Topics in Applied Econometrics

by

J. Mark Hou

A.B. Mathematics, Princeton University (2011)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2016

Signature redacted

Author .....
Department of Economics
August 15, 2016

Signature redacted

Certified by .....
Jerry A. Hausman
John & Jennie S. MacDonald Professor of Economics
Thesis Supervisor

Signature redacted

Certified by .....
Glenn Ellison
Gregory K. Palm Professor of Economics
Thesis Supervisor

Signature redacted

Accepted by .....
Ricardo Caballero
Ford International Professor of Economics
Chairman, Department Committee on Graduate Theses

Topics in Applied Econometrics

by

J. Mark Hou

Submitted to the Department of Economics
on August 15, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

Chapter 1 focuses on the problem of predicting equilibrium outcomes in large online auction markets. For online retailers, content publishers, and search engines, predicting how the behavior of their auction markets might respond to policy changes is an important business problem. However, this problem is challenging due to both the size and the complexity of such real-world markets. We introduce a method for predicting how various statistics of such markets adjust to changes in supply and demand by: (1) modeling the auction market mechanism as a Walrasian mechanism, (2) coarsening the resulting Walrasian market via a stochastic block model, (3) computing the Walrasian equilibrium of this coarsened market through sampling, and (4) using the resulting equilibrium, together with some reduced-form adjustments, to approximate the equilibrium of the initial auction market. We demonstrate the internal consistency of this method through formal proofs and synthetic experiments, and demonstrates its accuracy by comparison with the equilibrium outcomes of a more realistic pacing-based model of auction markets.

Chapter 2 introduces a model of consumer choice in which consumers simplify their latent high-dimensional preference vector into a low-dimensional one used for choosing products. This assumption induces a particular population structure over consumers' simplified preferences, which allows for tractable estimation in high dimensional settings. Estimation is performed via a stochastic gradient descent-based algorithm, and we evaluate its performance through a variety synthetic benchmarks. We also estimate the model on consumer consideration data, finding that the average consumer uses only 6 of 16 product attributes when forming their consideration set, and that this leads to a utility of loss of $2-3\%$ on average.

Chapter 3 uses admissions data from the University of Bologna's medical school to analyze how students' entrance exam rankings affect their subsequent academic performance. We find that: (1) worse rankings lead to worse academic performance, (2) this impact is more negative for worse-ranked students, (3) this impact on academic performance operates mostly through courseload rather than through GPA, and (4) male and female students' academic performance do not respond differentially to rank.

Thesis Supervisor: Jerry A. Hausman
Title: John & Jennie S. MacDonald Professor of Economics

Thesis Supervisor: Glenn Ellison
Title: Gregory K. Palm Professor of Economics

Acknowledgments

I am tremendously grateful to my advisors, Jerry Hausman and Glenn Ellison, whose guidance and encouragement over the last few years were absolutely crucial to the completion of this thesis. I benefited immensely from Jerry's deep knowledge of the various technical areas relevant to my work, as well as from his willingness to supervise me on topics of my choosing. I am also extremely lucky to have benefited from Glenn's broad experience in both economic theory and applied econometrics, as well as from his pithy and high-resolution advice on some of the fine details of this thesis.

In addition to my advisors, I would also like to thank all the faculty at MIT and elsewhere who provided feedback and other assistance over the past few years. In particular, I am very grateful to John Hauser and Victor Chernozhukov for their detailed feedback during the early stages of the second chapter of this thesis, and to Olivier Toubia and John Hauser for their kindness in providing the data used in that same chapter. I am also very much indebted to Muhamet Yildiz and Bob Gibbons, from whom I learned a great deal about microeconomic theory and organizational economics, and also to Ben Olken for giving me my first introduction to applied econometrics research.

A large portion of the first chapter of this thesis was completed during my internship at Facebook's economic research group. I am very grateful to everyone in economic research and core data science for all of their comments, technical advice, and camaraderie over the course of those very productive three months. In particular, I would like to thank Eric Sodomka and Nico Stier for all their hard work, both during the summer and after, as project could not have come to fruition without their extensive contributions, reassurances, and motivation. I am also greatly indebted to Mike Bailey, who brought me into the economic research group and also taught me a great deal about academic research in the private sector.

I am extremely lucky to have had many wonderful classmates over the last few years at MIT. In particular, I would like to thank Enrico Cantoni, who kindly lent me his first-hand understanding of the Italian university system, as well as the data used in the third chapter of this thesis. I am also very fortunate to have been in the company of some incredibly capable classmates outside of MIT economics. In particular, I would like to thank Peter Diao for our numerous illuminating discussions on the theory of dense graph limits, which formed an important part of the first chapter of this thesis.

Finally, I would like to thank my parents, Dingchen Hou and Xiaolan Zhuang, to whom my debt is inexpressibly great. It is only by their immense efforts over the last few decades, and their unwavering support over the last few years, that I have been granted the luxury of being able to complete this thesis.

Chapter 1

Equilibrium Prediction in Large Online Auction Markets

with Eric Sodomka and Nicolas E. Stier-Moses

1.1 Introduction

How should one model a complex, dynamic, real-world market when the goal is to make practical decisions in that market? How does the market structure, valuation environment, and decision space affect the accuracy of using a given model?

We frame our problem in the context of *online auction marketplaces*, a key differentiating factor of search engines and content publishers today. In such markets, agents express to the auctioneer or publisher (e.g., eBay, Amazon, Google, Facebook, Google) how much they are willing to pay in exchange for the opportunity to show products, text, images, links or videos to different users along with (possibly) some budget constraints. When a user arrives to the publisher's site, the publisher decides what content to show to the user based on the outcome of the auction. The most widely-used approach today is to run a separate auction every time a user searches for a keyword or loads a page, with per-auction bids set by a proxy bidder (run by the publisher on the agent's behalf) that attempts to maximize the agent's utility while satisfying any budget constraints. Agents can update their expressed preferences at any time.

In this work, we take the perspective of the publisher. The publisher is responsible for designing a content delivery mechanism that decides how users and agents express their preferences and constraints, what content is shown to users when they arrive, and

how agents pay when their content is delivered. There often exists much complexity in the real-world environment (e.g., stochasticity, partial observability, dynamic actions and revelation of information, heterogeneous participants) that propagates into complexity in the mechanism. Furthermore, the mechanism is continually evolving as the publisher better understands how to help agents and users better express their preferences and achieve their desired objectives. Given this complexity, even answering a basic question about what market outcomes arise for some given mechanism inputs can prove to be difficult.

1.1.1 A Key Challenge: Predicting Counterfactual Market Outcome Summary Statistics

This work does not directly tackle the mechanism design problem. Rather, for a given mechanism and its corresponding inputs, we aim to *predict relevant market outcome summary statistics.* In the general case, the mechanism inputs are a profile of agents and users along with distributions over agents' and users' private information. The publisher might use such predictions to: (1) improve the mechanism by understanding how parameters of the mechanism correlate with outcomes; (2) understand various agent groups and take marketing actions to increase the value provided to them, which will be reflected in new mechanism inputs; and (3) decide a course of action based on the mechanism outputs (e.g., internal hiring decisions, resource allocation, etc.).

