t=1}^{\tau}\hat{M}_t'\hat{V}_t^{-1}\hat{M}_t, \tag{12}$$

where $\hat{k} = \frac{1}{T}\sum_{t=1}^{T}\dot{k}_t(\hat{\gamma})$ is $\dim(\gamma)\times 1$, $\hat{K} = \frac{1}{T}\sum_{t=1}^{T} k_t(\hat{\gamma})Z_t\hat{\xi}_t(\hat{\gamma})$ is $\dim(z)\times 1$, $\hat{G} = \frac{1}{T}\sum_{t=1}^{T} Z_t\dot{\xi}_t(\hat{\gamma})$ is $\dim(z)\times\dim(\gamma)$, $\hat{M}_t = \dot{m}_t(\hat{\gamma})$ is $\dim(m)\times\dim(\gamma)$, and $\dot{k}_t(\gamma) = \frac{\partial k_t(\gamma)}{\partial\gamma}$, $\dot{\xi}_t(\gamma)' = \frac{\partial\hat{\xi}_t(\gamma)'}{\partial\gamma}$, and $\dot{m}_t(\gamma)' = \frac{\partial m_t(\gamma)'}{\partial\gamma}$ are $\dim(\gamma)\times 1$, $\dim(\gamma)\times J_t$, and $\dim(\gamma)\times\dim(m)$, respectively. The weights to plug into (10) are

$$\hat{c} = \hat{V}^{-1}(\hat{K} + \hat{G}\hat{H}^{-1}\hat{h}), \tag{13}$$

and

$$\hat{d}_t = \hat{V}_t^{-1}\hat{M}_t\hat{H}^{-1}\hat{h}, \quad t = 1, \ldots, \tau. \tag{14}$$

Without microdata, $\hat{d}_1 = \ldots = \hat{d}_\tau = 0$ and $\hat{H} = \hat{G}'\hat{V}^{-1}\hat{G}$.

We defer formal statements of the results sketched out above to Section 6.1, and instead highlight a few key properties of the corrected estimator.

Remark 2 (Easy to compute standard errors)**.** The asymptotic variance of the bias-corrected estimator can be estimated as follows:

$$\hat{V}_{bc} = \hat{s}_k^2 + \hat{c}'\hat{V}\hat{c} - 2\hat{c}'(\hat{K} - \bar{k}\bar{g}) + \sum_{t=1}^{\tau}\hat{d}_t'\hat{V}_t\hat{d}_t, \tag{15}$$

where $\hat{s}_k^2 = \frac{1}{T}\sum_{t=1}^T k_t(\hat{\gamma})^2 - \bar{k}^2$ and $\bar{k} = \frac{1}{T}\sum_{t=1}^T k_t(\hat{\gamma})$ are the sample variance and sample mean of $k_t(\hat{\gamma})$. Without microdata, the expression simplifies to

$$\hat{V}_{bc} = \hat{s}_k^2 + \hat{c}'\hat{V}\hat{c} - 2\hat{c}'(\hat{K} - \bar{k}\bar{g}). \tag{16}$$

In either case, standard errors for $\hat{\kappa}_{bc}$ are $\sqrt{\hat{V}_{bc}/T}$. Again, these are closed-form expressions involving quantities that are easy to compute given $\hat{\theta}$.

Remark 3 (Efficiency)**.** The weights $\hat{c}$ and $\hat{d}_1, \ldots, \hat{d}_\tau$ are chosen so that the asymptotic variance of $\hat{\kappa}_{bc}$ (and thus standard errors) are as small as possible: see Proposition 2 in Section 6.1 for a formal statement. Importantly, $\hat{\kappa}_{bc}$ remains efficient even if $\hat{\theta}$ is inefficient. Thus, there is no need to estimate θ using an optimal weighting and/or optimal instruments.

Remark 4 (Fine Tuning)**.** The bias correction in (10) and standard error formulas in Remark 2 allow the proxies $\tilde{e}$ to be sample-dependent. In the context of embeddings, this accommodates scenarios where an off-the-shelf algorithm has been fine tuned on the choice data to provide a better fit. Because $\hat{\kappa}_{bc}$ doesn't depend on $\tilde{e}$ to first order, there is no need to correct the standard errors for fine tuning.

2.3 Diagnostics

Next, we propose two diagnostics that practitioners can use to assess the suitability of a candidate set of proxies $\tilde{e}$. The first speaks to whether $\hat{\gamma} = \gamma(\hat{\theta}, \tilde{e})$ is sufficiently close to the truth $\gamma_0 = \gamma(\theta_0, e_0)$. The second addresses the question of whether the dimension of $\tilde{e}$ matches that of e_0. Both diagnostics are based on LM statistics evaluated at $\hat{\gamma}$, so they require minimal additional computation.

2.3.1 Diagnostic 1: Is $\hat{\gamma}$ Close To γ_0?

The bias correction is based on linearization and thus requires the discrepancy between $\hat{\gamma}$ and γ_0 to not be too large relative to sampling error, as discussed above. Here we show that a simple LM statistic can be used to validate this condition. This diagnostic is also helpful to guide the choice among sets of embeddings (e.g., embeddings obtained from various data source and/or ML model combinations).

The first diagnostic is

$$LM_1 = \|\sqrt{T}\hat{H}^{-1/2}\hat{S}\|^2, \tag{17}$$

where $\hat{S} = \hat{G}'\hat{V}^{-1}(\frac{1}{T}\sum_{t=1}^{T} Z_t\hat{\xi}_t(\hat{\gamma})) + \sum_{t=1}^{\tau} \hat{M}_t'\hat{V}_t^{-1}(m_t(\hat{\gamma}) - \bar{m}_t)$ represents the "score" at $\hat{\gamma}$ and $\hat{H}$ is given in (12). This diagnostic can be interpreted as the LM statistic in a test of the null hypothesis that γ_0 is in the set of composite parameters spanned by the candidate proxies $\tilde{e}$, ignoring the fact that $\tilde{e}$ is possibly stochastic.[12] Proposition 4 below shows that LM_1 behaves like $\|\sqrt{T}(\hat{\gamma} - \gamma_0)\|^2$ as the sample size grows large. In other words, researchers can validate the assumption that the discrepancy between $\hat{\gamma}$ and γ_0 is sufficiently small, as needed in Proposition 1, by checking whether LM_1 is below a threshold. In particular, for any sequence $C_T = o(T^{1/4})$, we have that $LM_1 \leq C_T^2$ implies $\|\hat{\gamma} - \gamma_0\| \leq \text{constant} \times C_T/\sqrt{T} = o(T^{-1/4})$ with probability approaching one. It can also be shown under a slight strengthening of the conditions of Proposition 4 that LM_1 can be used to bound $\|\tilde{e} - e_0\|$, providing a measure of how well the proxies $\tilde{e}$ capture the true latent attributes e driving consumer choices. This is especially useful as the true e_0 can never be observed. As a result, LM_1 also serves as a model-based criterion to target when fine tuning.

2.3.2 Diagnostic 2: Is The Dimension of $\tilde{e}$ Correct?

When estimating this type of model, practitioners also have to choose how many attributes to include. In our notation, this corresponds to choosing the dimension of $\tilde{e}$. This is particularly delicate in the cases where the candidate proxies don't have a natural economic interpretation, as is typically the case when they are obtained from black-box ML algorithms or by applying principal component analysis (PCA) to a rich set of numeric attributes. For example, in the application of Section 5, we use PCA to reduce the dimensionality of proxies obtained from pre-trained algorithms; the relevant question is then how many principal components to include in the model.

We provide guidance on this by again considering a diagnostic based on an LM statistic. The idea is to augment $\tilde{e}$ with a vector $\eta \in \mathbb{R}^J$ representing some excluded but potentially important product attributes and augment θ with an additional component $\psi \in \Psi$ representing coefficients on η. The second diagnostic is based on an LM statistic for the null that $\psi = 0$. Like LM_1, this diagnostic depends on $\hat{\theta}$ only and therefore requires minimal additional computation.

To introduce the diagnostic, we extend γ to $\zeta = (\gamma, \psi)$. With slight abuse of notation, we now write $\hat{\xi}_{jt}$ and m_t on this extended space as functions $\hat{\xi}_{jt}(\zeta; \eta)$ and

[12] Formally, a test of $\mathbb{H}_0 : \gamma_0 \in \Gamma(\tilde{e}) := \{\gamma(\theta, \tilde{e}) : \theta \in \Theta\}$ against the alternative $\mathbb{H}_1 : \gamma_0 \in \Gamma \setminus \Gamma(\tilde{e})$, where Θ and Γ are the parameter spaces for θ and γ, respectively.

$m_t(\zeta;\eta)$ of ζ and η, with the understanding that $\hat{\xi}_{jt}(\gamma) = \hat{\xi}_{jt}((\gamma,0);\eta)$ and $m_t(\gamma) = m_t((\gamma,0);\eta)$. Let

$$\hat{\lambda}(\eta) = \left(\frac{1}{T}\sum_{t=1}^{T} Z_t w_t(\hat{\gamma},\eta)\right)' \hat{V}^{-1}(I - \hat{G}(\hat{G}'\hat{V}^{-1}\hat{G})^{-1}\hat{G}'\hat{V}^{-1})\sqrt{T}\bar{g}, \tag{18}$$

$$\hat{\lambda}_t(\eta) = u_t(\hat{\gamma},\eta)'\hat{V}_t^{-1}(I - \hat{M}_t(\hat{M}_t'\hat{V}_t^{-1}\hat{M}_t)^{-1}\hat{M}_t'\hat{V}_t^{-1})\sqrt{T}(m_t(\hat{\gamma}) - \bar{m}_t), \tag{19}$$

where $w_t(\gamma,\eta) = (w_{jt}(\gamma,\eta)')_{j\in\mathcal{J}_t}$ is $|\mathcal{J}_t| \times \dim(\psi)$ and $u_t(\gamma,\eta)$ are $\dim(\psi) \times \dim(m)$, with

$$w_{jt}(\gamma,\eta) = \lim_{\psi\to 0} \frac{\partial \hat{\xi}_{jt}((\gamma,\psi);\eta)}{\partial\psi}, \quad j \in J_t, \qquad u_t(\gamma,\eta)' = \lim_{\psi\to 0} \frac{\partial m_t((\gamma,\psi);\eta)'}{\partial\psi}.$$

We take limits to deal with parameters that are at the boundary when $\psi = 0$, such as the variance of the random coefficients on η. For the intuition, the left-most terms in (18) and (19) are the Jacobian of the moments with respect to ψ. These can depend to first order on $\hat{\gamma}$. The second parts of these expressions eliminate this dependence with a similar correction to $\hat{\kappa}_{bc}$. We estimate the variance of $\hat{\lambda}(\eta)$ and $\hat{\lambda}_t(\eta)$ using

$$\hat{\Lambda}(\eta) = \left(\frac{1}{T}\sum_{t=1}^{T} Z_t w_t(\hat{\gamma},\eta)\right)' \hat{V}^{-1}(I - \hat{G}(\hat{G}'\hat{V}^{-1}\hat{G})^{-1}\hat{G}'\hat{V}^{-1}) \left(\frac{1}{T}\sum_{t=1}^{T} Z_t w_t(\hat{\gamma},\eta)\right), \tag{20}$$

$$\hat{\Lambda}_t(\eta) = u_t(\hat{\gamma},\eta)'\hat{V}_t^{-1}(I - \hat{M}_t(\hat{M}_t'\hat{V}_t^{-1}\hat{M}_t)^{-1}\hat{M}_t'\hat{V}_t^{-1})u_t(\hat{\gamma},\eta).$$

Finally, define

$$\hat{W}(\eta) = \left(\hat{\lambda}(\eta) + \sum_{t=1}^{\tau}\hat{\lambda}_t(\eta)\right)' \left(\hat{\Lambda}(\eta) + \sum_{t=1}^{\tau}\hat{\Lambda}_t(\eta)\right)^{-1} \left(\hat{\lambda}(\eta) + \sum_{t=1}^{\tau}\hat{\lambda}_t(\eta)\right).$$

Without microdata, $\hat{W}(\eta)$ simplifies to $\hat{W}(\eta) = \hat{\lambda}(\eta)'\hat{\Lambda}(\eta)^{-1}\hat{\lambda}(\eta)$. The statistic $\hat{W}(\eta)$ can be shown to behave like a $\chi^2_{\dim(\psi)}$ random variable under the null (i.e., when the true $\psi = 0$), but it depends on the nuisance parameter η. Thus, we define our second diagnostic as

$$LM_2 = \sup_{\eta\in S^J:\eta\perp C(\tilde{e})} \hat{W}(\eta), \tag{21}$$

where we take the supremum over η in the unit sphere (since the scale of η is not

important) that are orthogonal to the column span of $\tilde{e}$.

Following, e.g., Hansen (1996), critical values can be computed by simulation. Draw $(\varpi_t^*)_{t=1}^T, (\varpi_{i1}^*)_{i=1}^{N_1}, \ldots, (\varpi_{i\tau}^*)_{i=1}^{N_\tau}$ iid from a $N(0,1)$ distribution, and set

$$\hat{W}^*(\eta) = \left(\hat{\lambda}^*(\eta) + \sum_{t=1}^{\tau} \hat{\lambda}_t^*(\eta)\right)' \left(\hat{\Lambda}(\eta) + \sum_{t=1}^{\tau} \hat{\Lambda}_t(\eta)\right)^{-1} \left(\hat{\lambda}^*(\eta) + \sum_{t=1}^{\tau} \hat{\lambda}_t^*(\eta)\right),$$

where

$$\hat{\lambda}^*(\eta) = \left(\frac{1}{T}\sum_{t=1}^{T} Z_t w_t(\hat{\gamma}, \eta)\right)' \hat{V}^{-1}(I - \hat{G}(\hat{G}'\hat{V}^{-1}\hat{G})^{-1}\hat{G}'\hat{V}^{-1}) \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \varpi_t^* Z_t \hat{\xi}_t(\hat{\gamma})\right),$$

$$\hat{\lambda}_t^*(\eta) = u_t(\hat{\gamma}, \eta)'\hat{V}_t^{-1}(I - \hat{M}_t(\hat{M}_t'\hat{V}_t^{-1}\hat{M}_t)^{-1}\hat{M}_t'\hat{V}_t^{-1}) \left(\frac{1}{\sqrt{T}}\sum_{i=1}^{N_t} \varpi_{it}^*(m_t(\hat{\gamma}) - m_{it})\right).$$

For each collection of $N(0,1)$ draws, compute

$$LM_2^* = \sup_{\eta \in S^J : \eta \perp C(\tilde{e})} \hat{W}^*(\eta)^2.$$

Let $\hat{\xi}_{0.95}^*$ denote the 95th percentile of LM_2^* across a large number of independent draws. This quantity is easy to compute, as only the right-most terms in the expressions for $\hat{\lambda}^*$ and $\hat{\lambda}_t^*$ need to be recomputed for different draws and these terms do not depend on η. We reject the null that the dimension of $\tilde{e}$ is adequate if $LM_2 > \hat{\xi}_{0.95}^*$.

Example 1 (continued). *Let the random coefficient on η be $\delta_i = \psi_1 + \sqrt{\psi_2} Z_i$, where $Z_i \sim F_0$ has mean zero and unit variance, $\psi_1 \in \mathbb{R}$, and $\psi_2 \in [0, \infty)$. For simplicity, suppose $\bar{x}_t$ is empty and $|\mathcal{J}_t| = J$. Define the functions $\sigma_{jt}(\,\cdot\,; (\gamma, \psi), \eta)$ and $\varsigma_{jt}(\,\cdot\,; \gamma)$ from $\mathbb{R}^J$ to $\mathbb{R}$ by*

$$\sigma_{jt}(\xi_t; (\gamma, \psi), \eta) = \int \varsigma_{jt}(\xi_t + (\psi_1 + \sqrt{\psi_2} z)\eta; \gamma)\, dF_0(z),$$

$$\varsigma_{jt}(u; \gamma) = \int \frac{e^{-\alpha' p_{jt} + \beta' e_j + u_j}}{1 + \sum_{k \in \mathcal{J}_t} e^{-\alpha' p_{kt} + \beta' e_k + u_k}} dF(\alpha, \beta),$$

where we have suppressed dependence on p_t. Here $\hat{\xi}_{jt}(\zeta; \eta)$ solves $s_{jt} = \sigma_{jt}(\hat{\xi}; \zeta, \eta)$ for $j \in \mathcal{J}_t$. Evidently, $\hat{\xi}_{jt}(\gamma) = \hat{\xi}_{jt}((\gamma, 0); \eta)$. Using $\dot{\varsigma}_{jt}$ and $\ddot{\varsigma}_{jt}$ to denote the first and

second derivatives of ς_{jt} with respect to its first argument, we have

$$\lim_{\psi \to 0} \frac{\partial \sigma_{jt}(\hat{\xi}_t((\gamma,\psi);\eta);(\gamma,\psi),\eta)}{\partial \psi} = \begin{bmatrix} \eta' \dot{\varsigma}_{jt}(\hat{\xi}_t(\gamma);\gamma) \\ \frac{1}{2}\eta' \ddot{\varsigma}_{jt}(\hat{\xi}_t(\gamma);\gamma)\eta \end{bmatrix}.$$

It follows by the implicit function theorem that

$$w_t(\gamma,\eta) = -\left(\frac{\partial \sigma_t(\hat{\xi}_t(\gamma);(\gamma,0),\eta)}{\partial \xi'}\right)^{-1} \begin{pmatrix} \eta' \dot{\varsigma}_{jt}(\hat{\xi}_t(\gamma);\gamma) & \frac{1}{2}\eta' \ddot{\varsigma}_{jt}(\hat{\xi}_t(\gamma);\gamma)\eta \end{pmatrix}_{j \in \mathcal{J}_t},$$

where $\sigma_t(\xi_t;\zeta,\eta) = (\sigma_{jt}(\xi_t;\zeta,\eta))_{j\in\mathcal{J}_t}$. Plugging this into (18) and (20), one can compute $\hat{W}(\eta)$ and thus the LM_2 diagnostic.

2.4 Practitioner's Guide

Given a counterfactual of interest κ and a candidate set of proxies $\tilde{e}$:

1. Calculate the model parameter estimates $\hat{\theta}$ as usual, treating $\tilde{e}$ as the truth.
2. Compute weights $\hat{c}$ in (13) and, if microdata is available, weights $\hat{d}_t$ in (14).
3. Plug the weights in (10) to compute the corrected estimator $\hat{\kappa}_{bc}$. If only aggregate data is available, use (11) instead.
4. Compute standard errors for $\hat{\kappa}_{bc}$ using Remark 2.
5. To check whether a given set of proxies $\tilde{e}$ is adequate:
   (a) Compute the LM_1 statistic in (17). If it's below a threshold C_T^2, conclude that $\tilde{e}$ is sufficiently close to e_0. In Section 6.1.2, we motivate a threshold of $C_T^2 = \chi^2_{\dim(\gamma),0.95} \log T$ to deliver a rate of $\sqrt{(\log T)/T}$.
   (b) Compute the LM_2 statistic in (21). If it's below a threshold $\hat{\xi}^*_{0.95}$, conclude that $\tilde{e}$ is of adequate dimension. The bootstrap method in Section 2.3.2 can be used to compute $\hat{\xi}^*_{0.95}$.

2.5 What About Models with Standard Numeric Attributes?

Our approach also provides a data-driven way to robustly estimate counterfactuals and validate some of the assumptions implicitly made in the demand estimation lit-

erature in contexts where only standard numeric attributes are available. The typical workflow assumes that product attributes are measured without error and are of adequate dimension. Our bias correction allows practitioners to relax the assumption of correct measurement. To do so, the bias-corrected estimator can be implemented as above, where now $\tilde{e}$ simply represents the numeric attributes that may be mismeasured. The resulting bias-corrected estimator is robust to such mismeasurement. Similarly, our diagnostics may be used to choose among attributes and assess whether the dimension of a candidate set of attributes is adequate.

Even when mismeasurement is not a concern, our bias-corrected estimator and standard error formulas can be used to perform efficient inference on counterfactuals (see Proposition 2 below for a formal statement). In this case, there is no need to reparameterize the model to account for mismeasurement and both are implemented as described above with $\gamma \equiv \theta$. This approach offers a few advantages: (i) it yields efficient estimates of counterfactuals even when $\hat{\theta}$ is inefficient, (ii) standard errors are available in closed form, avoiding the need to bootstrap; and (iii) it allows for combined market-level and microdata.

3 Case 2: Individual-Level Price Variation and Product Fixed Effects

Next, we consider settings with individual-level price variation and choice data. Following an established literature (Dubé and Rossi, 2019), we assume the researcher includes product fixed effects to account for systematic differences across products and is willing to rule out any remaining price endogeneity.

3.1 Model and Data

The researcher has data on a large number n of consumers in a single market in which J goods are sold,[13] and identifies the outside option with $j = 0$. For each consumer i, the researcher observes individual choices $d_i = (d_{ij})_{j=1}^J$ where $d_{ij} = 1$ if i chooses good j and 0 otherwise, $p_i = (p_{ij})_{j=1}^J$ which collects prices and other variables (e.g.,

[13] As before, we assume that J is fixed but it is straightforward to extend our analysis to asymptotic thought experiments where J grows slowly with n. For ease of exposition, we present results for a single market, though our approach extends easily to settings with multiple markets.

rankings on the results page) that vary across consumers, and a vector of demographic variables y_i. For each product j, the data also may contain attributes x_j that are common across all consumers. The model predicts choice probabilities as a function of p_i, y_i, $x = (x_j)_{j=1}^J$, and a parameter vector θ:

$$Pr(d_{ij} = 1|p_i, y_i, x; \theta) = \sigma_j(p_i, y_i, x; \theta), \quad j = 1, \ldots, J. \tag{22}$$

This model subsumes many empirical examples. Here we give just one standard workhorse model.

Example 3 (Mixed Logit with Fixed Effects)**.** *The utility that individual i derives from good j is of standard mixed-logit form with microdata:*

$$\begin{aligned} u_{ij} &= \alpha_i' p_{ij} + \beta_i' x_j + x_j' \Pi y_i + \xi_j + \varepsilon_{ij}, \qquad j = 1, \ldots, J, \\ u_{i0} &= \varepsilon_{i0}, \end{aligned}$$

where ξ_j is a product fixed effect and the ε_{ij} are iid type 1 extreme value random variables. The vector x_j collects characteristics with random coefficients only; characteristics with non-random coefficients are absorbed into the product fixed effect ξ_j. Choice probabilities are

$$\sigma_j(p_i, y_i, x; \theta) = \int \frac{e^{\alpha' p_{ij} + \beta' x_j + x_j' \Pi y_i + \xi_j}}{1 + \sum_{k=1}^J e^{\alpha' p_{ik} + \beta' x_k + x_k' \Pi y_i + \xi_k}} dF(\alpha, \beta; \theta), \quad j = 1, \ldots, J,$$

where F is a parametric distribution and θ contains $(\xi_j)_{j=1}^J$ and other parameters.

As before, we partition $x_j \equiv (\bar{x}_j, e_j)$, where $\bar{x}_j$ is a vector of standard observed product attributes, such as product size, and e_j is an r-vector representing product attributes that are harder to capture using standard numeric data and which we treat as latent to the econometrician. We again assume the researcher has proxies $\tilde{e} = (\tilde{e}_j')_{j=1}^J$ for the true underlying $e = (e_j')_{j=1}^J$ and stay agnostic on the form that $\tilde{e}$ takes. A leading case is again the scenario where $\tilde{e}_j$ are embeddings computed to represent unstructured data U_j, though our approach may equally be used in scenarios where $\tilde{e}_j$ represents some potentially mismeasured product attributes.

3.2 Bias-Corrected Counterfactuals

We are interested in estimating a counterfactual of the form

$$\kappa = \mathbb{E}[k(p_i, y_i, \bar{x}, e; \theta)],$$

where the expectation is over the distribution of (p_i, y_i). Here κ might represent an average price-elasticity of consumers, average equilibrium price, or average welfare measure. It could also represent a quantity that doesn't depend on the distribution of (p_i, y_i), such as the price-elasticity or welfare measure for an individual with given y facing given prices p, in which case the expectation is redundant.

In the usual workflow, model parameters are estimated using the observed data and a candidate set of proxies $\tilde{e}$. For instance, one could use maximum likelihood. Given an estimate $\hat{\theta}$ of θ, the counterfactual κ is usually estimated as

$$\hat{\kappa} = \frac{1}{n} \sum_{i=1}^{n} k(p_i, y_i, \bar{x}, \tilde{e}; \hat{\theta}).$$

As before, we refer to this as the *naive estimator* of κ since it does not account for the fact that the proxies $\tilde{e}$ might differ from the true latent attributes. Mismeasurement of e affects the naive estimator of κ both directly, since $\tilde{e}$ is an argument of k, and indirectly through bias in the first-stage estimate $\hat{\theta}$, since $\tilde{e}$ enters the likelihood.

We now introduce a bias-corrected estimator of κ that is designed to mitigate the effects of using $\tilde{e}$ in place of the true e. As before, we consider models in which e and θ enter choice probabilities (22) via a composite parameter

$$\gamma \equiv \gamma(\theta, e).$$

As before, many common specifications have this property. We illustrate it in our leading example.

Example 3 (continued). *We partition* $\beta_i = (\beta_{\bar{x},i}, \beta_{e,i})$ *and* $\Pi = [\Pi_{\bar{x}}\ \Pi_e]$ *and write the utilities as:*

$$u_{ij} = \alpha_i' p_{ij} + \bar{x}_j' \Pi_{\bar{x}} y_i + e_j' \Pi_e y_i + \xi_j + \beta_{\bar{x},i}' \bar{x}_j + \beta_{e,i}' e_j + \varepsilon_{ij}, \quad j = 1, \ldots, J,$$

Suppose $\alpha_i \sim N(\bar{\alpha}, \Sigma_\alpha)$, $\beta_{\bar{x},i} \sim N(0, \Sigma_{\bar{x}})$, *and* $\beta_{e,i} \sim N(0, \Sigma_e)$, *and* α_i, $\beta_{\bar{x},i}$, *and* $\beta_{e,i}$

are independent. Recall the notation l and l_r from Example 1. Then,

$$\theta = (\bar{\alpha}, \xi, v(\Pi_{\bar{x}}), v(\Pi_e), l(\Sigma_\alpha), l(\Sigma_{\bar{x}}), l(\Sigma_e)),$$

where $\xi = (\xi_j)_{j=1}^J$, and $v(\Pi)$ stacks the entries of Π into a vector. Collecting $\beta'_{e,i} e_j$ across products, we have $e\beta_{e,i} \sim N(0, e\Sigma_e e')$, where $e\Sigma_e e'$ has rank $r \leq J$. Hence,

$$\gamma(\theta, e) = (\bar{\alpha}, \xi, v(\Pi_{\bar{x}}), v(e\Pi_e), l(\Sigma_\alpha), l(\Sigma_{\bar{x}}), l_r(e\Sigma_e e')).$$

As before, parameters that do not interact with e, such as $\bar{\alpha}$ and ξ, are left unchanged, whereas we expand the parameter space for those that interact with e to capture the effect of shifting e and/or θ.

We shall implicitly assume in what follows that the right-hand side of (22) depends on (θ, e) only via $\gamma(\theta, e)$. We similarly restrict attention to counterfactuals that depend on (θ, e) only via $\gamma(\theta, e)$ and write

$$\begin{aligned} \sigma_{ij}(\gamma(\theta, e)) &= \sigma_j(p_i, y_i, \bar{x}, e; \theta), \quad j = 1, \ldots, J, \\ k_i(\gamma(\theta, e)) &= k(p_i, y_i, \bar{x}, e; \theta). \end{aligned}$$

We then define the *bias-corrected estimator*

$$\hat{\kappa}_{bc} = \frac{1}{n} \sum_{i=1}^{n} \left(k_i(\hat{\gamma}) + \hat{c}_i'(d_i - \sigma_i(\hat{\gamma})) \right), \tag{23}$$

where $\hat{\gamma} = \gamma(\hat{\theta}, \tilde{e})$, and $\hat{c}_i$, $d_i = (d_{ij})_{j=1}^J$, and $\sigma_i(\gamma) = (\sigma_{ij}(\gamma))_{j=1}^J$ are $J \times 1$ vectors, with

$$\hat{c}_i = V_i(\hat{\gamma})^{-1} \dot{\sigma}_i(\hat{\gamma}) \hat{H}^{-1} \hat{k}, \tag{24}$$

where $V_i(\gamma) = \text{diag}(\sigma_i(\gamma)) - \sigma_i(\gamma)\sigma_i(\gamma)'$ and $\hat{H} = \frac{1}{n} \sum_{i=1}^n \dot{\sigma}_i(\hat{\gamma})' V_i(\hat{\gamma})^{-1} \dot{\sigma}_i(\hat{\gamma})$ are of dimension $J \times J$, $\hat{k} = \frac{1}{n} \sum_{i=1}^n \dot{k}_i(\hat{\gamma})$ and $\dot{k}_i(\gamma) = \frac{\partial k_i(\gamma)}{\partial \gamma}$ are $\dim(\gamma) \times 1$, and $\dot{\sigma}_i(\gamma)' = \frac{\partial \sigma_i(\gamma)'}{\partial \gamma}$ is $\dim(\gamma) \times J$. Importantly, $\hat{\kappa}_{bc}$ involves closed-form expressions of objects that can be easily computed given the estimate $\hat{\theta}$. As a result, it requires minimal computation beyond what is needed to estimate the model.

As before, $\hat{\kappa}_{bc}$ takes the naive estimator and adds an adjustment term that purges the effect of $\hat{\gamma}$ on $\hat{\kappa}_{bc}$ to first order. Proposition 5 in Section 6.2 shows that the asymptotic distribution of $\hat{\kappa}_{bc}$ is centered around the true counterfactual κ_0 and does

not depend on $\hat{\gamma}$. This means $\hat{\kappa}_{bc}$ is immune to any bias arising from proxying e with $\tilde{e}$, and its variance is not impacted by data-dependent $\tilde{e}$. We defer the formal statement of these results to Section 6.2 and instead highlight a few key properties.

Remark 5 (Easy to compute standard errors)**.** The asymptotic variance of $\hat{\kappa}_{bc}$ can be easily estimated using

$$\hat{V}_{bc} = \hat{s}_k^2 + \frac{1}{n}\sum_{i=1}^{n} \hat{c}_i' V_i(\hat{\gamma}) \hat{c}_i, \tag{25}$$

where $\hat{s}_k^2$ is the sample variance of $k_i(\hat{\gamma})$. Standard errors for $\hat{\kappa}_{bc}$ are then $\sqrt{\hat{V}_{bc}/n}$.

Remark 6 (Semiparametric Efficiency)**.** We may view model (22) as a conditional moment model based on the moment condition

$$\sigma_i(\gamma_0) = \mathbb{E}[d_i | p_i, y_i, \bar{x}]. \tag{26}$$

In Section 6.2, we show that the asymptotic variance of $\hat{\kappa}_{bc}$ is the semiparametric efficiency bound for estimating κ in model (26) (Brown and Newey, 1998; Ai and Chen, 2012). Thus, there do not exist regular estimators of κ in model (26) with smaller asymptotic variance than the bias-corrected estimator $\hat{\kappa}_{bc}$.

Remark 7 (Fine Tuning)**.** The bias correction in (23) allows the proxies $\tilde{e}$ to be sample-dependent, for instance because an ML algorithm has been fine tuned on the choice data to provide a better fit. As before, there is no need to correct the standard errors: one can use $\sqrt{\hat{V}_{bc}/n}$ with $\hat{V}_{bc}$ as above whether or not $\tilde{e}$ is data-dependent.

3.3 Diagnostics

Here we propose two diagnostics that can be used to assess the suitability of a candidate set of proxies $\tilde{e}$. The first can be used to assess whether $\hat{\gamma} = \gamma(\hat{\theta}, \tilde{e})$ is sufficiently close to the truth $\gamma_0 = \gamma(\theta_0, e_0)$ as required by our theory in Section 6.2. The second can be used to determine whether the true e are higher dimensional than the proxies $\tilde{e}$. Both diagnostics are again based on LM statistics so that they require minimal additional computation. Later, we will show that these two diagnostics perform well in finite samples via both simulations and in our empirical application.

The standard demand estimation workflow implicitly assumes that product attributes are measured without error and are of the correct dimension. Our diagnostics may be used to assess the validity of these implicit assumptions even in standard contexts with only quantifiable attributes.

3.3.1 Diagnostic 1: Is $\hat{\gamma}$ Close To γ_0?

As in Section 2.3.1, a simple LM statistic can be used to check whether $\hat{\gamma}$ is sufficiently close to γ_0. This diagnostic is also helpful to guide the choice among sets of embeddings obtained from various data source and/or ML model combinations. Let

$$LM_1 = \|\sqrt{n}\hat{H}^{-1/2}\hat{S}\|^2, \tag{27}$$

where $\hat{S} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}\frac{d_{ij}}{\sigma_{ij}(\hat{\gamma})}\dot{\sigma}_{ij}(\hat{\gamma})$ is the score and $\hat{H}$ is the (expected) Hessian, defined below (24). As before, this diagnostic can be interpreted as an LM statistic for a test of the null hypothesis that γ_0 is in the set of composite parameters spanned by $\tilde{e}$. Proposition 6 below shows that LM_1 behaves like $\|\sqrt{n}(\hat{\gamma}-\gamma_0)\|^2$ as n grows large. This allows researchers to validate the assumption that the discrepancy between $\hat{\gamma}$ and γ_0 is sufficiently small, as needed in Proposition 5, by checking whether LM_1 is below a threshold. In particular, for any sequence $C_n = o(n^{1/4})$, we have that $LM_1 \leq C_n^2$ implies $\|\hat{\gamma} - \gamma_0\| \leq \text{constant} \times C_n/\sqrt{n} = o(n^{-1/4})$ with probability approaching one. Further, Proposition 7 shows that LM_1 can be used to bound the discrepancy between $\tilde{e}$ and e_0.