To make informed decisions in any of those categories, the publisher may require predictions on inputs that differ significantly from historical data. Examples of relevant inputs include those that capture drastic changes in foreign currency exchange rates, user demographics, or available inventory to be auctioned. Lower-dimensional representations of the outcome space may be sufficient for making decisions as those above since only summary statistics are required. Broadly speaking, the publisher may measure how a particular market outcome will affect its short- and long-term utility. The long-term utility of the publisher depends on the short-term utilities of agents and users. At the coarsest level of granularity, the publisher could reason about a single metric for total agent and user value, or it could reason about metrics for particular segments of agents and users.

Because our focus is on a real-world prediction problem, it is worth noting that reality surfaces considerations that affect the value of a given methodology to test counterfactuals. Besides first-order impacts such as computational and engineering

resources, the approach should be *minimally obtrusive to agents and users.* If counterfactual experiments are run, for example, one must trade off the gains from improved prediction accuracy with the short-term costs of running those experiments. Additionally, there is a preference for *simplicity*: the methodology in place should require minimal maintenance when mechanism details or other dependencies change.

A natural first thought is, *if we are given the mechanism and its inputs, why not simply run the mechanism to compute its outputs?* Such a simulation-based solution would give us exactly the answer we are looking for. However, obtaining market outcomes requires running one or many days worth of auctions which makes the approach computationally costly. Further, finding accurate forecasts for the inputs at the level of granularity needed to run the actual mechanism is difficult, especially when all that is needed are some aggregate summary statistics of the output.

Another natural question is, *if we have historical inputs and outputs to the mechanism, why not use standard machine learning techniques to predict expected outputs?* The main challenge with this approach is that it may perform poorly for predicting on inputs that are vastly different from those in the training set. A black-box machine learning approach does not take advantage of the publisher's knowledge of its mechanism for out-of-sample predictions. The computational study we conduct provides evidence of how much we can improve basic prediction algorithms when augmented with the output of our model.

1.1.2 Our Approach: Market Abstractions, Machine Learning, and Counterfactual Measurements

Our approach combines market modeling, equilibrium computation, machine learning techniques, and counterfactual measurements, combining the merits of each technique while mitigating the aforementioned costs. Rather than computing outcomes using the specification of the actual mechanism, which is prohibitively expensive, the core of our methodology is to make *market abstractions* that capture the supply and demand dynamics of the original market while leaving residual complexity to be accounted for in a reduced-form way. We compute equilibria for the simplified, abstract market, and project the resulting summary statistics onto the original problem. Figure 1-1 provides an illustration of the core market abstraction methodology.

When choosing a market abstraction, we trade off computational tractability for prediction accuracy: with an overly-stylized model, we may abstract away features of

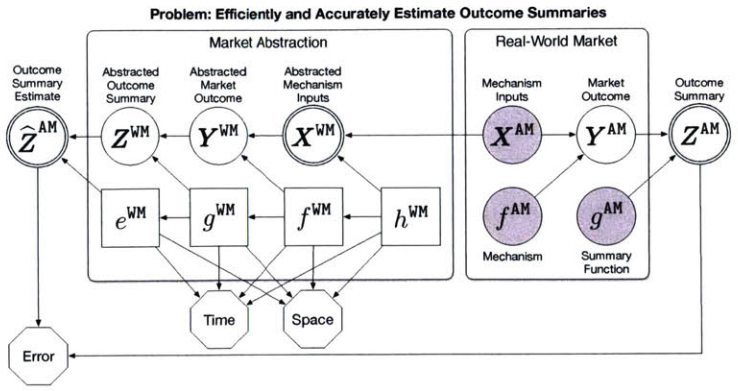


Figure 1-1: An overview of the approach taken. Octagons are utility nodes. Squares are decision nodes. Circles are chance nodes. Double-lined circles are deterministic given the node inputs. Shaded nodes are observed at the time at which decisions are made.

the original problem that have a significant effect on the outcome of interest. Combining the output of our market-based predictions with machine learning allows us to account for some systematic biases that arise from our market abstraction. Because historical data may lack diversity of inputs, machine-learning-based corrections applied directly to market-based predictions may be inaccurate for inputs that are far out of sample. Hence, the last component of the methodology is using counterfactual experiments to add diversity to the training data. By training on measurements that correspond to an altered amount of supply and demand in the system, we mitigate the potential errors from out-of-sample predictions.

The main focus of this paper is on the choice of market abstraction. The strategy we put forward is to transform the actual mechanism into a *Walrasian market*, solve for equilibrium, and then use the equilibrium outcome to predict summary statistics for the actual market. We propose to use a Walrasian market to simplify real-world mechanism dynamics, and a *Walrasian stochastic block model* as an optional second layer to further reduce the dimensionality of the input space. In addition, we propose a *pacing market* as an artificial model that represents an auction market where agents compete in multiple second-price single-slot auctions. Equilibria of that market provide us with artificial ground-truth summary statistics, which we use as a target of the computational simulations and evaluate the Walrasian approximations. The three abstractions are summarized below.

Walrasian markets (WM) A WM captures a situation with a fixed number of buyers and goods along with a specification of each buyer's valuation for each possible

bundle of goods. A *Walrasian market equilibrium* is a set of prices for each good and an allocation of goods to buyers such (1) that each buyer maximizes her utility at the given allocation and (2) the market clears. The Walrasian markets we consider satisfy strict gross substitutes and thus have a unique Walrasian equilibrium Arrow et al. (1959). To solve for equilibrium, we use an existing tatonnement algorithm that is guaranteed to converge Cole and Fleischer (2008).

Walrasian stochastic block model (WSBM) We cannot compute Walrasian equilibria directly on real-world data because of the intractability of the input size. Reducing the input size motivated us to consider a WSBM, which provides a more compact description of an instance, encoded as a weighted bipartite graph. We generate an instance by taking a general WM and clustering it into another WM with similar structural properties but of lower dimension. After the transformation, we proceed similarly to the steps described earlier: we solve for equilibrium and output the relevant summary statistics. Through computational simulations, we found that the summary statistics of interest for WSBMs converged to those in the original WM as the instance size increased. We provide a corresponding theoretical result on convergence that applies under some additional assumptions.