3.3.2 Diagnostic 2: Is The Dimension of $\tilde{e}$ Correct?

The construction follows similar ideas to Section 2.3.2. We augment $\tilde{e}$ with a vector η representing additional attributes not included in $\tilde{e}$ and augment θ with an additional component $\psi \in \Psi$ representing coefficients on η. Correspondingly, we extend γ to $\zeta = (\gamma, \psi)$. With slight abuse of notation, we now write choice probabilities on this extended space as $\sigma_{ij}(\zeta; \eta)$, with the understanding that $\sigma_{ij}(\gamma) = \sigma_{ij}((\gamma, 0); \eta)$.

Consider an LM test of the null hypothesis that $\psi = 0$. Such a test could be based on the score

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}\frac{d_{ij}}{\sigma_{ij}(\hat{\gamma})}w_{ij}(\hat{\gamma}, \eta), \tag{28}$$

where

$$w_{ij}(\gamma,\eta) = \lim_{\psi\to 0} \frac{\partial \sigma_{ij}((\gamma,\psi);\eta)}{\partial \psi}.$$

Example 3 (continued). *Let the random coefficient on η be $\delta_i \sim \psi Z_i$, where $\psi \in [0,\infty)$ and $Z_i \sim F_0$ has mean zero and unit variance. For simplicity, suppose $\Pi = 0$. Define the functions $\sigma_{ij}(\,\cdot\,;(\gamma,\psi),\eta)$ and $\varsigma_{ij}(\,\cdot\,;(\gamma,\psi),\eta)$ from $\mathbb{R}^J$ to $\mathbb{R}$ by*

$$\sigma_{ij}((\gamma,\psi);\eta) = \int \varsigma_{ij}(\xi + \sqrt{\psi}\eta z;\gamma)\, dF_0(z),$$
$$\varsigma_{ij}(u;\gamma) = \int \frac{e^{\alpha' p_{ij} + \beta' e_j + u_j}}{1 + \sum_{k=1}^J e^{\alpha' p_{ik} + \beta' e_k + u_k}} dF(\alpha,\beta),$$

where we have suppressed dependence on p_{ij} and $\bar{x}$. Using $\dot{\varsigma}_{ij}$ and $\ddot{\varsigma}_{ij}$ to denote first and second derivatives of ς_{ij} with respect to its first argument, we have

$$\frac{\partial \sigma_{ij}((\gamma,\psi);\eta)}{\partial \psi} = \frac{1}{2\sqrt{\psi}} \int z\eta' \dot{\varsigma}_{ij}(\xi + \sqrt{\psi}\eta z;\gamma)\, dF_0(z).$$

and so $w_{ij}(\gamma,\eta) = \frac{1}{2}\eta'\ddot{\varsigma}_{ij}(\xi;\gamma)\eta$. The term $w_{ij}(\hat{\gamma},\eta) = \frac{1}{2}\eta'\ddot{\varsigma}_{ij}(\hat{\xi};\hat{\gamma})\eta$ is plugged into the LM_2 statistic below.

The statistic (28) can still depend to first order on $\hat{\gamma}$. To eliminate this dependence, we perform a similar correction to $\hat{\kappa}_{bc}$. Let $g_i(\gamma;\eta) = w_i(\gamma,\eta)'V_i(\gamma)^{-1}(d_i - \sigma_i(\gamma))$, where $w_i(\gamma;\eta) = (w_{ij}(\gamma;\eta)')_{j=1}^J$ is $J \times \dim(\psi)$. Then define the $\dim(\psi) \times J$ matrix

$$\hat{c}_i(\eta)' = \left(\frac{1}{n}\sum_{l=1}^n w_l(\hat{\gamma};\eta)'V_l(\hat{\gamma})^{-1}\dot{\sigma}_l(\hat{\gamma}) + \dot{g}_l(\hat{\gamma};\eta)\right)\hat{H}^{-1}\dot{\sigma}_i(\hat{\gamma})'V_i(\hat{\gamma})^{-1},$$

where $\dot{g}_i(\gamma;\eta)' = \frac{\partial g_i(\gamma;\eta)}{\partial \gamma}$ is $\dim(\gamma) \times \dim(\psi)$, and let

$$\hat{\lambda}(\eta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{c}_i(\eta)'(d_i - \sigma_i(\hat{\gamma})), \qquad \hat{\Lambda}(\eta) = \frac{1}{n}\sum_{i=1}^n \hat{c}_i(\eta)'V_i(\hat{\gamma})\hat{c}_i(\eta).$$

Finally, let

$$\hat{W}(\eta) = \hat{\lambda}(\eta)'\hat{\Lambda}(\eta)^{-1}\hat{\lambda}(\eta).$$

It can be shown that the statistic $\hat{W}(\eta)$ behaves like a $\chi^2_{\dim(\psi)}$ random variable under the null (i.e., when the true $\psi = 0$), but it depends on the nuisance parameter η.

Thus, we again define our second diagnostic as

$$LM_2 = \sup_{\eta \in S^J : \eta \perp C(\tilde{e})} \hat{W}(\eta).$$

Critical values can be computed by simulation: for iid $N(0,1)$ random variables $(\varpi_i^*)_{i=1}^n$, compute

$$LM_2^* = \sup_{\eta \in S^J : \eta \perp C(\tilde{e})} \hat{W}^*(\eta)^2,$$

where $\hat{W}^*(\eta) = \hat{\lambda}^*(\eta)' \hat{\Lambda}(\eta)^{-1} \hat{\lambda}^*(\eta)$, with $\hat{\lambda}^*(\eta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varpi_i \hat{c}_i(\eta)'(d_i - \sigma_i(\hat{\gamma}))$. Note $\lambda^*(\eta)$ factors into the product of terms involving η, which only need to be computed once, and terms involving $(\varpi_i^*)_{i=1}^n$, which are trivial to compute. Let $\hat{\xi}_{0.95}^*$ denote the 95th percentile of LM_2^* across a large number of independent sequences $(\varpi_i^*)_{i=1}^n$. The null that the dimension of $\tilde{e}$ is adequate can be rejected if $LM_2 > \hat{\xi}_{0.95}^*$.

3.4 Practitioner's Guide

Given a counterfactual of interest κ and a candidate set of proxies $\tilde{e}$:

1. Calculate the model parameter estimates $\hat{\theta}$ as usual.
2. Compute weights $\hat{c}$ in (24).
3. Plug the weights in (23) to compute the corrected estimator $\hat{\kappa}_{bc}$.
4. Compute standard errors for $\hat{\kappa}_{bc}$ using Remark 5.
5. To check whether a given set of proxies $\tilde{e}$ is adequate:
   (a) Compute the LM_1 statistic in (27). If it's below a threshold C_n^2, conclude that $\tilde{e}$ is sufficiently close to e_0. In Section 6.2.2, we motivate a threshold of $C_n^2 = \chi^2_{\dim(\gamma),0.95} \log n$ to deliver a rate of $\sqrt{(\log n)/n}$.
   (b) Compute the LM_2 statistic in (28). If it's below a threshold $\hat{\xi}_{0.95}^*$, conclude that $\tilde{e}$ is of adequate dimension. The bootstrap method in Section 3.3.2 can be used to compute $\hat{\xi}_{0.95}^*$.

4 Simulations

We first illustrate our approach in simulations. We consider the model in Section 3 with 10 products and 10,000 consumers. These figures are in line with the data used in the empirical application in Section 5. We model utility as a function of price and two-dimensional latent attributes e (in addition to idiosyncratic shocks). The latent attributes e_j are drawn iid $N(0,1)$ across products. Individual-level prices are drawn iid from a $N(5,1)$ distribution and vary across simulations. The random coefficient on price is $N(-1, 0.3^2)$ and the coefficients on the latent attributes are iid $N(0, 0.75^2)$. We keep the latent attributes e fixed in the data generating process and vary the amount of mismeasurement in the proxies $\tilde{e}$ that are used in estimation. Specifically, for every j, we let:

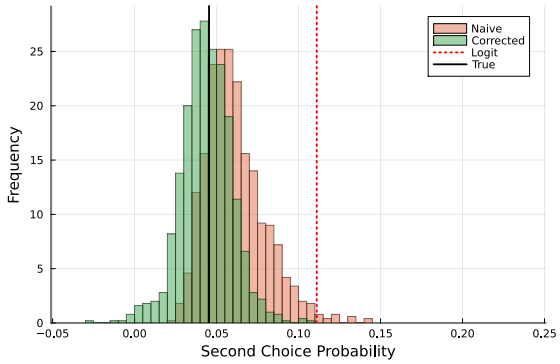
$$\tilde{e}_j = (1-\rho)e_j + \sqrt{1-(1-\rho)^2}\eta_j, \tag{29}$$

where η_j is a two-dimensional standard normal random vector drawn iid across simulations and ρ determines the amount of mismeasurement in $\tilde{e}$. When $\rho = 0$, the proxies exactly match the latent attributes, whereas as ρ increases towards 1 the proxies are increasingly mismeasured. We note that this simulation design imposes very few restrictions on the form of mismeasurement. In particular, depending on the draw of η_j, each element of $\tilde{e}_j$ could be smaller or larger than the corresponding element of e_j and this can freely vary across goods j.[14]

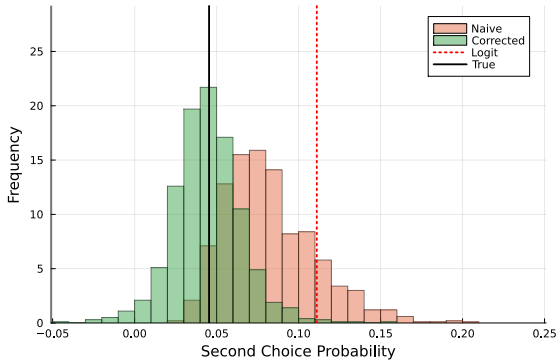
We focus on estimation of the fraction of consumers that switch from one product to another one when the former is removed from the choice set. Figure 1 plots histograms of the naive estimator that takes the proxies $\tilde{e}$ as true and the distribution of our bias corrected estimator across simulations. As the level of mismeasurement increases, the distribution of the naive estimator moves away from the true value of the counterfactual (roughly 0.05), whereas the corrected estimator remains centered around the true value. Interestingly, for larger levels of mismeasurement, the distribution of the naive estimator ends up being centered around the counterfactual prediction of the logit model with no random coefficients. This is intuitive: as the proxies $\tilde{e}$ become increasingly noisy, they capture less of the substitution patterns in the data, and the estimated variance of their random coefficients shrinks towards

[14]Note that (29) is such that $\tilde{e}_j$ has roughly the same amount of variation across goods j as e_j does. This allows us to isolate the effect of mismeasurement in a way that is not confounded by changes in the scale of the proxies $\tilde{e}_j$ used in estimation.

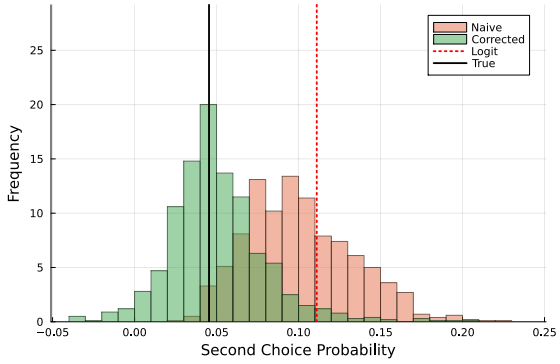
Figure 1: Distribution of the naive and bias-corrected estimators



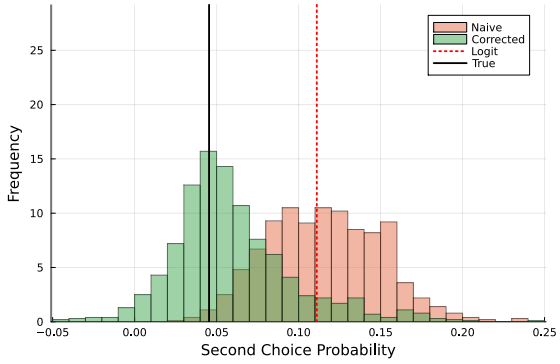
(a) Low mismeasurement ($\rho = 0.1$)



(b) Medium mismeasurement ($\rho = 0.2$)



(c) High mismeasurement ($\rho = 0.3$)

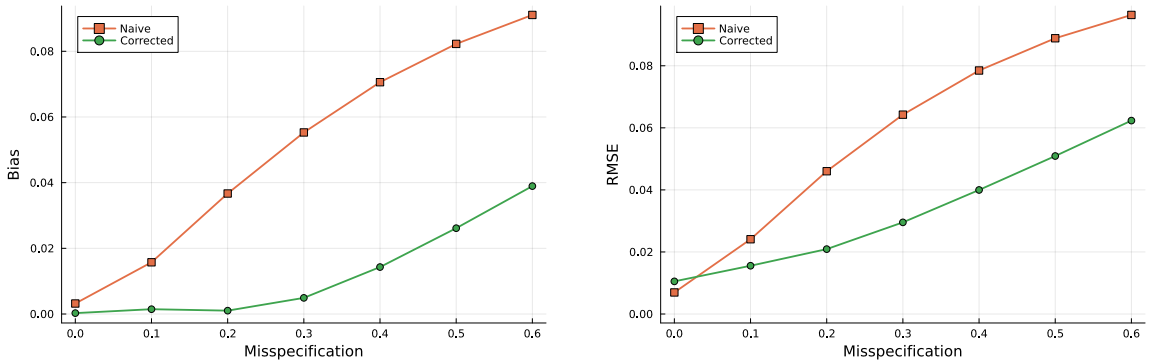


(d) Very high mismeasurement ($\rho = 0.4$)

zero. This finding also serves as a warning that mismeasurement in the proxies can defeat the purpose of estimating a random coefficients model in the first place: if the mismeasurement bias is not properly accounted for, the model may revert to the restrictive substitution patterns that the model was specifically intended to relax.

To better assess the trade-offs involved in our bias correction, Figure 2 shows how the bias and RMSE of the two estimators vary with the amount of mismeasurement ρ. When $\tilde{e}$ is measured with no error, both the naive and the bias-corrected estimator have very low bias. Our estimator has a marginally higher RMSE, indicating that its variance is slightly higher than that of the naive estimator. This is intuitive: since the naive estimator leverages the assumption that the proxies $\tilde{e}$ are correct and ours does not, we obtain slightly less precise estimates when that assumption happens to be correct. However, this is a knife-edge case. When mismeasurement is present, our estimator consistently achieves lower bias and RMSE than the naive estimator. The

Figure 2: Bias and RMSE of the naive and bias-corrected estimators



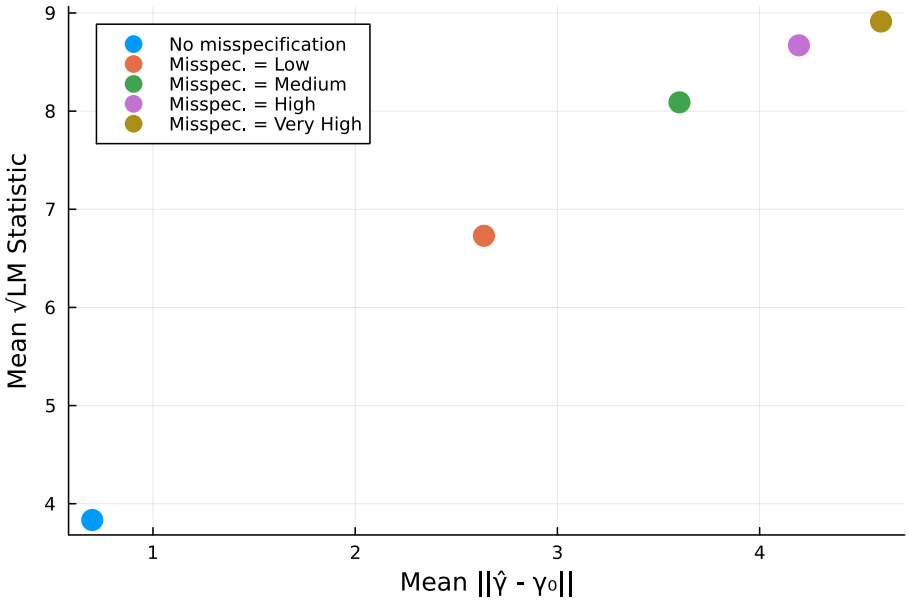
comparison is especially striking for small to moderate mismeasurement ($\rho \in [0, 0.3]$), where the bias correction is able to remove essentially all of the bias. As expected, when the mismeasurement becomes very large ($\rho > 0.5$), the bias correction starts to also perform worse. This is because the bias correction requires that $\hat{\gamma}$ be within a vicinity of γ_0 that is roughly double the order of sampling error.

Finally, Figure 3 shows that the LM_1 diagnostic discussed in Section 3.3.1 is able to correctly rank proxies. In particular, the average LM_1 statistic increases monotonically with the average distance between the $\hat{\gamma}$ induced by the proxies and γ_0. This confirms that the diagnostic can be valuable in guiding researchers towards proxies that are relatively close to the true latent attributes.

5 Empirical Application

We now apply our method to the experimental data from Compiani et al. (2025). The data records the choices made by 9,265 participants when faced with a choice of ten e-books. In a first task, participants were asked to choose their preferred e-book based on information displayed to them, including (randomized) prices, standard attributes (author, year of publication, genre and number of pages), and unstructured information (cover images, titles, plot descriptions and reviews). In a second task, each participant's first choice was removed and they were asked to choose again from the remaining nine books. Compiani et al. (2025) estimate a range of models on the first choice data and compare their performance in predicting second choices. This gives a direct measure of how well different models capture counterfactual substitution patterns. Specifically, the paper compares mixed logit models based on standard

Figure 3: Distance between $\hat{\gamma}$ and γ_0 versus LM_1 diagnostic



attributes with mixed logit models that leverage proxies extracted from unstructured data. The key findings are that (i) unstructured data is predictive of substitution patterns, and (ii) book descriptions and reviews, when processed with transformer-based text models, perform particularly well at predicting substitution. Figure 4 displays the covers of the ten books used in the experiment.

The results in Compiani et al. (2025) treat the proxies as if they were correctly specified. However, there are good reasons to believe that mismeasurement might play an important role. First, the unstructured data are processed using pre-trained ML models that are not targeted towards predicting substitution patterns.[15] While the resulting proxies are found to be predictive of substitution patterns, they may not perfectly capture the underlying attributes that drive consumers' choices. Second, the dimension of the proxies is reduced via PCA before inputting them into the demand model, which is likely to introduce further mismeasurement. This also raises the question of how many principal components should be included in the model.

Here we investigate whether applying our bias correction method and diagnostics helps better capture substitution patterns. We use the approach from Section 3 since

[15]Specifically, images are processed via classification models trained to assign each image to one of many classes; texts are processed using bag-of-words and transformer models: the former simply capture word frequency, whereas the latter are trained to predict the next word in a text.

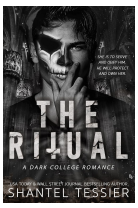
Figure 4: Covers of the ten e-books in the experiment.



Please Tell Me



The Housemaid



The Ritual



The Inmate



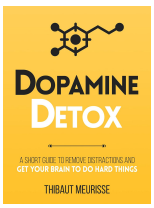
The Serpent & the Wings of Night



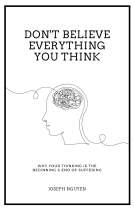
The Ashes & the Star Cursed King



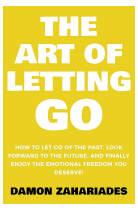
Court of Ravens and Ruin



Dopamine Detox



Don't Believe Everything You Think



The Art of Letting Go

Notes: Books 1–4 are mystery & thriller, books 5–7 are sci-fi & fantasy, and books 8–10 are self-help.

we have individual-level data and prices are randomized, so endogeneity is not a concern. We focus on the ability of different models to correctly predict the closest substitute for any given book. The second choice data give us a direct measure of this: for a given book A, its closest substitute is the book that most people switch to when A is removed from the choice set.

We note that this exercise sets a high bar for our approach. Unlike in a simulation, the demand model might be misspecified even if the proxies are correctly specified. For instance, the model assumes a normal distribution for the random coefficients but the true distribution of preference heterogeneity might be different. As a result, this exercise tests whether our approach works well even in cases where all assumptions needed for the theoretical results might not hold exactly. Further, by looking at substitution patterns in response to product removals that are not part of the estimation data, this provides a direct test of the model's ability to correctly predict counterfactuals.

Figure 5 shows the results. For each specification—defined as a combination of unstructured data source and ML model used to extract proxies from it—we report the fraction of the ten books for which the model correctly identifies the closest substitute

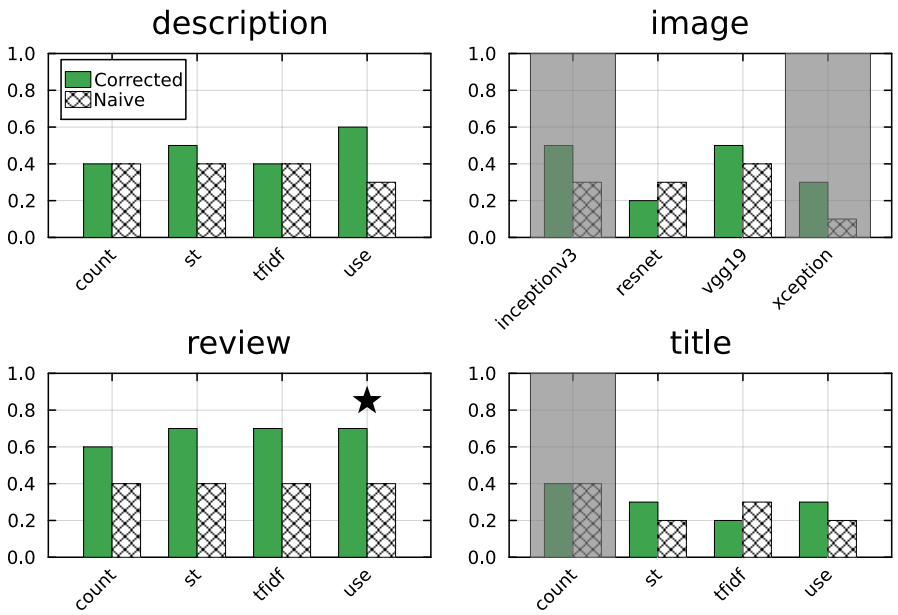
(as measured by the second choice data).[16] The hashed bars show the performance of the naive approach that uses the estimates from Compiani et al. (2025), whereas the green solid bars show the performance of the bias-corrected estimator. Three specifications are ruled out by the diagnostic LM_2, indicating that they don't feature a sufficient number of random coefficients. For 11 out of the 13 remaining specifications (around 85%), the bias correction weakly improves performance and the magnitude of the improvement is large in several cases. In particular, for specifications using reviews data, the fraction of correctly predicted substitutes goes from 40% to 60-70%. For comparison, a coin flip would achieve a hit rate of 11%. Further, the specification fitting the data best as measured by the LM_1 diagnostic is among those achieving the best counterfactual performance (70% with bias correction). These results confirm that our approach is able to meaningfully improve counterfactual predictions and guide researchers towards the best-performing specifications.

5.1 Why the Correction Helps: Breaking Away from Logit IIA Substitution Patterns

A key challenge in estimating demand for differentiated products is that the standard logit model exhibits the Independence of Irrelevant Alternatives (IIA) property: when a product is removed from the choice set, consumers substitute toward remaining products in proportion to their market shares, regardless of how similar or dissimilar those products are to the removed one (McFadden, 1974). Mixed logit models with random coefficients relax IIA by allowing substitution to depend on product similarity (Berry et al., 1995; McFadden and Train, 2000). However, random coefficients can only generate realistic substitution patterns if the product attributes entering the model adequately capture the dimensions along which products are differentiated. When the proxies used as attributes are mismeasured or of insufficient dimension, the model's ability to differentiate between products is impaired, and substitution patterns can revert toward the IIA benchmark even in the presence of random coefficients. The bias correction addresses precisely this issue, and the experimental data provide a clear illustration.

[16] The model prediction of the closest substitute is the product with the highest average (across consumers) second-choice probability once product A is removed. Here, κ corresponds to the probability that B is a consumer's second choice conditional on A being their first choice, averaged across consumers, which we compute across all (A, B) pairs.

Figure 5: Rates of correct closest substitutes predictions.

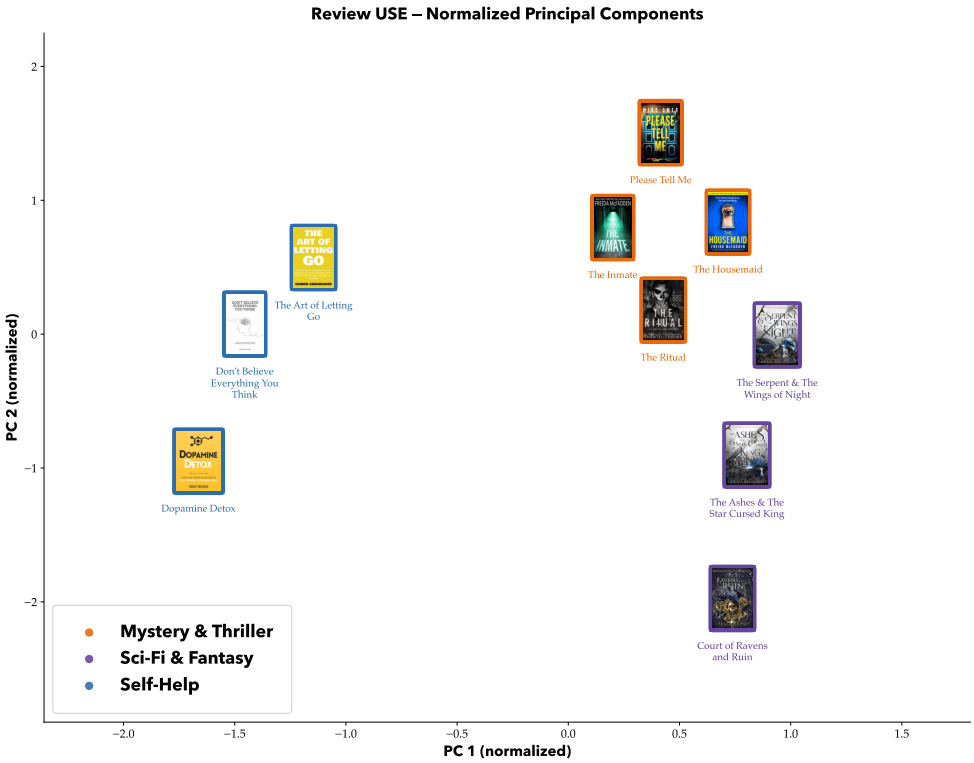


Notes: Solid bars show the fraction of books for which the bias-corrected estimator correctly identifies the closest substitute. Hashed bars show the corresponding figures for the naive estimator. The specifications ruled out by the LM_2 diagnostic are grayed out. A star flags the specification with the smallest LM_1 diagnostic.

Product positioning in the embedding space. To build intuition, Figure 6 plots the ten books in the space of the first two principal components of the review USE (Universal Sentence Encoder) embeddings. The books separate clearly by genre: mystery and thriller titles (orange) cluster on the right, sci-fi and fantasy titles (purple) on the right with lower PC2 values, and self-help titles (blue) cluster on the left. This spatial structure reflects the dimensions of differentiation that consumers care about: when a book is removed, we expect consumers to substitute primarily toward nearby books—that is, toward books in the same genre or with similar thematic content—rather than toward the most popular book overall.

Substitution predictions: naive versus corrected. Figure 7 compares closest-substitute predictions across four approaches for each of the ten books. The first row shows the focal book. The second row ("Empirical") shows the true closest substitute as revealed by the second-choice data: when a focal book is removed, the empirical closest substitute is the book that the largest share of consumers switch to. The third row ("Naive") shows the prediction of the mixed logit model estimated using review

Figure 6: Ten books positioned in the space of the first two principal components extracted from review USE embeddings.



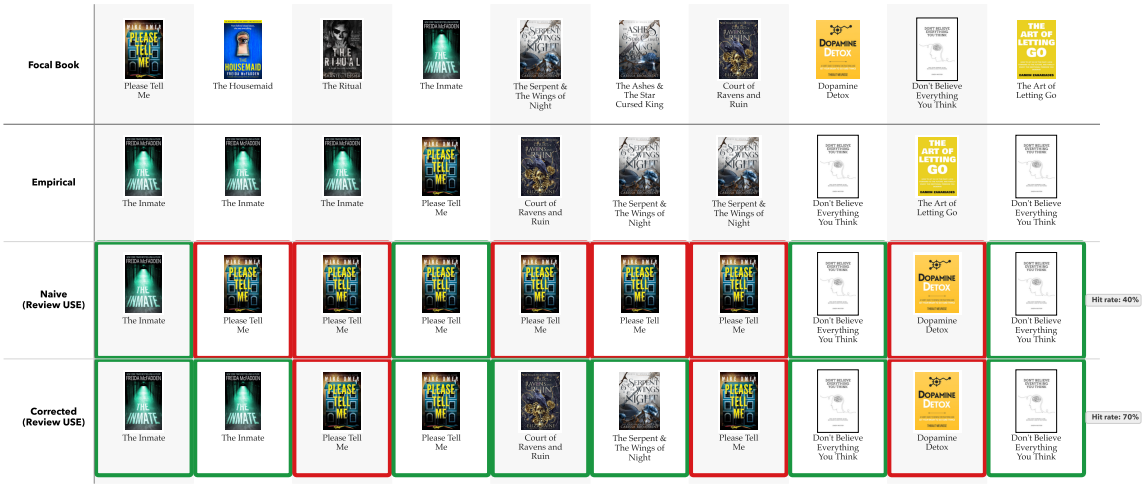
Notes: Each point represents a book, positioned according to its first two principal component scores from review USE embeddings. Colors indicate genre: orange for mystery & thriller, purple for sci-fi & fantasy, and blue for self-help.

USE embeddings *without* the bias correction. The fourth row ("Corrected") shows the prediction *with* the bias correction.

The pattern in the naive predictions is striking: the model predicts *Please Tell Me*—a mystery/thriller title—as the closest substitute for seven out of ten focal books, regardless of genre. This includes sci-fi/fantasy titles (*The Serpent & The Wings of Night*, *The Ashes & The Star Cursed King*) and even self-help titles (*Don't Believe Everything You Think*, *Dopamine Detox*). This is a hallmark of near-IIA behavior: *Please Tell Me* is the most popular book in the choice set, and the naive model channels a disproportionate share of substitution toward it, much as a logit model would. Despite the presence of random coefficients, the mismeasured proxies do not provide sufficient information for the model to differentiate between genres, causing substitution to revert toward the share-proportional IIA pattern.

The corrected model, by contrast, breaks this pattern. Substitution is redirected

Figure 7: Closest substitute predictions for all ten books: empirical, naive (review USE), and corrected (review USE).



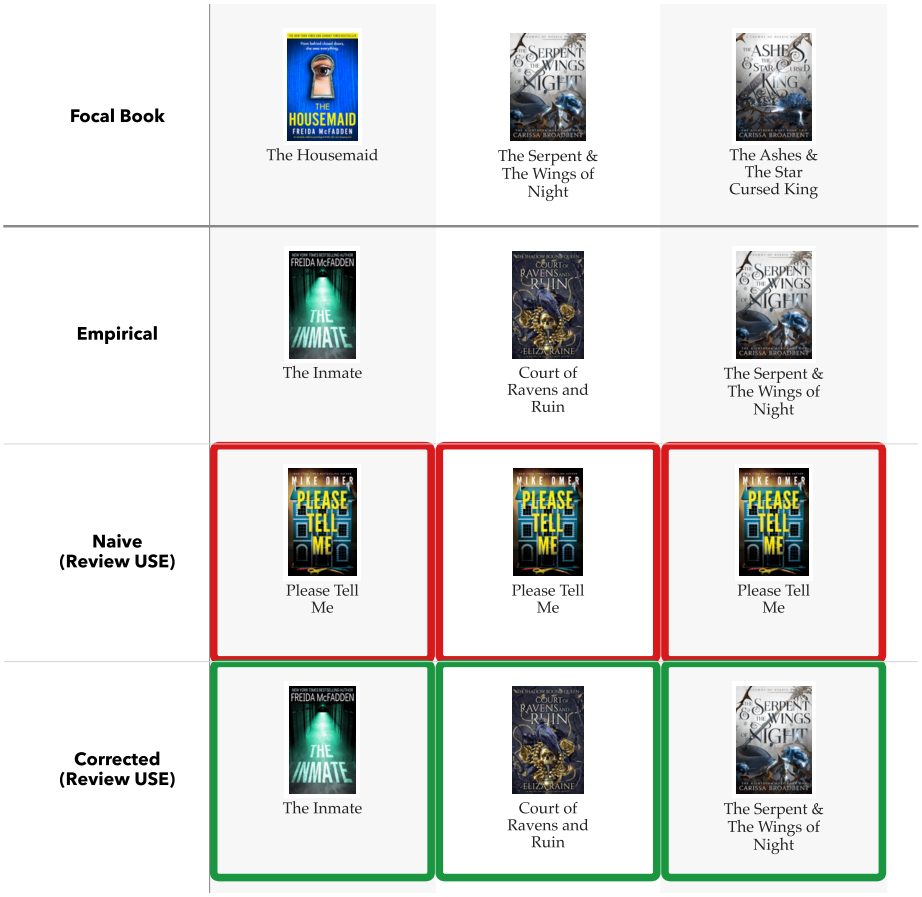
Notes: Each column corresponds to a focal book (top row). The second row shows the empirically observed closest substitute from the second-choice data. The third row shows the naive model prediction (red border = incorrect, green border = correct). The fourth row shows the bias-corrected prediction. Hit rates are reported on the right.

toward books that are genuinely similar to the focal book. For instance, when *The Housemaid* (mystery/thriller) is removed, the corrected model predicts substitution toward *The Inmate* (another mystery/thriller), matching the empirical pattern. When *The Serpent & The Wings of Night* (sci-fi/fantasy) is removed, the corrected model predicts *Court of Ravens and Ruin* (also sci-fi/fantasy) rather than *Please Tell Me*. Overall, the corrected model correctly identifies the closest substitute for 7 out of 10 books (70%), compared to 4 out of 10 (40%) for the naive model.