Pacing market abstraction Our computational simulations are performed in the context of a marketplace that runs multiple second-price single-slot auctions, which we refer to as a *pacing market.* Under this setting, we can compute both Walrasian and pacing equilibria, and compare the results to understand how well one can be used to predict the other. In the pacing market, there are buyers (agents that participate in the auction) with budgets, and goods (auctions), and each buyer has some valuation for each good. However, unlike a WM, in the pacing market each buyer's action consists of a single pacing multiplier that controls how much a bidder scales down her bids relative to her valuations. The effect of this multiplier is to pace the spending and exhaust the budget only at the end of the auctions, while keeping the mechanism incentive compatible. This definition was inspired by some mechanisms put in place by several online marketplaces, and hence is a good yardstick to evaluate our predictors in a more controlled environment.

1.1.3 Main Results and Contributions

We formalize the market outcome summary statistic prediction problem and present a method for making predictions that combines equilibrium computation in abstract markets, machine learning, and counterfactual measurements. Our first main result is that, although the Walrasian market abstraction may overestimate the publisher's utility, it is able to predict the shape of the curve as one uniformly varies the budgets across agents (Figure 1-5). When the Walrasian market output is combined into an ML model with other market features, the bias is corrected and the method significantly improves predictions of the auctioneer's utility over alternative approaches (Figure 1-6, top). Compared to these alternative approaches, the Walrasian market approach suffers less from extrapolation errors when data from training and test sets are drawn from different distributions (Figure 1-6, bottom), which is the practical problem of interest for estimating counterfactuals. We find that even a small amount of counterfactual data is sufficient to accurately estimate the full publisher's utility curve (Figure 1-7). The proposed method thus appears to be most promising in answering what-if questions that require out-of-sample analysis and for which running a large number of experiments is expensive, impractical or time-consuming.

Paper Organization The remainder of the paper proceeds as follows. Section 1.2 comments on related work in the literature. In Sect. 1.3, we describe our model of online auction markets. Section 1.4 introduces the main approach, consisting of Walrasian market and Walrasian stochastic block model abstractions, supplemented with machine learning and counterfactual measurements. We test the internal consistency of the approach in Sect. 1.5 by showing that the algorithms converge as expected. In Sect. 1.6, we present the experimental setup by specifying the pacing market that we use as ground truth and a suite of baseline predictors. Section 1.7 evaluates our approach on the synthetic problem instances introduced earlier. We conclude in Sect. 1.8 with discussion and future research directions.

1.2 Related Work

Our market abstraction approach is related to a growing literature on *game abstractions*; see Sandholm (2015) for a review. The goal of that work is to solve for Nash equilibria in a game for which doing so is intractable because the game is either

too large, in a class for which solutions are hard to compute, or of such complexity that the game itself is difficult to fully specify. At a high level, the approach takes a game as input, converts it into an abstracted game, solves that abstracted game for equilibrium, and projects abstracted game strategies into full game strategies. Main developments include automating the abstraction process Ganzfried and Sandholm (2014) and providing theoretical bounds on resulting strategies Sandholm and Singh (2012); Kroer and Sandholm (2014). Our work most closely maps to this line of work when we consider that the proxy bidders in our domain take the bids submitted by agents as their ground truth values, and then potentially find proxy bids that result in equilibrium. A key difference in our work stems from the objective: we are interested in computing relevant summary statistics of the outcome rather than the full equilibrium strategy profile. This lower-dimensional objective provides more flexibility in choice of abstraction, since we do not have to project abstracted-game strategies back into full-game strategies. Another difference is that much of the literature above is focused on abstractions for general games, whereas there is considerable market structure in our domain that we are willing to exploit.

Our market abstraction approach is perhaps most directly related to that of Wellman et al. (2004), who used a Walrasian equilibrium abstraction to predict market prices in TAC Travel (a simulated market game in which agents participate in simultaneous and sequential auctions of various forms). Despite no guarantees of existence of Walrasian equilibrium in that domain, nor guarantees that such an equilibrium would be reached in practice, the authors showed that their approach accurately predicted market prices. We similarly simplify many complexities of our mechanism through the Walrasian market assumption. One difference in our work is that, because our real-world domain is so large, we require additional abstractions beyond the Walrasian market abstraction (specifically, the Walrasian stochastic block model). Regardless, this work can be seen as treating the positive TAC Travel results as a hypothesis that such a WM abstraction may be useful in a real-world market environment; our results provide supporting evidence for that hypothesis.

1.3 Auction Market Models

This section describes a family of stylized models that capture some of the intricacies of real-world auction markets. In an online marketplace, agents make decisions in

a high-dimensional action space. An agent can state its *budget*, which specifies the maximum amount that it is willing to spend on auctions within a specified time range. For a set of auctions, the agent states its *targeting criteria*, which represents keywords or users in which the agent is interested, and *per-auction bids*, indicating the maximum amount the agent is willing to pay if the item is selected (or other events depending on the cirumstance, e.g., a conversion, a click, a video view or some other engagement metrics). The agent can pause or un-pause its participation or modify any charasteristic of its participation in the auction marketplace at any time.

When users visit the publisher's site, enter a keyword or interact with the pages, the publisher runs the corresponding auctions and subsequently show the outcome to the users. At auction time, the publisher also decides the payments from the agent according to the pre-established mechanism. Different platforms make use of different mechanisms at this stage. Some publishers explicitly aim for a system that is incentive compatible and maximizes social welfare, while others employ mechanisms that maximize its utility. The motivation for incentive compatibility is to enable agents to invest their resources into generating compelling content rather than developing sophisticated bidding strategies. The motivation for maximizing social welfare is to maximize agents' return on investment to generate a healthy ecosystem with more inflow of agents, which in turn improves user experience (i.e., allowing for better matchings of agents' content to users).

To respect each agent's budget constraint, each proxy bidder uses a *pacing parameter* to update its per-auction bid with the goal of spending its budget smoothly over time. The pacing mechanism as a whole can be thought of as a game between proxy bidders and its solution is referred to as a *pacing equilibrium.*

1.3.1 A Stylized Representation of an Auction Market

We consider a simplified representation of the real-world mechanism alluded to above, as well as to the representations of its inputs and outputs. The model simplifies the following aspects:

1. Agent actions are chosen in a single stage rather than changing over time.
2. Each agent has a single budget constraint.
3. Each user makes a single request (i.e., there are no cross-auction dependencies aside from the agent's budget constraint).

4. Predictions about event probabilities are perfect.

5. Each auction is for a single slot.

6. There is a single event type (i.e., a click).

7. The mechanism is not evolving over time.

8. The proxy bidders' decisions are made offline (i.e., under full-information).

More formally, from the perspective of the publisher, we are given a set of auction items and impression opportunities as input. Each auction item has a prespecified per-click bid, budget, and targeting criteria (which determines which auction items can be shown for which queries). Each query can accommodate a single auction item (i.e., auctions are for a single slot). We are given predictions about the probability a given auction item will be clicked on if it is shown for a given query. An *auction market* consists of the following input data:

Definition 1 (AM). *An* auction market $\boldsymbol{x}^{AM} = (n, m, \boldsymbol{v}, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{\gamma})$ *is a tuple where* $n \in \mathbb{N}$ *is the number of auction items,* $m \in \mathbb{N}$ *is the number of queries,* $\boldsymbol{v} = (v_i \in \mathbb{R}_{\geq 0})_{i \in [n]}$ *is each auction item's stated value per click,* $\boldsymbol{b} = (b_i \in \mathbb{R}_{\geq 0})_{i \in [n]}$ *is each agent's stated budget,* $\boldsymbol{t} = (t_{i,j} \in \{0, 1\})_{i \in [n], j \in [m]}$ *is the targeting criteria for each (auction item, query) pair, and* $\boldsymbol{\gamma} = (\gamma_{i,j} \in [0, 1])_{i \in [n], j \in [m]}$ *is the estimated click-through rate (CTR) for each (auction item, query) pair.*

Taking the auction market as input, the publisher determines pacing multipliers for each bidder, which determines who wins each auction and how much they must pay. Users then probabilistically click on the auction items. The result of the publisher's decision about the allocation and payments, as well as the result of the user's decision to click on auction items, is called the *auction market outcome.*

Definition 2 (Auction Market Outcome). *An outcome* $\boldsymbol{y}^{AM} = (\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{c})$ *represents an allocation* $\boldsymbol{a} = (a_{i,j} \in \{0, 1\})_{i \in [n], j \in [m]}$, *where* $a_{i,j}$ *indicates whether auction item* i *was allocated to query* j*; the conditional payments* $\boldsymbol{p} = (p_{i,j} \in \mathbb{R}_{\geq 0})_{i \in [n], j \in [m]}$, *where* $p_{i,j}$ *is auction item* i*'s payment for query* j *if it receives a click; and the realized clicks* $\boldsymbol{c} = (c_{i,j} \in \{0, 1\})_{i \in [n], j \in [m]}$, *where* $c_{i,j}$ *indicates whether a click occurred for auction item* i *on query* j.

The algorithm that generates these outputs is defined by function $f^{\mathtt{AM}} : \boldsymbol{X}^{\mathtt{AM}} \to \boldsymbol{Y}^{\mathtt{AM}}$, which depends on the publisher's mechanism and user clicking behavior. We also consider a function $g^{\mathtt{AM}} : \boldsymbol{Y}^{\mathtt{AM}} \to \boldsymbol{Z}^{\mathtt{AM}}$ that, given an auction market outcome $\boldsymbol{y}^{\mathtt{AM}} \in \boldsymbol{Y}^{\mathtt{AM}}$, outputs a summary statistic of interest (of arbitrary dimension) to the publisher. In subsequent experiments, we treat the publisher's utility as the summary statistic of interest, $z^{\mathtt{AM\text{-}Ut}} = g^{\mathtt{AM\text{-}Ut}}(\boldsymbol{y}^{\mathtt{AM}}) := \sum_{i=1}^{n}\sum_{j=1}^{m} p_{i,j}c_{i,j}$, given by the sum of payments the publisher receives from the agents.

1.4 Prediction with Walrasian Market Abstractions

With the concept of auction market in place, we now describe our method for predicting market outcome summary statistics through a Walrasian market abstraction. For completeness, we start with the definition of a Walrasian market.

Definition 3 (WM). *A Walrasian market $\boldsymbol{x}^{\mathtt{WM}} = (n, m, \boldsymbol{e}, \boldsymbol{u})$ is a tuple where $n \in \mathbb{N}$ is the number of buyers, $m \in \mathbb{N}$ is the number of goods, $\boldsymbol{e} = (e_{i,j} \in \mathbb{R}_{\geq 0})_{i\in[n],j\in[m]}$ is the initial endowment for each (buyer i, good j) pair, and $\boldsymbol{u} = (u_i : \mathbb{R}^m_{\geq 0} \to \mathbb{R})_{i\in[n]}$ denotes the concave utility functions for each buyer.*

A Walrasian market outcome $\boldsymbol{y}^{\mathtt{WM}} = (\boldsymbol{a}, \boldsymbol{p})$ is an allocation of goods $\boldsymbol{a} = (a_{i,j})_{i\in[n],j\in[m]}$ and prices $\boldsymbol{p} = (p_j)_{j\in[m]}$, where $a_{i,j} \in \mathbb{R}_{\geq 0}$ is buyer i's allocated quantity of good j and $p_j \in \mathbb{R}_{\geq 0}$ is the unit price for good j. A *Walrasian market equilibrium* is a vector of prices (one per good) and an allocation of goods to buyers such that (1) each buyer is maximizing its utility at the given allocation and (2) the market clears.

Definition 4 (Walrasian Equilibrium (WME)). *A Walrasian outcome $\boldsymbol{y}^{\mathtt{WM}} = (\boldsymbol{a}, \boldsymbol{p})$ is an equilibrium for WM $\boldsymbol{x}^{\mathtt{WM}} = (n, m, \boldsymbol{e}, \boldsymbol{u})$ if and only if, for all $i \in [n]$,*

$$a_i \in \text{argmafx}\{u_i(\boldsymbol{a}_i) : \boldsymbol{a}_i \in \mathbb{R}^m_{\geq 0}, \boldsymbol{p} \cdot \boldsymbol{a}_i \leq \boldsymbol{p} \cdot \boldsymbol{e}_i\}, \tag{1.1}$$

where $\boldsymbol{e}_i = (e_{i,j})_{j\in[m]}$ and $\boldsymbol{a}_i = (a_{i,j})_{j\in[m]}$.

1.4.1 Transformation to a Walrasian Market

We now describe the transformation from AM to WM. Each auction item and query in the AM represents a buyer and good in the WM, respectively. The WM also contains an additional buyer representing the publisher and an additional good representing

money. Each buyer in the WM other than the publisher is given an initial endowment of money equal to its budget and no endowment for any other good. The publisher is given no endowment of money and a unit endowment of every other good. We start by considering buyers with linear utility functions. Given an allocation, a buyer's expected utility is defined as the sum of all per-impression values corresponding to goods the buyer was allocated. This transformation is summarized in Definition 5. When there is a possibility of confusion, we add a superscript to indicate whether a term is associated with one market or the other.