Zooming in: where the correction matters. Figure 8 isolates the three cases where the naive and corrected models disagree, which most clearly illustrate the mechanism. In all three cases—*The Housemaid*, *The Serpent & The Wings of Night*, and *The Ashes & The Star Cursed King*—the naive model predicts *Please Tell Me* as the closest substitute, while the corrected model predicts a within-genre alternative. In all three cases, the corrected model matches the empirical data and the naive model does not.

Consider *The Housemaid*, a mystery/thriller novel. The empirical closest substitute is *The Inmate*, another mystery/thriller that is nearby in the embedding space (Figure 6). The naive model instead predicts *Please Tell Me*, which, while also a

Figure 8: Closest substitute predictions for the three books where the naive and corrected models disagree.



Notes: The three columns correspond to the focal books for which the naive and corrected models make different closest-substitute predictions. Red borders indicate incorrect predictions; green borders indicate correct predictions.

mystery/thriller, is the overall most popular book. The corrected model predicts *The Inmate*, correctly capturing the finer within-genre substitution pattern. A similar logic applies to the two sci-fi/fantasy titles: the naive model defaults to the highest-share book overall, while the corrected model directs substitution toward other sci-fi/fantasy books, as the empirical data indicate.

The mechanism. Why does the correction break away from IIA-like behavior? The bias arises because the proxies—principal components of review embeddings—are an imperfect representation of the true latent attributes driving consumer choices. When the random coefficients are estimated using these mismeasured proxies, the

resulting heterogeneity in preferences over the embedding dimensions is attenuated: the model cannot fully distinguish how differently consumers value books that are close versus far apart in the true attribute space. This attenuation compresses the model's substitution predictions toward the share-proportional IIA benchmark.

The bias correction works by adjusting the predicted counterfactuals to account for this attenuation. Intuitively, it reweights the first-order effects of product removal to amplify substitution toward products that are similar in the proxy space and dampen substitution toward products that are dissimilar. This restores the role of product proximity in driving substitution, which is exactly the dimension along which the logit IIA property fails empirically. The result is a set of counterfactual predictions that better reflect the genre-based and thematic clustering visible in the embedding space (Figure 6), and that match the substitution patterns actually observed in the experimental data.

6 Theory

6.1 Case 1: Endogenous Prices

Let Γ denote the set of all values of $\gamma(\theta, e)$ as θ varies over the parameter space Θ and e varies over all $J \times r$ matrices with linearly independent rows. We shall implicitly assume in what follows that the true latent attributes e_0 and the proxies $\tilde{e}$ are $J \times r$ with linearly independent rows. We shall also implicitly assume that Γ is convex and open. This is true for the $\gamma(\theta, e)$ given in Examples 1 and 2, for which Γ is the product of copies of $\mathbb{R}$ and $(0, \infty)$ and is therefore convex and open.[17] In Section 2, the counterfactual function k_t, $\hat{\xi}_t$ from (8), and micro-moments m_t depended on (θ, e) only through the value of $\gamma(\theta, e)$. We can therefore view k_t, $\hat{\xi}_t$, and m_t as random functions defined on Γ.

6.1.1 Theory for Bias Correction

We first give results for the case of combined market-level data and microdata. We let $\chi_t = (p_t, \xi_t, \bar{x}_t, e)$ and let $\mathcal{M}$ denote the σ-algebra generated by $\chi_1, \ldots, \chi_\tau$. Let

[17] For instance, the operation l stacking the lower-triangular entries of the Cholesky factor maps the manifold of symmetric positive definite matrices into the product of copies of $\mathbb{R}$ and $(0, \infty)$. A similar result holds for the reduced-rank Cholesky decomposition l_r (Neuman et al., 2023).

$V = \text{Var}(Z_t\hat{\xi}_t(\gamma_0))$ be the $\dim(z)\times\dim(z)$ covariance matrix of the aggregate moments at the true parameters, and let $V_t = \text{Var}\left(m_{it}|\chi_t\right)$ be the $\dim(m)\times\dim(m)$ covariance matrix of the micro moments for market t conditional on χ_t. Let

$$H = G'V^{-1}G + \sum_{t=1}^{\tau} M_t'(r_tV_t)^{-1}M_t, \tag{30}$$

where $r_1, \ldots, r_\tau$ are defined in Assumption 1 below, and let $h = \mathbb{E}[\dot{k}_t(\gamma_0)] - G'V^{-1}K$, where $K = \mathbb{E}[k_t(\gamma_0)Z_t\hat{\xi}_t(\gamma_0))]$ is $\dim(z)\times 1$, $G = \mathbb{E}[Z_t\dot{\xi}_t(\gamma_0)]$ is $\dim(z)\times\dim(\gamma)$, and $M_t = \dot{m}_t(\gamma_0)$ is $\dim(m)\times\dim(\gamma)$. Here V and h are deterministic whereas $V_1, \ldots, V_\tau$ and H are $\mathcal{M}$-measurable random matrices. Let N be a neighborhood of γ_0.

Assumption 1. Let the following hold:

(i) $k_t(\cdot)$, $m_t(\cdot)$, and $Z_t\hat{\xi}_t(\cdot)$ are twice continuously differentiable in γ on N (almost surely), and elements of the functions and their first and second derivatives are uniformly (for $\gamma \in N$) bounded by a random variable D_t with finite fourth moment;

(ii) $\mathbb{E}[\|m_{it}\|^{2+\delta}|\mathcal{M}] \leq C$ (almost surely) for some $0 < \delta, C < \infty$;

(iii) V is positive definite and $\lambda_{\min}(V_1), \ldots, \lambda_{\min}(V_\tau), \lambda_{\min}(H) \geq \epsilon$ (almost surely) for some $\epsilon > 0$;

(iv) $T/N_t \to r_t \in (0, \infty)$ for $1 \leq t \leq \tau$.

Assumption 1(i)-(iii) are standard smoothness, moment, and rank conditions, respectively. Assumption 1(iv) treats the sample size T of the aggregate data and the sample sizes of the microdata as comparable. This is designed to give a meaningful approximation to common empirical scenarios where T is in the high tens or hundreds and N_t is in the hundreds or thousands for each market (e.g., Petrin (2002) and Grieco et al. (2024)).

The next result shows that the bias-corrected estimator $\hat{\kappa}_{bc}$ from (10) is asymptotically centered at the true counterfactual $\kappa_0 = \mathbb{E}[k_t(\gamma_0)]$ and its asymptotic variance is independent of $\hat{\gamma}$, $\hat{\theta}$, and $\tilde{e}$. Because the effect of market-level variables in markets for which there is microdata persists in the limit, we use the notion of stable convergence. We say a sequence of random variables Z_T converges in distribution to Z ($\mathcal{M}$-stably) if $\lim_{T\to\infty}\Pr(Z_T \leq z, A) = \Pr(Z \leq z, A)$ for all continuity points z of

the distribution of Z and all $\mathcal{M}$-measurable events A. Convergence in distribution is a special case corresponding to replacing $\mathcal{M}$ with the trivial σ-algebra $\{\emptyset, \Omega\}$.

Proposition 1. *Let Assumption 1 hold and $\hat{\gamma} = \gamma_0 + o_p(T^{-1/4})$. Then $\sqrt{T}(\hat{\kappa}_{bc} - \kappa_0)$ converges in distribution ($\mathcal{M}$-stably) as $T \to \infty$ to a mixed Gaussian random variable with mean zero and $\mathcal{M}$-measurable variance*

$$V_{bc} = \text{Var}\left(k_t(\gamma_0)\right) + h' H^{-1} h - K' V^{-1} K. \tag{31}$$

The (random) asymptotic variance V_{bc} can easily be estimated using $\hat{V}_{bc}$ in equation (15). Standard errors $\text{s.e.}(\hat{\kappa}_{bc}) = \sqrt{\hat{V}_{bc}/T}$ are consistent under Assumption 1 and valid inference can be performed based on t-statistics $(\hat{\kappa}_{bc} - \kappa_0)/\text{s.e.}(\hat{\kappa}_{bc})$ using the standard $N(0,1)$ critical values.

Remark 8. Proposition 1 requires $\|\hat{\gamma} - \gamma_0\| = o_p(T^{-1/4})$, which is a standard condition used in asymptotic theory for plug-in estimators (e.g., Newey, 1994, Assumption 5.1(ii)). Of course, asymptotics are only useful insofar as they deliver accurate approximations to the finite-sample distribution of $\hat{\kappa}_{bc}$ encountered in practice. In practical terms, in any finite sample, this condition requires that $\hat{\gamma}$ is within a vicinity of γ_0 that is roughly double the order of sampling uncertainty. Indeed, the simulations reported in Section 4 show that $\hat{\kappa}_{bc}$ has negligible bias up to moderate amounts of mismeasurement of $\tilde{e}$ (which translates to a range of $\|\hat{\gamma} - \gamma_0\|$) for a fixed sample size T. We also note that $\hat{\gamma} = \gamma(\hat{\theta}, \tilde{e})$, where $\hat{\theta}$ is estimated on the choice data and $\tilde{e}$ can be computed using fine tuning on the same choice data. For these reasons, it is plausible to adopt an asymptotic framework in which $\hat{\gamma}$ approaches γ_0 as the sample size T increases.

We next provide a sense in which $\hat{\kappa}_{bc}$ is efficient. The proof of Proposition 1 shows that $\hat{\kappa}_{bc}$ belongs to the class $\mathcal{K}$ of estimators $\hat{\kappa}$ of κ_0 that satisfy

$$\sqrt{T}(\hat{\kappa} - \kappa_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(k_t(\gamma_0) - \kappa_0 - c' Z_t \hat{\xi}_t(\gamma_0) \right) + \sqrt{T} \sum_{t=1}^{\tau} d_t' \left(\bar{m}_t - m_t(\gamma_0) \right) + o_p(1), \tag{32}$$

where c and $d_1, \ldots, d_\tau$ are any $\mathcal{M}$-measurable random vectors that satisfy

$$\mathbb{E}[\dot{k}_t(\gamma_0)] - G'c - \sum_{t=1}^{\tau} M_t' d_t = 0, \tag{33}$$

(almost surely). For instance, similar arguments as in the proof of Proposition 1 show that any $\hat{\kappa}$ obtained by plugging-in any $\hat{\gamma} = \gamma_0 + o_p(T^{-1/4})$ into (10) for some arbitrary weights $\hat{c}$ and $\hat{d}_1, \ldots, \hat{d}_\tau$ converging to c and $d_1, \ldots, d_\tau$ belongs to this class. Condition (33) typically ensures that such an estimator satisfies (32) uniformly for γ local to γ_0. In general, there are many different weights $\hat{c}$ and $\hat{d}_1, \ldots, \hat{d}_\tau$ whose probability limits c and $d_1, \ldots, d_\tau$ will correspond to different (random) asymptotic variances. The following result shows that $\hat{\kappa}_{bc}$ has the smallest asymptotic variance among this class of estimators of κ_0.

Proposition 2. *Let Assumption 1 hold and $\hat{\gamma} = \gamma_0 + o_p(T^{-1/4})$. Then $\hat{\kappa}_{bc}$ has the smallest asymptotic variance among the class of estimators $\mathcal{K}$ of κ_0.*

An important practical take-away from Proposition 2 is that microdata should be used, when available, to improve the efficiency of estimators of counterfactuals. Any estimator that discards the microdata by implicitly setting $d_1, \ldots, d_\tau = 0$ will have an unnecessarily large variance (and hence standard errors).

Of course, in many scenarios microdata may not be available. Here we state a simpler version of Proposition 1 tailored to this case. Recall that the bias-corrected estimator in this case is given in (11), where $\hat{c}$ is given in (13) with $\hat{H} = \hat{G}'\hat{V}^{-1}\hat{G}$. To introduce the assumptions, let V and G be as above, and let $H = G'V^{-1}G$.

Assumption 2. Let the following hold:

(i) $k_t(\cdot)$ and $Z_t\hat{\xi}_t(\cdot)$ are twice continuously differentiable in γ on N (almost surely), and the functions and all elements of their first and second derivatives are uniformly (for $\gamma \in N$) bounded by a random variable D_t with finite fourth moment;

(ii) V and H are positive definite.

The next result is a special case of Proposition 1 and is stated without proof.

Proposition 3. *Let Assumption 3 hold and* $\hat{\gamma} = \gamma_0 + o_p(T^{-1/4})$. *Then,*

$$\sqrt{T}(\hat{\kappa}_{bc} - \kappa_0) \to_d N(0, V_{bc})$$

with V_{bc} *as in* (31) *with* $H = G'V^{-1}G$.

The asymptotic variance can be easily estimated using the formula $\hat{V}_{bc}$ in (16). Standard errors are then computed as $\sqrt{\hat{V}_{bc}/T}$. These are consistent under the conditions of Proposition 3.

6.1.2 Theory for LM_1

We now present a result that provides a formal sense in which the diagnostic LM_1 in (17) behaves like $\|\sqrt{T}(\hat{\gamma} - \gamma_0)\|^2$ as the sample size grows large. We first state the result then discuss its implications. In what follows, we abbreviate "with probability approaching one" to "wpa1." Recall H from (30) and let $\lambda_{\min}(H)$ denote its smallest singular value, which is uniformly bounded away from zero by Assumption 1(iii).

Proposition 4. *Let Assumption 1 hold and let* $\hat{\gamma} = \gamma_0 + o_p(1)$. *Fix any sequence* $C_T \uparrow \infty$ *and any* $\epsilon > 0$. *Then wpa1, we have*

$$\frac{1+\epsilon}{1+2\epsilon}\left(\sqrt{LM_1} - \epsilon C_T\right) \leq \|H^{1/2}(\sqrt{T}(\hat{\gamma} - \gamma_0))\| \leq \frac{1+2\epsilon}{1+\epsilon}\left(\sqrt{LM_1} + \epsilon C_T\right).$$

In particular, wpa1 we have that $LM_1 \leq C_T^2$ *implies*

$$\|\sqrt{T}(\hat{\gamma} - \gamma_0)\| \leq \frac{(1+2\epsilon)C_T}{\sqrt{\lambda_{\min}(H)}}.$$

Moreover, if $\hat{\gamma} = \gamma_0 + o_p(C_T/\sqrt{T})$, *then wpa1 we have*

$$LM_1 \leq (1+\epsilon)^2 C_T^2.$$

Proposition 4 shows LM_1 behaves like $\|\sqrt{T}(\hat{\gamma} - \gamma_0)\|$. With $C_T = o(T^{1/4})$, wpa1 we have that $LM_1 \leq C_T^2$ implies $\|\hat{\gamma} - \gamma_0\| \leq \text{constant} \times C_T/\sqrt{T} = o(T^{-1/4})$.

The proof of Proposition 4 shows that the "wpa1" qualifier depends on whether a $\chi^2_{\dim(\gamma)}$ random variable is less than $\epsilon^2 C_T^2$. With $\epsilon = 1$, say, this suggests taking C_T^2 to

be at least as large as the 95th or 99th percentile of the $\chi^2_{\dim(\gamma)}$ distribution. To check a convergence rate of $\sqrt{(\log T)/T}$, for instance, one could use $C_T^2 = \chi^2_{\dim(\gamma),0.95} \log T$.

6.2 Case 2: Individual-Level Price Variation and Product Fixed Effects

We again let Γ denote the set of all values of $\gamma(\theta, e)$ as θ varies over the parameter space Θ and e varies over all $J \times r$ matrices with linearly independent rows, assume e_0 and $\tilde{e}$ are $J \times r$ with linearly independent rows, and that Γ is convex and open. This is true for the $\gamma(\theta, e)$ given in Example 3, for which Γ is the product of copies of $\mathbb{R}$ and $(0, \infty)$ and is therefore convex and open (see footnote 17). In Section 3, the counterfactual function k_i and choice probabilities σ_i depended on (θ, e) only through the value of $\gamma(\theta, e)$. We therefore treat k_i and σ_i as random functions on Γ.

6.2.1 Theory for Bias Correction

We now derive the theoretical properties of the bias-corrected estimator $\hat{\kappa}_{bc}$ from (23). We first outline some standard smoothness, moment, and rank assumptions. In what follows, we use a dot (as above) and double dot to denote first and second derivatives with respect to γ. Let $H = \mathbb{E}[\dot{\sigma}_i(\gamma_0)' V_i(\gamma_0)^{-1} \dot{\sigma}_i(\gamma_0)]$ and N be a neighborhood of γ_0.

Assumption 3. Let the following hold:

(i) $k_i(\cdot)$ and $\sigma_i(\cdot)$ are twice continuously differentiable in γ on N (almost surely), and all elements of the functions and their first and second derivatives are uniformly (for $\gamma \in N$) bounded by a random variable D_i with finite second moment;

(ii) $\dot{\sigma}_i(\cdot)' V_i(\cdot)^{-1}$ is continuously differentiable (almost surely) and all elements of the function and its derivative are uniformly (for $\gamma \in N$) bounded by a random variable D_i with finite higher-than-second moment;

(iii) H is positive definite.