Definition 5 (Walrasian Market Transformation (Linear Utilities)). *Let* $h^{\text{WM-Linear}} : X^{\text{AM}} \to X^{\text{WM}}$ *be a function that takes as input an auction market* $\boldsymbol{x}^{\text{AM}} = (n^{\text{AM}}, m^{\text{AM}}, \boldsymbol{v}^{\text{AM}}, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{s}, \boldsymbol{\gamma})$ *and outputs a Walrasian market* $\boldsymbol{x}^{\text{WM}} = (n^{\text{WM}}, m^{\text{WM}}, \boldsymbol{e}, \boldsymbol{u})$ *such that* $n^{\text{WM}} := n^{\text{AM}} + 1$, $m^{\text{WM}} := m^{\text{AM}} + 1$, $\boldsymbol{e} = (e_{i,j})_{i \in [n^{\text{WM}}], j \in [m^{\text{WM}}]}$ *given by* $e_{i,j} := \{b_i^{\text{AM}}$ *if* $i \in [n^{\text{AM}}]$ *and* $j = m^{\text{AM}} + 1$; 1 *if* $i = n^{\text{AM}} + 1$ *and* $j \in [m^{\text{AM}}]$; 0 *otherwise*$\}$, *and* $\boldsymbol{u} = (u_i)_{i \in [n^{\text{WM}}]}$ *given by* $u_i(\boldsymbol{a}_i) = v_i \sum_{j=1}^{m^{\text{WM}}} a_{i,j} t_{i,j} \gamma_{i,j} + a_{i,m^{\text{WM}}}$.

We consider *constant elasticity of substitution* (CES) utilities, which take an additional parameter σ and approximate linear utilities as σ approaches infinity. To solve for WME, we use the distributed tatonnement algorithm of Cole and Fleischer (2008), which we refer to as f^{WM}, and whose API is given in Alg. 1. All inputs to f^{WM} other than $\boldsymbol{x}^{\text{WM}}$ are tuning parameters that we subsequently suppress for readability. Appendix 1.B provides further details on the CES approximation while Appendix 1.C describes f^{WM}.

Algorithm 1: $f^{\text{WM}}(\boldsymbol{x}^{\text{WM}}; \boldsymbol{p}^{\text{init}}, \lambda, \epsilon, \kappa)$

Input: $\boldsymbol{x}^{\text{WM}} = (n, m, \boldsymbol{e}, \boldsymbol{u})$: Walrasian market
$\boldsymbol{p}^{\text{init}}$: initial price vector
λ : learning parameter
ϵ : convergence tolerance
κ : max num iterations
Output: $\boldsymbol{y}^{\text{WM}} = (\boldsymbol{a}, \boldsymbol{p})$: Walrasian outcome

The WM overall social welfare $g^{\text{WM-SW}}(\boldsymbol{y}^{\text{WM}}) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{i,j} p_j$ is simply the sum of all payments from each buyer in the WM. Because goods and buyers in the WM map directly to users and agents in the AM, converting from WM social welfare to AM publisher's utility is the identity function $e^{\text{WM-AM}}(x) = x$. We combine these transformations and equilibrium-solving algorithms into function $\phi^{\text{WM}}(\boldsymbol{x}^{\text{AM}})$ to get the estimated

AM publisher's utility:

$$\phi^{\texttt{WM}}(\boldsymbol{x}^{\texttt{AM}}) := e^{\texttt{WM-AM}}(g^{\texttt{WM-SW}}(f^{\texttt{WM}}(h^{\texttt{WM-Linear}}(\boldsymbol{x}^{\texttt{AM}})))).$$

1.4.2 Walrasian Stochastic Block Model Abstraction

We now describe a method that encodes a WM as a *stochastic block model* and uses such a representation to compute the corresponding output summary statistics for the original AM.

Definition 6 (WSBM). *A* Walrasian stochastic block model *is a tuple* $\boldsymbol{w}^{WM} = (n, m, c, d, \xi^{buyer}, \zeta^{good}, \boldsymbol{g}^{budget}, \boldsymbol{g}^{value})$ *where* $n \in \mathbb{N}$ *is the number of buyers,* $m \in \mathbb{N}$ *is the number of goods,* $c \in \mathbb{N}$ *is the number of buyer types,* $d \in \mathbb{N}$ *is the number of good types,* $\xi^{buyer} : [c] \to [0, 1]$ *is an i.i.d. distribution over buyer types,* $\zeta^{good} : [d] \to [0, 1]$ *is an i.i.d. distribution over good types,* $\boldsymbol{g}^{budget} = (g_i^{budget} : \mathbb{R}_{\geq 0} \to [0, 1])_{i \in [c]}$ *is a tuple of i.i.d. distributions over budgets for buyers of each type, and* $\boldsymbol{g}^{value} = (g_{i,j}^{value} : \mathbb{R}_{\geq 0} \to [0, 1])_{i \in [c], j \in [d]}$ *is a tuple of i.i.d. distributions over values for each (buyer type, good type) pair.*

To construct the WSBM, we choose partitions over auction items and queries equal to the market segments used for computing summary statistics, and then compute stochastic block model parameters from empirical distributions of the input WM. Then, for that WSBM, we sample multiple WMs of a particular market size, solve each for equilibrium via $f^{\texttt{WM}}$, compute the mean of the relevant summary statistics, and scale up the summary statistics for the larger market. We repeat the above procedure, sampling WMs of increasingly larger size from the WSBM, until the summary statistics converge or the size of the original WM is reached. A formal specification of the procedure is given in Appendix 1.D, Algorithms 2-6.

1.4.3 Combining Market Abstractions and Machine Learning

Our WM and WSBM abstractions simplify away some of the complexity in the system, which could introduce estimation error. We thus combine the abstraction-based predictions with standard machine learning techniques, where the abstraction-based prediction is a feature in the machine learning model.

Letting $\hat{\pi}_S(\boldsymbol{x}^{\text{AM}})$ be the predicted publisher's utility for auction market input $\boldsymbol{x}^{\text{AM}}$, the predictors we evaluate are linear regression models defined by a set of basis functions $S = \{\phi_1, \ldots, \phi_k\}$ according to the expression

$$\hat{\pi}_S(\boldsymbol{x}^{\text{AM}}) = \hat{\beta}_0 + \sum_{j=1}^{k} \phi_j(\boldsymbol{x}^{\text{AM}})\hat{\beta}_j \,. \tag{1.2}$$

We consider auction market features that summarize agents, users, edges, and other graph structure parameters. The specific features we use for our experiments are described in Sect. 1.6.4.

1.4.4 Training with Counterfactual Experiments

Supply and demand variations in real life do not always provide a rich enough dataset to accurately predict counterfactual outcomes. This section describes a procedure for adding diversity to training data through counterfactual experimentation.

Suppose the publisher wishes to predict counterfactual utility if all agents increased their budgets by a factor of α. The intuition for the counterfactual experiment is shown in Figure 1-2. For simplicity, suppose there is only one agent in the system whose budget is normalized to 1, and suppose there is some number of opportunities to participate in auctions, also normalized to 1 (this is the state shown on the LHS of Fig. 1-2).

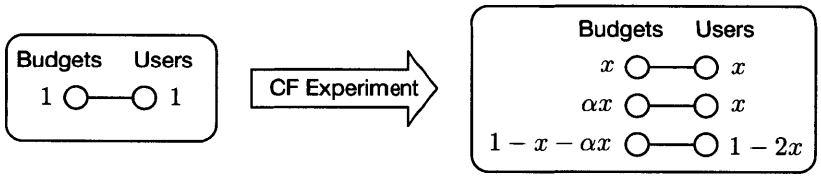


Figure 1-2: An illustrative example of a counterfactual experiment in which budgets are increased by multiplicative factor α.

A counterfactual experiment could be conducted by splitting each agent into multiple *sub-agents*, where the sub-agents' budgets sum to the original agent's budget. The population of impressions is similarly split and the mechanism is run independently for these groups. In the RHS of the figure, the ratio between the resulting publisher's utility from the top two groups indicates how much utility would be achieved if budgets increased by α.

This procedure can be applied to simulate different levels of supply and demand,

which can be used as training data for market inputs that would otherwise not be observed. However, running such an experiment takes significant time and computational resources, and also potentially decreases market efficiency (since users from one group are not exposed to sub-agents in another group). It is thus preferable to accurately predict counterfactual outcomes with as little counterfactual experimentation data as possible. We explore prediction accuracy as a function of counterfactual training data in Sect. 1.6.

1.5 Evaluation of Internal Consistency

Prior to evaluating the accuracy of our approach, we ran experiments to evaluate internal consistency of our WM- and WSBM-finding algorithms. Details of these experiments are presented in Appendix 1.E. At a high level, we found the following, as illustrated in Figure 1-3:

1. For sufficiently low convergence tolerance thresholds, $f^{\texttt{WM}}$ produced consistent summary statistic values, regardless of initial price vector (Figure 1-3a).

2. For sufficiently high CES parameters σ, the tatonnement algorithm $f^{\texttt{WM}}$ produced consistent summary statistic values, regardless of initial price vector (Figure 1-3b).

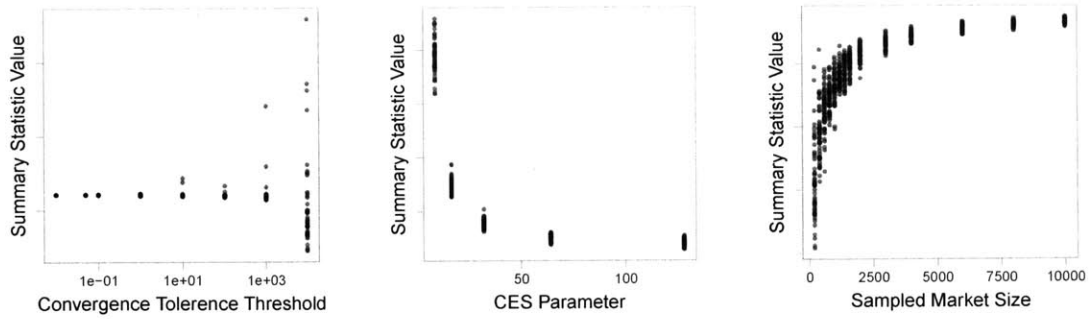
3. As the WSBM algorithm sampled larger markets, the amount of variation in the summary statistic decreased and began to converge (Figure 1-3c).

We used the results of these experiments to set algorithm tuning parameters. The particular values chosen are described in Appendix 1.E.

1.6 Experimental Setup

This section describes the sets of experiments used to validate our approach. We provide further details on the ground truth mechanism, the problem instance distribution, the market outcome summary statistic, and relevant baseline predictors. This study helps answer the following questions (pointers to figures referring to those questions are in parenthesis):

1. How well does the WM abstraction capture the structure of the more complex pacing model (Figure 1-5)?



(a) Tatonnement Convergence (b) CES Approximation (c) Sampling Consistency

Figure 1-3: Example of a synthetic output. a The average price of all goods within a representative good type, computed via $f^{\texttt{WM}}$ for different initial prices $\boldsymbol{p}^{\texttt{init}}$ and convergence tolerance thresholds ϵ. b The equilibrium average price of all goods within a representative good type, computed for a variety of different elasticity of substitution parameters σ. c The equilibrium average price of all goods within a representative good type, computed from sampled markets of increasing size.

2. How well do the extra features, in addition to the WM, predict publisher's utility compared to competitive baselines (Figure 1-6 top)? How well can we predict when considering inputs that are generated from a distribution different from what has been observed previously (Figure 1-6 bottom)?

3. How well does supplementing training data with counterfactual experiments improve predictions (Figure 1-7)?

The high-level description of the evaluation procedure is as follows: We first generated a set of AM inputs, which we call the *base problem instances*. For each base problem instance, we also generated a set of *counterfactual problem instances*, which are exactly the same as the base instances but with budgets uniformly multiplied by a constant. For each problem instance $\boldsymbol{x}^{\texttt{AM}}$, we generated the ground truth summary statistic $\pi^*(\boldsymbol{x}^{\texttt{AM}}) := g^{\texttt{AM-Ut}}(f^{\texttt{AM}}(\boldsymbol{x}^{\texttt{AM}}))$ and extracted features used by various predictors. For each predictor $\hat{\pi}_S$, defined by a set of features S including or not the publisher's utility in the WM, we ran K-fold cross validation (for $K \in \{2, 5, 10, 20\}$) with the following types of train/test splits: (1) training and test data contain only base instances, (2) training data contains base instances while test data contains counterfactual instances, (3) training and test data both contain counterfactual instances.

We evaluated the performance of a given predictor $\hat{\pi}_S$ over test instances $\boldsymbol{x}_1^{\texttt{AM}}, \ldots, \boldsymbol{x}_N^{\texttt{AM}}$

through mean absolute percent error:

$$\texttt{MAPE(S)} = \frac{100}{N}\sum_{k=1}^{N}\frac{|\hat{\pi}_S(\boldsymbol{x}_k^{\texttt{AM}}) - \pi^*(\boldsymbol{x}_k^{\texttt{AM}})|}{\pi^*(\boldsymbol{x}_k^{\texttt{AM}})}.$$

1.6.1 Problem Instances

We opted to evaluate on synthetic problem instances for this computational study because it allowed us to explicitly specify the values of all experiment parameters. We evaluate the robustness of results by considering multiple input distributions. For readability, in this section we focus on a single problem instance distribution. Results for different distributions were qualitatively similar and some are shown in Appendix 1.A.

Recall that an auction market input (i.e., a problem instance) takes the form

$$\boldsymbol{x}^{\texttt{AM}} = (n, m, (v_i, b_i, (t_{i,j}, \gamma_{i,j})_{j\in[m]})_{i\in[n]}).$$

We refer to the kth sampled instance by adding a superscript (k) to each component. For the kth sampled instance, the number of agents is $n^{(k)} = 5$ and the number of auctions is $m^{(k)} = 20$. Agent i's per-click value was independently chosen to be $v_i^{(k)} \sim \mathcal{U}(10, 50)$ and budget to be $b_i^{(k)} \sim \mathcal{U}(75, 150)$, where $\mathcal{U}$ denotes the uniform distribution. For each agent $i \in [n^{(k)}]$ and impression $j \in [m^{(k)}]$, we set all agents to target all auctions ($t_{i,j}^{(k)} = 1$) and the CTR for each auction item was drawn from a Beta distribution $\gamma_{i,j}^{(k)} \sim \beta(3, 3, 0)$. We generated a set of 100 base problem instances i.i.d. in this manner.