Let $\kappa_0 = \mathbb{E}[k_i(\gamma_0)]$ denote the true value of the counterfactual. The following result shows that the asymptotic distribution of $\hat{\kappa}_{bc}$ is centered at κ_0 and its variance is independent of $\hat{\gamma}$, $\hat{\theta}$, and $\tilde{e}$.

Proposition 5. *Let Assumption 3 hold and* $\hat{\gamma} = \gamma_0 + o_p(n^{-1/4})$. *Then*

$$\sqrt{n}(\hat{\kappa}_{bc} - \kappa_0) \to_d N(0, V_{bc}),$$

as $n \to \infty$, *where*

$$V_{bc} = \mathrm{Var}\big(k_i(\gamma_0)\big) + \mathbb{E}[\dot{k}_i(\gamma_0)]' H^{-1} \mathbb{E}[\dot{k}_i(\gamma_0)]. \tag{34}$$

The asymptotic variance V_{bc} can be easily estimated using $\hat{V}_{bc}$ in (25). Standard errors are then computed as $\sqrt{\hat{V}_{bc}/n}$. These are consistent under the conditions of Proposition 5. We note that, as before, Proposition 5 requires that $\hat{\gamma}$ be in a vicinity of γ_0, and refer the reader to Remark 8 for a discussion.

6.2.2 Theory for LM_1

The following result is analogous to Proposition 4 and shows LM_1 behaves like $\|\sqrt{n}(\hat{\gamma} - \gamma_0)\|$.

Proposition 6. *Let Assumption 3 hold and let* $\hat{\gamma} = \gamma_0 + o_p(1)$. *Fix any sequence* $C_n \uparrow \infty$ *and any* $\epsilon > 0$. *Then wpa1, we have*

$$\frac{1+\epsilon}{1+2\epsilon}\left(\sqrt{LM_1} - \epsilon C_n\right) \le \|H^{1/2}(\sqrt{n}(\hat{\gamma} - \gamma_0))\| \le \frac{1+2\epsilon}{1+\epsilon}\left(\sqrt{LM_1} + \epsilon C_n\right).$$

In particular, wpa1 we have that $LM_1 \le C_n^2$ *implies*

$$\|\sqrt{n}(\hat{\gamma} - \gamma_0)\| \le \frac{(1+2\epsilon)C_n}{\sqrt{\lambda_{\min}(H)}}.$$

Moreover, if $\hat{\gamma} = \gamma_0 + o_p(C_n/\sqrt{n})$, *then wpa1 we have*

$$LM_1 \le (1+\epsilon)^2 C_n^2.$$

The implications of Proposition 6 are similar to before. In particular, for $C_n = o(n^{1/4})$, we have wpa1 that $LM_1 \le C_n^2$ implies $\|\hat{\gamma} - \gamma_0\| \le$ constant $\times\, C_n/\sqrt{n} = o(n^{-1/4})$. The proof of Proposition 6 shows that the "wpa1" qualifier depends on whether a $\chi^2_{\dim(\gamma)}$ random variable is less than $\epsilon^2 C_n^2$. To check a convergence rate of $\sqrt{(\log n)/n}$, for instance, one could use something like $C_n^2 = \chi^2_{\dim(\gamma),0.95} \log n$.

With some additional structure, we can also use LM_1 to deduce a similar bound on the proxies $\tilde{e}$. To introduce the assumptions, let θ_* and e_* be such that $\gamma(\theta_*, e_*) = \gamma_0$. We do not require that θ_* and e_* are the true structural parameters and attributes, only that they induce γ_0. Let $\hat{G}_\theta = \frac{\partial \gamma(\hat{\theta}, \tilde{e})'}{\partial \theta}$, $\hat{G}_e = \frac{\partial \gamma(\hat{\theta}, \tilde{e})'}{\partial \text{vec}(e)}$, $G_\theta = \frac{\partial \gamma(\theta_*, e_*)'}{\partial \theta}$, and $G_e = \frac{\partial \gamma(\theta_*, e_*)'}{\partial \text{vec}(e)}$ (these are well defined under Assumption 4 below). Also let $C(G_\theta)$ denote the column span of G_θ and $M = I - H^{1/2} G_\theta' (G_\theta H G_\theta')^{-1} G_\theta H^{1/2}$ denote the projection onto $C(G_\theta)^\perp$.

Assumption 4. Let the following hold:

(i) $\hat{\theta} \to_p \theta_*$ and $\tilde{e} \to_p e_*$ with $\gamma_0 = \gamma(\theta_*, e_*)$;

(ii) $\gamma(\theta, e)$ is continuously differentiable in both its arguments at (θ_*, e_*) and G_θ has full row rank;

(iii) $\hat{\theta}$ satisfies the first-order condition $0 = \hat{G}_\theta \hat{S}$ and there exists a constant C such that $\|\hat{\theta} - \theta_*\| \leq C \|\tilde{e} - e_*\|$ wpa1;

(iv) $MH^{1/2} G_e'$ has full rank.

Let $\sigma_{\min}(MH^{1/2} G_e')$ denote the smallest singular value of the matrix $MH^{1/2} G_e'$. Note this is positive by Assumption 4(iv).

Proposition 7. *Let Assumptions 3 and 4 hold. Fix any sequence $C_n \uparrow \infty$ and any $\epsilon > 0$. Then wpa1, we have*

$$\frac{1+3\epsilon}{1+\epsilon} \left(\sqrt{LM_1} - \epsilon C_n \right) \leq \|MH^{1/2} G_e' \sqrt{n} (\text{vec}(\tilde{e} - e_*))\| \leq \frac{1+3\epsilon}{1+\epsilon} \left(\sqrt{LM_1} + \epsilon C_n \right).$$

In particular, wpa1 we have that $LM_1 \leq C_n^2$ implies

$$\|\sqrt{n} (\text{vec}(\tilde{e} - e_*))\| \leq \frac{(1+2\epsilon) C_n}{\sigma_{\min}(MH^{1/2} G_e')}.$$

The proof of Proposition 6 shows that the "wpa1" qualifier depends on whether a $\chi^2_{\text{rank}(M)}$ random variable is less than $\epsilon^2 C_n^2$. With $\epsilon = 1$, say, this suggests taking C_n^2 to be at least as large as the 95th or 99th percentile of the $\chi^2_{\text{rank}(M)}$ distribution.

7 Conclusion

In this paper, we develop a toolkit to correct bias and perform valid inference on counterfactuals when the product attributes used in demand estimation may only

imperfectly capture the latent attributes that drive substitution. A leading case is when consumer choice is driven by difficult-to-quantify characteristics and unstructured data, such as product images, descriptions, review text, or consumer surveys, are converted into numerical variables using ML methods. As e-commerce continues to expand and such data play an increasingly central role in driving consumer choices, the need to incorporate these sources into demand estimation will only grow. In addition, our methods may be applied as simple post-estimation robustness checks even with standard numeric attributes when mismeasurement is a concern. All our methods require minimal additional computation once model parameters are estimated and can be easily integrated in the canonical demand estimation workflow.

References

Ai, C. and X. Chen (2012): "The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions," *Journal of Econometrics*, 170, 442–457.

Allcott, H. and N. Wozny (2014): "Gasoline prices, fuel economy, and the energy paradox," *Review of Economics and Statistics*, 96, 779–795.

Allon, G., D. Chen, Z. Jiang, and D. Zhang (2023): "Machine learning and prediction errors in causal inference," *The Wharton School Research Paper*.

Andrews, D. W. (1994a): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica: Journal of the Econometric Society*, 62, 43–72.

——— (1994b): "Empirical process methods in econometrics," *Handbook of econometrics*, 4, 2247–2294.

——— (2005): "Cross-section regression with common shocks," *Econometrica*, 73, 1551–1585.

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023): "Prediction-powered inference," *Science*, 382, 669–674.

Bach, P., V. Chernozhukov, S. Klaassen, M. Spindler, J. Teichert-Kluge, and S. Vijaykumar (2024): "Adventures in demand analysis using AI," *arXiv preprint arXiv:2501.00382*.

Backus, M., C. Conlon, and M. Sinkinson (2021): "Common Ownership and Competition in the Ready-To-Eat Cereal Industry," *NBER Working Paper 28350*.

Battaglia, L., T. Christensen, S. Hansen, and S. Sacher (2024): "Inference for Regression with Variables Generated by AI or Machine Learning," *arXiv preprint arXiv:2402.15585*.

Bayer, P., F. Ferreira, and R. McMillan (2007): "A unified framework for measuring preferences for schools and neighborhoods," *Journal of Political Economy*, 115, 588–638.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

——— (2004): "Differentiated Products Demand System from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, 112, 68–105.

Berry, S. T. and P. A. Haile (2014): "Identification in differentiated products markets using market level data," *Econometrica*, 82, 1749–1797.

——— (2021): "Foundations of demand estimation," in *Handbook of industrial organization*, Elsevier, vol. 4, 1–62.

——— (2024): "Nonparametric identification of differentiated products demand using micro data," *Econometrica*, 92, 1135–1162.

Brown, B. W. and W. K. Newey (1998): "Efficient semiparametric estimation of expectations," *Econometrica*, 66, 453–464.

Carlson, J. and M. Dell (2025): "A Unifying Framework for Robust and Efficient Inference with Unstructured Data," *arXiv preprint arXiv:2505.00282*.

Chen, X., H. Hong, and E. Tamer (2005): "Measurement error models with auxiliary data," *The Review of Economic Studies*, 72, 343–366.

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric efficiency in GMM models with auxiliary data," *The Annals of Statistics*, 36, 808–843.

Compiani, G., I. Morozov, and S. Seiler (2025): "Demand estimation with text and image data," *arXiv preprint arXiv:2503.20711*.

Conlon, C. and J. Gortmaker (2025): "Incorporating Micro Data into Differentiated Products Demand Estimation with PyBLP," *Journal of Econometrics*, 105926.

Dubé, J.-P. and P. E. Rossi (2019): *Handbook of the Economics of Marketing*, vol. 1, North Holland.

Egami, N., M. Hinck, B. Stewart, and H. Wei (2023): "Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models," *Advances in Neural Information Processing Systems*, 36, 68589–68601.

Fan, Y. (2013): "Ownership consolidation and product characteristics: A study of the US daily newspaper market," *American Economic Review*, 103, 1598–1628.

Fong, C. and M. Tyler (2021): "Machine learning predictions as regression covariates," *Political Analysis*, 29, 467–484.

Freyberger, J. (2015): "Asymptotic theory for differentiated products demand models with many markets," *Journal of Econometrics*, 185, 162–181.

Goldberg, P. K. (1995): "Product differentiation and oligopoly in international markets: The case of the US automobile industry," *Econometrica*, 891–951.

Grieco, P. L., C. Murry, J. Pinkse, and S. Sagl (2025): "Optimal Estimation

of Discrete Choice Demand Models with Consumer and Product Data," *NBER Working Paper 33397*.

Grieco, P. L., C. Murry, and A. Yurukoglu (2024): "The evolution of market power in the us automobile industry," *The Quarterly Journal of Economics*, 139, 1201–1253.

Hahn, J., G. Kuersteiner, and M. Mazzocco (2022): "Joint time-series and cross-section limit theory under mixingale assumptions," *Econometric Theory*, 38, 942–958.

Han, S. and K. Lee (2025): "Copyright and Competition: Estimating Supply and Demand with Unstructured Data," *arXiv preprint arXiv:2501.16120*.

Hansen, B. E. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, 64, 413–430.

Hausman, J. A. (1994): *Valuation of new goods under perfect and imperfect competition*, National Bureau of Economic Research Cambridge, Mass., USA.

Lee, K. (2025): "Generative brand choice," *Working Paper*.

Lee, R. S. (2013): "Vertical integration and exclusivity in platform and two-sided markets," *American Economic Review*, 103, 2960–3000.

Magnolfi, L., J. McClure, and A. Sorensen (2025): "Triplet embeddings for demand estimation," *American Economic Journal: Microeconomics*, 17, 282–307.

McFadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*.

McFadden, D. and K. Train (2000): "Mixed MNL models for discrete response," *Journal of applied Econometrics*, 15, 447–470.

Neilson, C. (2017): "Targeted vouchers, competition among schools, and the academic achievement of poor students," *Working Paper*.

Neuman, A. M., Y. Xie, and Q. Sun (2023): "Restricted Riemannian geometry for positive semidefinite matrices," *Linear Algebra and its Applications*, 665, 153–195.

Nevo, A. (2000): "Mergers with differentiated products: The case of the ready-to-eat cereal industry," *The RAND Journal of Economics*, 395–421.

——— (2001): "Measuring market power in the ready-to-eat cereal industry," *Econometrica*, 69, 307–342.

Newey, W. K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica: Journal of the Econometric Society*, 62, 1349–1382.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Petrin, A. (2002): "Quantifying the benefits of new products: The case of the minivan," *Journal of political Economy*, 110, 705–729.

Zhang, J., W. Xue, Y. Yu, and Y. Tan (2023): "Debiasing ML-or AI-Generated Regressors in Partial Linear Models," *SSRN Working Paper 4636026*.

A Proofs

A.1 Proofs for Section 6.1

Proof of Proposition 1. We have $\hat{\gamma} \in N$ wpa1 by the assumed consistency of $\hat{\gamma}$. By Assumption 1(i), wpa1 we may take a mean value expansion around γ_0 to obtain

$$\begin{aligned}\sqrt{T}(\hat{\kappa}_{bc} - \kappa_0) = &\frac{1}{\sqrt{T}}\sum_{t=1}^{T} k_t(\gamma_0) - \kappa_0 - c' Z_t \hat{\xi}_t(\gamma_0) + \sqrt{T}\sum_{t=1}^{\tau} d_t' \left(\bar{m}_t - m_t(\gamma_0)\right) \\ &+ \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left((\hat{c} - c)' Z_t \hat{\xi}_t(\gamma_0) + \sum_{s=1}^{\tau}(\hat{d}_s - d_s)'\left(\bar{m}_s - m_s(\gamma_0)\right)\right) \\ &+ \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\dot{k}_i(\tilde{\gamma})' - \hat{c}' Z_t \dot{\xi}_t(\tilde{\gamma}) - \sum_{s=1}^{\tau}\hat{d}_s' \dot{m}_s(\tilde{\gamma})\right)(\hat{\gamma} - \gamma_0) \\ =: &\, T_{1,T} + T_{2,T} + T_{3,T},\end{aligned}$$

where $\tilde{\gamma}$ is in the segment between $\hat{\gamma}$ and γ_0, and

$$c = V^{-1}(K + GH^{-1}h), \qquad d_t = (r_t V_t)^{-1} M_t H^{-1} h, \quad t = 1, \ldots, \tau. \tag{35}$$

Note that c and $d_1, \ldots, d_\tau$ are well defined by virtue of Assumption 1(iii).

For $T_{1,T}$, define the $(\dim(z)+1) \times 1$ random vector $\zeta_t = (k_t(\gamma_0) - \kappa_0, Z_t \hat{\xi}_t(\gamma_0))$. By Theorem 2 of Hahn et al. (2022) (noting Assumptions 1(i)(ii) and independence within and across markets are sufficient for their integrability and dependence conditions) and Assumption 1(iv), for any $\mathcal{M}$-measurable random vectors $d_1, \ldots, d_\tau$, we have

$$\begin{pmatrix} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\zeta_t \\ \sum_{t=1}^{\tau} d_t' \sqrt{T}\left(\bar{m}_t - m_t(\gamma_0)\right) \end{pmatrix} \to_d \begin{pmatrix} Z_A \\ \left(\sum_{t=1}^{\tau} r_t d_t' V_t d_t\right)^{1/2} Z_M \end{pmatrix}$$

$\mathcal{M}$-stably, where the random vector Z_A and random variable Z_M are jointly normally distributed and independent with mean zero, $\mathrm{Var}(Z_A) = \mathrm{Var}(\zeta_t)$, and $\mathrm{Var}(Z_M) = 1$. Moreover, (Z_A, Z_M) are independent of any $\mathcal{M}$-measurable random variable. Hence, the asymptotic distribution of $T_{1,T}$ is mixed Gaussian with mean zero and random variance

$$\mathrm{Var}\left(k_t(\gamma_0)\right) + c' V c - 2c' K + \sum_{t=1}^{\tau} r_t d_t' V_t d_t.$$

Substituting the above formulas for c and $d_1, \ldots, d_\tau$ gives the form of the variance in display (31). It remains to show that $T_{2,T}$ and $T_{3,T}$ are asymptotically negligible.

For term $T_{2,T}$, first recall the expressions for $\hat{h}$ and $\hat{H}$ in (12). Note that by Assumption 1(i)(ii) and consistency of $\hat{\gamma}$, we can deduce by standard arguments (e.g., Lemma 2.4 of Newey and McFadden (1994)) that $\hat{k} \to_p \mathbb{E}[\dot{k}_t(\gamma_0)]$, $\hat{K} \to_p K$, $\hat{G} \to_p G$, and $\hat{V} \to_p V$. Hence, $\hat{h} \to_p h$ by Assumption 1(iii) and Slutsky's theorem. Note that for each $1 \leq t \leq \tau$, the m_{it} are iid conditional on $\mathcal{M}$. It follows by Lemma 1 of Andrews (2005) and Assumption 1(ii)(iv) that $\hat{V}_t \to_p r_t V_t$. Hence, $\hat{V}_t^{-1} \to_p (r_t V_t)^{-1}$ for $1 \leq t \leq \tau$ by Assumption 1(iii). Finally, Assumption 1(i) implies $\hat{M}_t \to_p M_t$ for $1 \leq t \leq \tau$. Hence, $\hat{H} \to_p H$ and so $\hat{H}^{-1} \to_p H^{-1}$, $\hat{c} \to_p c$, and $\hat{d}_t \to_p d_t$ for $1 \leq t \leq \tau$ by Assumption 1(iii) and Slutsky's theorem.

Now write

$$T_{2,T} = (\hat{c} - c)' \frac{1}{\sqrt{T}} \sum_{t=1}^{T} Z_t \hat{\xi}_t(\gamma_0) + \sum_{t=1}^{\tau} \frac{\sqrt{T}}{\sqrt{N_t}} (\hat{d}_t - d_t)' \sqrt{N_t} \, (\bar{m}_t - m_t(\gamma_0))$$
$$=: T_{2,T,a} + T_{2,T,b}.$$

Term $T_{2,T,a} \to_p 0$ because $\hat{c} \to_p c$ and $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Z_t \xi_t(\gamma_0) = O_p(1)$ by Assumption 1(i). Similarly, $T_{2,T,b} \to_p 0$ because $\sqrt{T/N_t} \to r_t \in (0, \infty)$ by Assumption 1(iv), $\hat{d}_t \to_p d_t$, and $\sqrt{N_t}(\bar{m}_t - m_t(\gamma_0))$ converges in distribution $\mathcal{M}$-stably to a mixed normal limit with mean zero and variance V_t (by Assumption 1(ii)) and is therefore tight by Assumption 1(ii).

For term $T_{3,T}$, we first let $m_{tl}(\gamma)$ denote the l-th element of $m_t(\gamma)$, and let $\rho_{tl}(\gamma)$ denote the l-th element of $Z_t \hat{\xi}_t(\gamma)$. Similarly, we let $\hat{c}_l$ and $\hat{d}_{tl}$ denote the l-th elements of $\hat{c}$ and $\hat{d}_t$. Then we may write

$$T_{3,T} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(\dot{k}_t(\tilde{\gamma})' - \sum_{l=1}^{\dim(z)} \hat{c}_l \dot{\rho}_{tl}(\tilde{\gamma})' - \sum_{s=1}^{\tau} \sum_{l=1}^{\dim(m)} \hat{d}_{sl} \dot{m}_{sl}(\tilde{\gamma})' \right) (\hat{\gamma} - \gamma_0).$$

By construction, $\hat{c}$ and $\hat{d}_1, \ldots, \hat{d}_\tau$ satisfy the in-sample orthogonality condition

$$\hat{k} - \hat{G}' \hat{c} - \sum_{s=1}^{\tau} \hat{M}_s' \hat{d}_s = 0, \tag{36}$$

wpa1. By Assumption 1(i) we may take a second mean-value expansion, this time of

$\tilde{\gamma}$ around $\hat{\gamma}$, to arrive at

$$T_{3,T} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(\dot{k}_t(\hat{\gamma})' - \sum_{l=1}^{\dim(z)} \hat{c}_l \dot{\rho}_{tl}(\hat{\gamma})' - \sum_{s=1}^{\tau} \sum_{l=1}^{\dim(m)} \hat{d}_{sl} \dot{m}_{sl}(\hat{\gamma})' \right) (\hat{\gamma} - \gamma_0)$$
$$+ T^{1/4}(\tilde{\gamma} - \hat{\gamma})' \left(\frac{1}{T} \sum_{t=1}^{T} \left(\ddot{k}_t(\check{\gamma}) - \sum_{l=1}^{\dim(z)} \hat{c}_l \ddot{\rho}_{tl}(\check{\gamma}) - \sum_{s=1}^{\tau} \sum_{l=1}^{\dim(m)} \hat{d}_{sl} \ddot{m}_{sl}(\check{\gamma}) \right) \right) T^{1/4}(\hat{\gamma} - \gamma_0)$$
$$=: T_{3,T,a} + T_{3,T,b},$$

wpa1, where $\check{\gamma}$ is in the segment between $\hat{\gamma}$ and $\tilde{\gamma}$, $\ddot{k}_t(\gamma) = \frac{\partial^2 k_t(\gamma)}{\partial \gamma \partial \gamma'}$, $\ddot{\rho}_{tl}(\gamma) = \frac{\partial^2 \rho_{tl}(\gamma)}{\partial \gamma \partial \gamma'}$, and $\ddot{m}_{tl}(\gamma) = \frac{\partial^2 m_{tl}(\gamma)}{\partial \gamma \partial \gamma'}$.

We have

$$T_{3,T,a} = \left(\hat{k} - \hat{G}'\hat{c} - \sum_{s=1}^{\tau} \hat{M}_s' \hat{d}_s \right) \sqrt{T}(\hat{\gamma} - \gamma_0) = 0$$

wpa1 by the in-sample orthogonality condition (36).

To show $T_{3,T,b} \to_p 0$, in view of the condition $\hat{\gamma} = \gamma_0 + o_p(T^{-1/4})$, it is enough to show that the central term in parentheses is $O_p(1)$. To this end, standard arguments (e.g., Lemma 2.4 of Newey and McFadden (1994)) using Assumption 1(i) and consistency of $\hat{\gamma}$ yield $\frac{1}{T}\sum_{t=1}^{T} \ddot{k}_t(\check{\gamma}) \to_p \mathbb{E}[\ddot{k}_t(\gamma_0)]$ and $\frac{1}{T}\sum_{t=1}^{T} \ddot{\rho}_{tl}(\check{\gamma}) \to_p \mathbb{E}[\ddot{\rho}_{tl}(\gamma_0)]$, both of which are finite. It also follows by the fact that $\hat{d}_t \to_p d_t$ for $1 \leq t \leq \tau$, Assumption 1(i), and consistency of $\hat{\gamma}$ that $\hat{d}_{sl}\ddot{m}_{sl}(\check{\gamma}) \to_p d_{sl}\ddot{m}_{sl}(\gamma_0)$ for $1 \leq s \leq \tau$ and $1 \leq l \leq L$. Finally, Assumption 1(i)-(iii) implies c and $d_1, \ldots, d_\tau$ are tight. □

Proof of Proposition 2. Arguing as in the proof of Proposition 1, the asymptotic distribution of any estimator of the form (32) is mixed Gaussian with mean zero and variance

$$\mathrm{Var}\left(k_t(\gamma_0)\right) + c'Vc - 2c'K + \sum_{t=1}^{\tau} r_t d_t' V_t d_t.$$

Conditioning on $\mathcal{M}$, we may minimize this expression with respect to the vectors c and $d_1, \ldots, d_\tau$ subject to (33) to obtain the weights c and $d_1, \ldots, d_\tau$ in (35). Substituting into the above display yields the minimum variance V_{bc} given in (31). □

Before proving Proposition 4, we first state and prove a lemma.

Lemma 1. *Let Assumption 1 hold and let* $\hat{\gamma} = \gamma_0 + o_p(1)$. *Then*

$$\sqrt{T}\hat{S} = Z_T + o_p(1) + (H + o_p(1))(\sqrt{T}(\hat{\gamma} - \gamma_0)),$$

where $Z_T := G'V^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\hat{\xi}_t(\gamma_0) + \sum_{t=1}^{\tau} M_t'(r_tV_t)^{-1}\sqrt{T}(m_t(\gamma_0) - \bar{m}_t)$ *converges in distribution (*$\mathcal{M}$*-stably) to a mixed Gaussian random variable with mean zero and* $\mathcal{M}$*-measurable variance* H.

Proof of Lemma 1. By definition of $\hat{S}$, we have

$$\begin{aligned}\sqrt{T}\hat{S} = \hat{G}'\hat{V}^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\hat{\xi}_t(\gamma_0) + \sum_{t=1}^{\tau}\hat{M}_t'\hat{V}_t^{-1}\sqrt{T}(m_t(\gamma_0) - \bar{m}_t)\\ + \hat{G}'\hat{V}^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t(\hat{\xi}_t(\hat{\gamma}) - \hat{\xi}_t(\gamma_0))\right)\\ + \sum_{t=1}^{\tau}\hat{M}_t\hat{V}_t^{-1}\sqrt{T}(m_t(\hat{\gamma}) - m_t(\gamma_0)) =: T_{1,T} + T_{2,T} + T_{3,T} + T_{4,T},\end{aligned}$$

where $\hat{G} \to_p G$, $\hat{M}_t \to_p M_t$ for $1 \le t \le \tau$, $\hat{V}^{-1} \to_p V^{-1}$, and $\hat{V}_t^{-1} \to_p (r_tV_t)^{-1}$ for $1 \le t \le \tau$ (all by the proof of Proposition 1).

For $T_{1,2}$ and $T_{2,T}$, we have by the proof of Proposition 1 that $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\hat{\xi}_t(\gamma_0)$ and $\sqrt{N_t}(m_t(\gamma_0) - \bar{m}_t)$, $1 \le t \le \tau$, are all $O_p(1)$. It follows by Assumption 1(iv) that $T_{1,T} + T_{2,T} = Z_T + o_p(1)$. Hence, by similar arguments to the proof of Proposition 1 we may invoke Theorem 2 of Hahn et al. (2022) to conclude that Z_T converges $\mathcal{M}$-stably to a mixed Gaussian limit with mean zero and variance H.

For $T_{3,T}$ and $T_{4,T}$, a mean-value expansion in $\hat{\gamma}$ around γ_0 yields

$$T_{3,T} = \hat{G}'\hat{V}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} Z_t\dot{\xi}_t(\tilde{\gamma})\right)\sqrt{T}(\hat{\gamma} - \gamma_0), \quad T_{4,T} = \sum_{t=1}^{\tau}\hat{M}_t\hat{V}_t^{-1}\dot{m}_t(\tilde{\gamma})'\sqrt{T}(\hat{\gamma} - \gamma_0),$$

for $\tilde{\gamma}$ in the segment between $\hat{\gamma}$ and γ_0. It follows by Assumption 1(i) and standard arguments that $\frac{1}{T}\sum_{t=1}^{T} Z_t\dot{\xi}_t(\tilde{\gamma}) \to_p G$ and $\dot{m}_t(\tilde{\gamma}) \to_p M_t$, $1 \le t \le \tau$. Hence, $T_{3,T} + T_{4,T} = (H + o_p(1))\sqrt{T}(\hat{\gamma} - \gamma_0)$. □

Proof of Proposition 4. The proof of Proposition 1 shows that $\hat{H} \to_p H$. Combined

with Lemma 1 and the triangle inequality, we have

$$\begin{aligned}&\|(H^{1/2}+o_p(1))(\sqrt{T}(\hat{\gamma}-\gamma_0))\|+\|(H^{-1/2}+o_p(1))(Z_T+o_p(1))\| \\ &\quad\geq \sqrt{LM_1}\geq \|(H^{1/2}+o_p(1))(\sqrt{T}(\hat{\gamma}-\gamma_0))\|-\|(H^{-1/2}+o_p(1))(Z_T+o_p(1))\|.\end{aligned}$$

As $\|(H^{-1/2}+o_p(1))(Z_T+o_p(1))\|^2\to_d \chi^2_{\dim(\gamma)}$ by the proof of Lemma 1, we have

$$\|(H^{-1/2}+o_p(1))(Z_T+o_p(1))\|\leq \epsilon C_T$$

wpa1. Moreover, we have that

$$\frac{1+2\epsilon}{1+\epsilon}\|H^{1/2}(\sqrt{T}(\hat{\gamma}-\gamma_0))\|\geq \|(H^{1/2}+o_p(1))(\sqrt{T}(\hat{\gamma}-\gamma_0))\|\geq \frac{1+\epsilon}{1+2\epsilon}\|H^{1/2}(\sqrt{T}(\hat{\gamma}-\gamma_0))\|$$

wpa1. The first result follows by combining the above three displays and rearranging. The second and third results are implications of the first. □

A.2 Proofs for Section 6.2

Proof of Proposition 5. Let $c_i'=(\mathbb{E}[\dot{k}_i(\gamma_0)])'H^{-1}\dot{\sigma}_i(\gamma_0)'V_i(\gamma_0)^{-1}$ denote the population counterpart of $\hat{c}_i'$. We have $\hat{\gamma}\in N$ wpa1 by consistency of $\hat{\gamma}$. Hence, wpa1, we may take a mean value expansion around γ_0 to obtain

$$\begin{aligned}\sqrt{n}(\hat{\kappa}_{bc}-\kappa_0)&=\frac{1}{\sqrt{n}}\sum_{i=1}^n k_i(\gamma_0)-\kappa_0+c_i'(d_i-\sigma_i(\gamma_0)) \\ &\quad+\frac{1}{\sqrt{n}}\sum_{i=1}^n(\hat{c}_i-c_i)'(d_i-\sigma_i(\gamma_0)) \\ &\quad+\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\dot{k}_i(\tilde{\gamma})'-\hat{c}_i'\dot{\sigma}_i(\tilde{\gamma})\right)(\hat{\gamma}-\gamma_0) \\ &=: T_{1,n}+T_{2,n}+T_{3,n},\end{aligned}$$

where $\tilde{\gamma}$ is in the segment between $\hat{\gamma}$ and γ_0. This expansion is valid by Assumption 3(i). Term $T_{1,n}$ is asymptotically $N(0,V_{bc})$. It remains to show that $T_{2,n}$ and $T_{3,n}$ are both asymptotically negligible.

For $T_{2,n}$, first define the $1 \times \dim(\gamma)$ vectors

$$\hat{a} = \bar{k}'\hat{H}^{-1}, \qquad a = \mathbb{E}[\dot{k}_i(\gamma_0)]'H^{-1},$$

where $\bar{k} = \frac{1}{n}\sum_{i=1}^n \dot{k}_i(\hat{\gamma})$. Also define the $\dim(\gamma) \times J$ random element $b_i(\gamma) = \dot{\sigma}_i(\gamma)'V_i(\gamma)^{-1}$, and let $e_i = d_i - \sigma_i(\gamma_0)$, which is $J \times 1$. Then we may write

$$\begin{aligned} T_{2,n} &= \hat{a}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (b_i(\hat{\gamma}) - b_i(\gamma_0))e_i\right) + (\hat{a} - a)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n b_i(\gamma_0)e_i\right) \\ &=: T_{2,n,a} + T_{2,n,b}. \end{aligned}$$

By Assumptions 3(i)(ii) and consistency of $\hat{\gamma}$, it follows by standard arguments (e.g., Newey and McFadden, 1994, Lemma 2.4) that $\bar{k} \to_p \mathbb{E}[\dot{k}_i(\gamma_0)]$ and $\hat{H} \to_p H$. Hence, $\hat{a} \to_p a$ by Assumption 3(iii).

To show $T_{2,n,a} \to_p 0$, first note that $\hat{a} = O_p(1)$ and $\mathbb{E}[b_i(\gamma)e_i] = 0$. Consider the empirical process $\nu_n(\gamma) = \frac{1}{\sqrt{n}}\sum_{i=1}^n b_i(\gamma)e_i$ defined for $\gamma \in N$. For $\gamma_1, \gamma_2 \in N$, we have by a mean-value expansion that $b_i(\gamma_1)e_i - b_i(\gamma_2)e_i = (\gamma_1 - \gamma_2)'\dot{b}_i(\tilde{\gamma})e_i$ for $\tilde{\gamma}$ in the segment between γ_1 and γ_2 (with possibly different values for each element), where $\dot{b}_i(\gamma)e_i = \frac{\partial}{\partial\gamma}(\dot{b}_i(\tilde{\gamma})e_i)'$. This expansion is valid in view of Assumption 3(ii). The elements of e_i are bounded by ± 1 and the elements of $\dot{b}_i(\gamma)$ are uniformly (for $\gamma \in N$) bounded by some random variable with finite second moment, again by Assumption 3(ii). Hence, $\|b_i(\gamma_1)e_i - b_i(\gamma_2)e_i\| \le B_i\|\gamma_1 - \gamma_2\|$ for $\gamma_1, \gamma_2 \in N$, for some random variable B_i with finite second moment. Thus, $\{b_i(\gamma)e_i : \gamma \in N\}$ is a type-II class of Andrews (1994b). It follows by Theorems 1 and 2 of Andrews (1994b) (using Assumption 3(ii) to verify the moment condition on the envelope function) that $\nu_n(\cdot)$ is stochastically equicontinuous. Also note by the Lipschitz condition the pseudometric corresponding to this process is dominated by the Euclidean metric. Hence, by consistency of $\hat{\gamma}$ we have $\frac{1}{\sqrt{n}}\sum_{i=1}^n (b_i(\hat{\gamma}) - b_i(\gamma_0))e_i \to_p 0$.

To show $T_{2,n,b} \to_p 0$, first note $\hat{a} \to_p a$. Moreover, $\mathbb{E}[\|b_i(\gamma_0)e_i\|^2] < \infty$ by Assumption 3(ii) and $\mathbb{E}[b_i(\gamma_0)e_i] = 0$, so $\frac{1}{\sqrt{n}}\sum_{i=1}^n b_i(\gamma_0)e_i = O_p(1)$ by Chebyshev's inequality.