For each base instance, we generated a set of 5 counterfactual instances, resulting in 600 total instances. Each counterfactual instance was identical to a base instance except that all agents' budgets were scaled by a multiplicative factor, which we refer to as the *counterfactual budget multiplier*. We generated such instances for counterfactual budget multipliers $\{0.5, 1.5, 2.0, 2.5, 3.0\}$ on top of the original instances that correspond to multiplier 1.0.

1.6.2 Ground Truth Mechanism

As described in Section 1.3.1, we consider a direct-revelation mechanism in which agents submit their bids, budgets, and targeting criteria, and each agent's proxy bidder attempts to find bids to maximize the corresponding agent's utility. Each agent i's proxy bidder chooses a pacing multiplier λ_i that scales all of i's per-impression bids by

that multiplicative factor. Each impression is allocated to the highest per-impression bidder, with ties broken randomly. The resulting payments for each impression are determined by a second-price auction.

We first introduce some notation to more precisely define agent i's utility for a profile of pacing multipliers $\boldsymbol{\lambda}$. Let $v_{i,j}^{\texttt{imp}} = \gamma_{i,j} v_i t_{i,j}$ be agent i's expected value for receiving an impression for auction j (which is computed as i's probability of receiving a click $\gamma_{i,j}$ times its value for receiving a click from a targeted user v_i times its targeting indicator $t_{i,j}$). Given agent i's pacing multiplier λ_i, let $d_{i,j}^{\texttt{imp}}(\lambda_i) = v_{i,j}^{\texttt{imp}} \lambda_i$ be i's per-auction bid. Let $H_j(\boldsymbol{\lambda})$ be the set of agents with the highest per-auction bids for auction j:

$$H_j(\boldsymbol{\lambda}) = \underset{i \in [n]}{\text{argmafx}} \{d_{i,j}^{\texttt{imp}}(\lambda_i)\}.$$

Let $w_{i,j}(\boldsymbol{\lambda})$ be the probability that agent i wins auction j, where ties amongst highest bidders are broken randomly:

$$w_{i,j}(\boldsymbol{\lambda}) = \begin{cases} 1/|H_j(\boldsymbol{\lambda})| & \text{if } i \in H_j(\boldsymbol{\lambda}) \\ 0 & \text{otherwise} \end{cases}.$$

Let $p_{i,j}^{\texttt{imp}}(\boldsymbol{\lambda}_{-i})$ be agent i's expected price per impression, conditional on winning auction j. This price is determined by a second-price auction, so the winner's expected price per impression is equal to the per-auction bid of the highest other bidder:

$$p_{i,j}^{\texttt{imp}}(\boldsymbol{\lambda}_{-i}) = \max_{i' \in [n] \setminus i} \{d_{i',j}^{\texttt{imp}}(\lambda_{i'})\}.$$

Agent i's expected payment $C_i(\boldsymbol{\lambda})$ is the sum of its expected costs per impression:

$$C_i(\boldsymbol{\lambda}) = \sum_{j=1}^{m} w_{i,j}(\boldsymbol{\lambda}) p_{i,j}^{\texttt{imp}}(\lambda_i).$$

Similarly, agent i's expected value $V_i(\boldsymbol{\lambda})$ is the sum of its expected values per impression:

$$V_i(\boldsymbol{\lambda}) = \sum_{j=1}^{m} w_{i,j}(\boldsymbol{\lambda}) v_{i,j}^{\texttt{imp}}.$$

Agent i's expected utility $U_i(\boldsymbol{\lambda})$ is quasi-linear provided that its expected spend does

not exceed its budget, and otherwise its utility is arbitrarily low:

$$U_i(\boldsymbol{\lambda}) = \begin{cases} V_i(\boldsymbol{\lambda}) - C_i(\boldsymbol{\lambda}) & \text{if } C_i(\boldsymbol{\lambda}) \leq b_i \\ -\infty & \text{otherwise} \end{cases}.$$

For auction market $\boldsymbol{x}^{\texttt{AM}}$, the mechanism tries to find multipliers $\boldsymbol{\lambda} = (\lambda_i, \boldsymbol{\lambda}_{-i})$ such that each agent is best-responding to other-agent multipliers:

$$\forall i \in [n], \lambda_i', \in \mathbb{R}_{\geq 0}, \quad U_i(\boldsymbol{\lambda}) \geq U_i((\lambda_i', \boldsymbol{\lambda}_{-i})).$$

While the objective is to output equilibrium pacing multipliers, there are two issues to address: (1) there may be multiple equilibria, so we need to specify which will be output, and (2) the algorithm may only output an approximate rather than exact equilibrium. Both of these issues are addressed in the next subsection.

Computing Approximate Equilibria

Given an auction market input, the ground truth mechanism searches for a profile of equilibrium pacing multipliers $\boldsymbol{\lambda}^*$ using a best-response dynamic. At each iteration, all agents update their pacing multipliers to best-respond to the bids from the previous iteration. The algorithm terminates after the change in multipliers across iterations becomes sufficiently small or if a maximum number of iterations is reached. Agents' utilities and auction outcomes are determined by the output profile of pacing multipliers.

The specifics of the best-response update, stopping condition, and initial pacing multipliers are as follows. Given other-agent multipliers $\boldsymbol{\lambda}_{-i}$, let $\lambda_{i,j}^{\texttt{critical}}(\boldsymbol{\lambda}_{-i})$ be the *critical pacing multiplier* for auction j, in that any higher (or lower) multiplier causes i to win (or lose) auction j:

$$\lambda_{i,j}^{\texttt{critical}}(\boldsymbol{\lambda}_{-i}) = p_{i,j}^{\texttt{imp}}(\boldsymbol{\lambda}_{-i})/v_{i,j}^{\texttt{imp}}.$$

Let $\mathcal{L}_i(\boldsymbol{\lambda}_{-i})$ be the set of critical pacing multipliers for agent i:

$$\mathcal{L}_i(\boldsymbol{\lambda}_{-i}) = \{x \in \mathbb{R}_{\geq 0} : \exists j \in [m] \text{ s.t. } t_{i,j} = 1 \wedge x = \lambda_{i,j}^{\texttt{critical}}\}.$$

Agent i's best response is computed as the highest multiplier that maximizes its ap-

proximate utility:

$$BR_i(\boldsymbol{\lambda}_{-i}) = \sup\{\underset{\lambda_i' \in \mathcal{L}_i(\boldsymbol{\lambda}_{-i})}{\text{argmafx}}\{\widehat{U}_i((\lambda_i', \boldsymbol{\lambda}_{-i}))\}\},$$

where $\widehat{U}_i$ is an approximation of i's utility function that assumes agent i wins all tiebreakers.