For term $T_{3,n}$, we first write

$$T_{3,n} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\dot{k}_i(\tilde{\gamma})' - \sum_{j=1}^J \hat{c}_{ij}\dot{\sigma}_{ij}(\tilde{\gamma})'\right)(\hat{\gamma} - \gamma_0),$$

where $\hat{c}_i = (\hat{c}_{i1}, \ldots, \hat{c}_{iJ})$, and $\dot{\sigma}_{ij}(\gamma) = \frac{\partial \sigma_{ij}(\gamma)}{\partial \gamma}$. Note by construction that $\hat{c}_i$ satisfies the in-sample orthogonality condition

$$\frac{1}{n}\sum_{i=1}^{n} \dot{k}_i(\hat{\gamma}) - \dot{\sigma}_i(\hat{\gamma})'\hat{c}_i = 0. \tag{37}$$

A second mean-value expansion, this time of $\tilde{\gamma}$ around $\hat{\gamma}$, yields

$$\begin{aligned} T_{3,n} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\dot{k}_i(\hat{\gamma})' - \hat{c}_i'\dot{\sigma}_i(\hat{\gamma})\right)(\hat{\gamma}-\gamma_0) \\ &\quad + n^{1/4}(\tilde{\gamma}-\hat{\gamma})'\left(\frac{1}{n}\sum_{i=1}^{n}\left(\ddot{k}_i(\check{\gamma}) - \sum_{j=1}^{J}\hat{c}_{ij}\ddot{\sigma}_{ij}(\check{\gamma})\right)\right) n^{1/4}(\hat{\gamma}-\gamma_0) \\ &=: T_{3,n,a} + T_{3,n,b}, \end{aligned}$$

where $\check{\gamma}$ is in the segment between $\hat{\gamma}$ and $\tilde{\gamma}$, $\ddot{k}_i(\gamma) = \frac{\partial^2 k_i(\gamma)}{\partial\gamma\partial\gamma'}$, and $\ddot{\sigma}_{ij}(\gamma) = \frac{\partial^2\sigma_{ij}(\gamma)}{\partial\gamma\partial\gamma'}$. We have $T_{3,n,a} = 0$ by the in-sample orthogonality condition (37).

To show $T_{3,n,b} \to_p 0$, in view of the condition $\hat{\gamma} = \gamma_0 + o_p(n^{-1/4})$, it is enough to show that the central term in parentheses is $O_p(1)$. To this end, standard arguments (e.g., Newey and McFadden, 1994, Lemma 2.4) using Assumption 3(i) and consistency of $\hat{\gamma}$ yield $\frac{1}{n}\sum_{i=1}^{n}\ddot{k}_i(\check{\gamma}) \to_p \mathbb{E}[\ddot{k}_i(\gamma_0)]$, which is finite. We may similarly deduce by the fact that $\hat{a} \to_p a$ and Assumption 3(i)-(iii) that $\frac{1}{n}\sum_{i=1}^{n}\hat{c}_{ij}\ddot{\sigma}_{ij}(\check{\gamma}) \to \mathbb{E}[c_{ij}\ddot{\sigma}_{ij}(\gamma_0)]$, which is finite, for $j = 1, \ldots, J$. □

Proof of Proposition 6. Analogous to the proof of Proposition 4, using Lemma 2 below in place of Lemma 1. □

Lemma 2. *Let Assumption 3 hold and let $\hat{\gamma} = \gamma_0 + o_p(1)$. Then*

$$\sqrt{n}\hat{S} = Z_n + o_p(1) - (H + o_p(1))(\sqrt{n}(\hat{\gamma}-\gamma_0)),$$

where $Z_n := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\sigma}_i(\gamma_0)'V_i(\gamma_0)^{-1}(d_i - \sigma_i(\gamma_0)) \to_d N(0, H)$.

Proof of Lemma 2. First note that since $\sum_{j=0}^{J}\dot{\sigma}_{ij}(\gamma) = 0$ for all $\gamma \in \Gamma$, we may write

$$\begin{aligned} \sqrt{n}\hat{S} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=0}^{J}\frac{d_{ij}-\sigma_{ij}(\gamma_0)}{\sigma_{ij}(\hat{\gamma})}\dot{\sigma}_{ij}(\hat{\gamma}) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=0}^{J}\frac{\sigma_{ij}(\gamma_0)-\sigma_{ij}(\hat{\gamma})}{\sigma_{ij}(\hat{\gamma})}\dot{\sigma}_{ij}(\hat{\gamma}) \\ &=: T_{1,n} + T_{2,n}. \end{aligned}$$

For $T_{1,n}$, we may rewrite this term using the notation from the proof of Proposition 5 as

$$\begin{aligned} T_{1,n} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_i(\gamma_0) e_i + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(b_i(\hat{\gamma}) - b_i(\gamma_0) \right) e_i \\ &=: T_{1,n,a} + T_{1,n,b} \end{aligned}$$

where $b_i(\gamma) = \dot{\sigma}_i(\gamma)' V_i(\gamma)^{-1}$ and $e_i = d_i - \sigma_i(\gamma_0)$. The summands in $T_{1,n,a}$ have mean zero and variance H, which is finite and non-singular by Assumption 3(iii). Hence, $T_{1,n,a} \to_d N(0, H)$. The proof of Proposition 5 shows that the empirical process $\nu_n(\gamma) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_i(\gamma) e_i$ defined for $\gamma \in N$, a suitable neighborhood of γ_0, is stochastically equicontinuous under Assumption 3(ii). Hence, $T_{1,n,b} \to_p 0$.

For $T_{2,n}$, a mean-value expansion in γ_0 around $\hat{\gamma}$ yields

$$T_{2,n} = \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{J} \frac{\dot{\sigma}_{ij}(\hat{\gamma}) \dot{\sigma}_{ij}(\tilde{\gamma})'}{\sigma_{ij}(\hat{\gamma})} \right) \sqrt{n}(\gamma_0 - \hat{\gamma}),$$

for $\tilde{\gamma}$ in the segment between γ_0 and $\hat{\gamma}$ (with possibly different values for each element). This expansion is valid in view of Assumption 3(i). For the term in parentheses, note

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{J} \frac{\dot{\sigma}_{ij}(\hat{\gamma}) \dot{\sigma}_{ij}(\tilde{\gamma})'}{\sigma_{ij}(\hat{\gamma})} = \frac{1}{n} \sum_{i=1}^{n} \dot{\sigma}_i(\hat{\gamma})' V_i(\hat{\gamma})^{-1} \dot{\sigma}_i(\tilde{\gamma}).$$

Standard arguments (e.g., Lemma 2.4 of Newey and McFadden (1994)) then yield that $\frac{1}{n} \sum_{i=1}^{n} \dot{\sigma}_i(\hat{\gamma})' V_i(\hat{\gamma})^{-1} \dot{\sigma}_i(\tilde{\gamma}) \to_p H$ under Assumption 3(i)(ii). □

Proof of Proposition 7. First note that Assumption 4(i)(ii) implies $\hat{\gamma} \to_p \gamma_0$. Moreover, $\hat{H} \to_p H$ by the proof of Proposition 5, H is positive definite by Assumption 3(iii), and $\hat{G}_\theta \to_p G_\theta$ by Assumption 4(i)(ii). It follows by Assumption 4(ii) that $\hat{M} = I - \hat{H}^{1/2} \hat{G}_\theta' (\hat{G}_\theta \hat{H} \hat{G}_\theta')^{-1} \hat{G}_\theta \hat{H}^{1/2}$ exists wpa1 and $\hat{M} \to_p M$. Now by Assumption 4(iii) and Lemma 2,

$$\sqrt{n}\hat{S} = \sqrt{n} \hat{M} \hat{H}^{-1/2} \hat{S} = \hat{M} \hat{H}^{-1/2} (Z_n + o_p(1)) - \hat{M} \hat{H}^{-1/2} (H + o_p(1)) (\sqrt{n}(\hat{\gamma} - \gamma_0)),$$

where $\hat{M} = I - \hat{H}^{1/2} \hat{G}_\theta' (\hat{G}_\theta \hat{H} \hat{G}_\theta')^{-1} \hat{G}_\theta \hat{H}^{1/2}$. Hence,

$$\sqrt{n}\hat{S} = (M + o_p(1))(H^{-1/2} Z_n + o_p(1)) - (M H^{1/2} + o_p(1))(\sqrt{n}(\hat{\gamma} - \gamma_0)).$$

A mean-value expansion of $\hat{\gamma}$ around (θ_*, e_*) yields

$$\sqrt{n}(\hat{\gamma} - \gamma_0) = (G'_\theta + o_p(1))\sqrt{n}(\hat{\theta} - \theta_*) + (G'_e + o_p(1))\sqrt{n}(\text{vec}(\tilde{e} - e_*)).$$

Since $MH^{1/2}G'_\theta = 0$, we have by Assumption 4(iii)(iv) that

$$\|(MH^{1/2} + o_p(1))(G'_\theta + o_p(1))\sqrt{n}(\hat{\theta} - \theta_*)\| \leq \frac{\epsilon}{1+3\epsilon}\|MH^{1/2}G'_e\sqrt{n}(\text{vec}(\tilde{e} - e_*))\|$$

wpa1. Moreover,

$$\begin{aligned}\frac{1+2\epsilon}{1+3\epsilon}&\|MH^{1/2}G'_e\sqrt{n}(\text{vec}(\tilde{e} - e_*))\| \\ &\leq \|(MH^{1/2} + o_p(1))(G'_e + o_p(1))\sqrt{n}(\text{vec}(\tilde{e} - e_*))\| \\ &\leq \frac{1+2\epsilon}{1+\epsilon}\|MH^{1/2}G'_e\sqrt{n}(\text{vec}(\tilde{e} - e_*))\|\end{aligned}$$

wpa1. We also have $\|MH^{-1/2}Z_n\|^2 \to_d \chi^2_{\text{rank}(M)}$ by Lemma 2, which implies that the inequality $\|(M + o_p(1))(H^{-1/2}Z_n + o_p(1))\| \leq \epsilon C_n$ holds wpa1. Hence, wpa1,

$$\begin{aligned}\frac{1+3\epsilon}{1+\epsilon}&\|MH^{1/2}G'_e\sqrt{n}(\text{vec}(\tilde{e} - e_*))\| + \epsilon C_n \\ &\geq \sqrt{LM_1} \geq \frac{1+\epsilon}{1+3\epsilon}\|MH^{1/2}G'_e\sqrt{n}(\text{vec}(\tilde{e} - e_*))\| - \epsilon C_n.\end{aligned}$$

The first result follows by rearranging. The second is an immediate implication. □

B Additional Results for Section 5

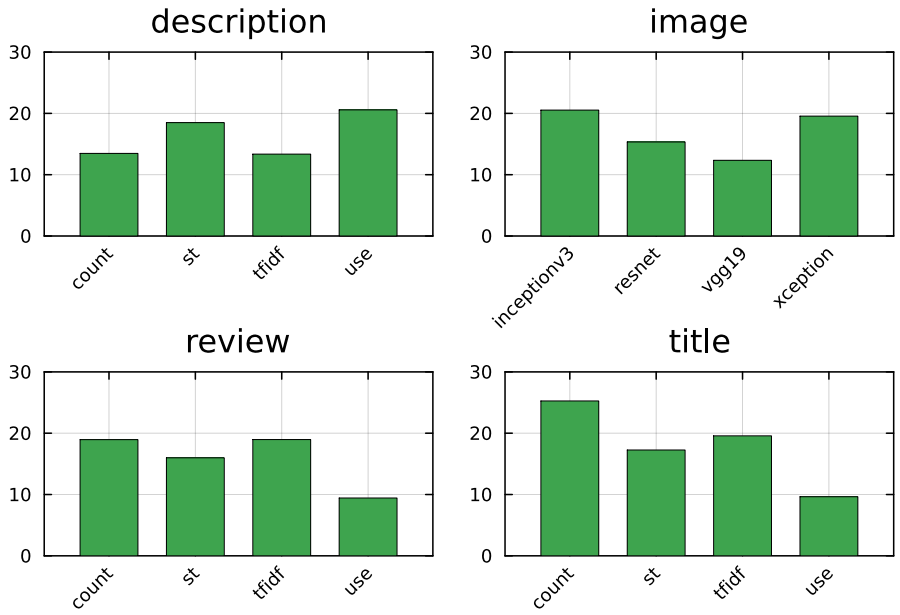
Figures 9 and 10 plot the LM_1 and LM_2 diagnostics, respectively, for each specification.

C Reparameterization for Micro BLP

Consider Example 2. We partition $\beta_i = (\beta_{\bar{x},i}, \beta_{e,i})$ and $\Pi = [\Pi_{\bar{x}}\ \Pi_e\ \Pi_p]$, and write the utilities as:

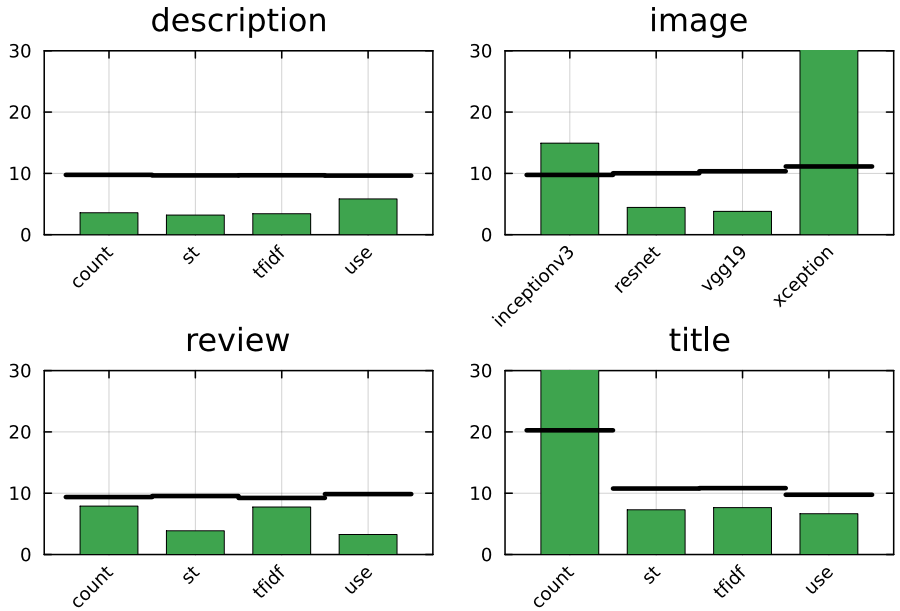
$$u_{ijt} = \beta'_{\bar{x},i}\bar{x}_{jt} + \beta'_{e,i}e_j - \alpha_i p_{jt} + (\bar{x}_{jt}, p_{jt})'[\Pi_{\bar{x}}\ \Pi_p]y_{it} + e'_j\Pi_e y_{it} + \pi'\bar{y}_{ijt} + \xi_{jt} + \varepsilon_{ijt}, \quad j \in \mathcal{J}_t.$$

Figure 9: LM_1 diagnostic in empirical application



Note: Each bar shows the value of the LM_1 statistic for the corresponding specification.

Figure 10: LM_2 diagnostic in empirical application



Note: Each bar shows the value of the LM_2 statistic for the corresponding specification. The horizontal segments show the associated critical values.

Suppose $\alpha_i \sim N(\bar{\alpha}, \sigma_\alpha^2)$, $\beta_{\bar{x},i} \sim N(\bar{\beta}_{\bar{x}}, \Sigma_{\bar{x}})$, $\beta_{e,i} \sim N(\bar{\beta}_e, \Sigma_e)$, and α_i, $\beta_{\bar{x},i}$ and $\beta_{e,i}$ are independent. Then,

$$\theta = \left(\bar{\alpha}, \sigma_\alpha, \bar{\beta}_{\bar{x}}, \bar{\beta}_e, \pi, v(\Pi_{\bar{x}}), v(\Pi_p), v(\Pi_e), l(\Sigma_{\bar{x}}), l(\Sigma_e)\right),$$

where we use the same notation v and l as for Examples 1 and 3.[18] Note that e_j only enters via $\beta'_{e,i}e_j$ and $e'_j\Pi_e y_{it}$. As before, collecting $\beta'_{e,i}e_j$ across products, we have $e\beta_{e,i} \sim N(e\bar{\beta}_e, e\Sigma_e e')$, where $e\Sigma_e e'$ has rank $r \leq J$ because e is $J \times r$. Hence,

$$\gamma(\theta, e) = \left(\bar{\alpha}, \sigma_\alpha, \bar{\beta}_{\bar{x}}, e\bar{\beta}_e, \pi, v(\Pi_{\bar{x}}), v(\Pi_p), v(e\Pi_e), l(\Sigma_{\bar{x}}), l_r(e\Sigma_e e')\right),$$

with l_r as in Examples 1 and 3.

[18] As with Example 1, if $\Sigma_{\bar{x}}$ and/or Σ_e are diagonal, then we replace $l(\Sigma_{\bar{x}})$ and/or $l(\Sigma_e)$ with vectors containing their diagonal entries.