The best-response dynamic terminates when either (1) every agent's multiplier is sufficiently close to what it's been in the recent past, or (2) a certain number of iterations have elapsed. Specifically, after computing the best response multiplier λ_i^t for each agent i at iteration t, the algorithm terminates with multipliers $\boldsymbol{\lambda}^t$ if $\delta^t < \kappa_1$ or $t > \kappa_2$, where the convergence threshold is

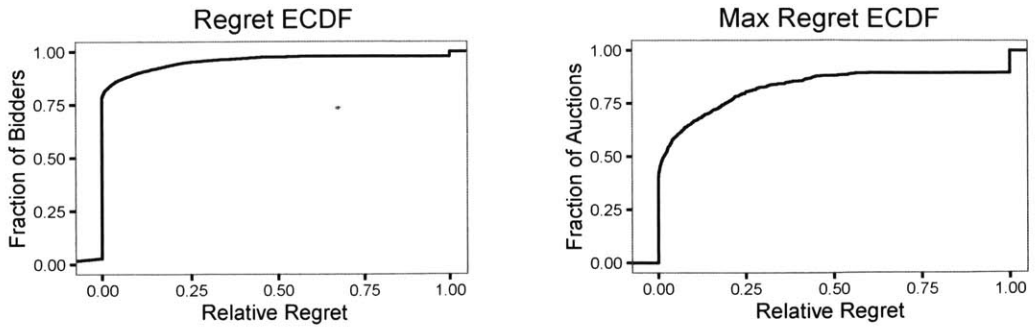
$$\delta^t = \min_{k \in [4]}\{\max_{i \in [n]}\{|\lambda_i^t - \lambda_i^{t-k}|\}\}$$

and $\lambda_i^t = \infty$ for $t < 1$. In our experiments, $\kappa_1 = 0.001$ and $\kappa_2 = 30$.

Initial multipliers to this procedure are themselves determined by a best response dynamic in which, at each iteration, each agent i chooses the largest multiplier such that the sum of its winning bids is under its budget.

Empirical Regret Figure 1-4 shows how well the ground truth mechanism converges to equilibrium for the given problem instances. Figure 1-4a shows the empirical CDF for each agent's regret at the terminating multipliers. Figure 1-4b shows the fraction of auctions which had at least one agent with a given amount of regret (this corresponds to the equilibrium approximation ratio). The jump at relative regret 1.0 is from instances in which the output pacing multipliers result in an agent losing all auctions (for utility 0) but a best response where it wins at least one auction for positive utility.

While an equilibrium profile was output by the ground truth mechanism on almost half of all instances, the failure to always reach exact equilibrium is not detrimental to our experiments. Indeed, equilibria may not always be reached in the real-world mechanism, either, and we aim to predict summary statistics accurately even in cases where an exact equilibrium is not found.



(a) (b)

Figure 1-4: Empirical CDFs for a per-agent regret and b per-auction max regret.

1.6.3 Relevant summary statistics

We consider expected publisher's utility for the ground truth mechanism:

$$g^{\texttt{AM-Ut}}(\boldsymbol{\lambda}) := \sum_{i=1}^{n} C_i(\boldsymbol{\lambda}).$$

1.6.4 Predictors

The features we selected are summary statistics of the auction market. They consist of mean budget, maximum CTR in an auction averaged out across auctions, the average ratio of budget to maximum possible total value to be gained by an agent, and the WM output:

$$\phi^{\texttt{meanBudget}}(\boldsymbol{x}^{\texttt{AM}}) := \frac{1}{n}\sum_{i=1}^{n} b_i \tag{1.3}$$

$$\phi^{\texttt{meanMaxCTR}}(\boldsymbol{x}^{\texttt{AM}}) := \frac{1}{m}\sum_{j=1}^{m} \max_{i\in[n]}\{t_{i,j}\gamma_{i,j}\} \tag{1.4}$$

$$\phi^{\texttt{meanBudgetToMaxImpBidTotal}}(\boldsymbol{x}^{\texttt{AM}}) := \frac{1}{n}\sum_{i=1}^{n} \frac{b_i}{\sum_{j=1}^{m} v_i t_{i,j}\gamma_{i,j}} \tag{1.5}$$

$$\phi^{\texttt{WM}}(\boldsymbol{x}^{\texttt{AM}}) := e^{\texttt{WM-AM}}(g^{\texttt{WM-Ut}}(f^{\texttt{WM}}(h^{\texttt{AM-WM}}(\boldsymbol{x}^{\texttt{AM}})))) \tag{1.6}$$

Each baseline predictor is a linear regression model as given in (1.2), with feature sets defined in Table 1.1. We compare these baselines to the raw WM output $\phi^{\texttt{WM}}(\boldsymbol{x}^{\texttt{AM}})$, as well as models that add $\phi^{\texttt{WM}}$ to each of the baseline feature sets.

Table 1.1: Predictor names and model details

Predictor Name	Feature Set
ML1	$\varnothing$
ML2	$\{\phi^{\texttt{meanBudget}}\}$
ML3	$\{\phi^{\texttt{meanMaxCTR}}\}$
ML4	$\{\phi^{\texttt{meanBudgetToMaxImpBidTotal}}\}$
ML5	$\{\phi^{\texttt{meanBudget}}, \phi^{\texttt{meanMaxCTR}}, \phi^{\texttt{meanBudgetToMaxImpBidTotal}}\}$
WM_MLx	MLx $\cup \{\phi^{\texttt{WM}}\}$

1.7 Experimental Results

In this section we report the results of the computational study with the setup provided by the previous section. Every paragraph below roughly corresponds to the questions posed earlier.

Performance of Raw WM versus Ground Truth To understand the accuracy of the WM model by itself we first plot its output on 20 sample instances. Figure 1-5 compares publisher's utility between the ground truth and WM model. Although WM generally predicts higher publisher's utilities than ground-truth, the shape of both curves is similar as the budget multiplier varies. This means that even though the estimator is slightly biased, it can be easily corrected by rescaling and that the WM is correctly capturing the market dynamics at different competition levels by estimating correct budget elasticities. These results suggest that even a simple ML model using only WM as a feature might accurately predict publisher's utility.

To understand why WM overestimates publisher's utility, consider an instance with a single agent. Because the agent bids in second-price auctions with no competition, it pays nothing in the ground truth mechanism. But in the WM model, the agent is a price taker, resulting in positive per-auction prices. We suspect that, as the number of bidders per auction increases, the publisher's utility output by WM model will converge to that of the ground truth mechanism. Such analysis will be the subject of further studies.

Within-sample versus out-of-sample predictions With the previous evidence as motivation, we now incorporate features with and without the WM output, and predict publisher's utility. Figure 1-6 shows MAPE for all predictors when using 10-fold cross validation (2, 5, and 20 folds had similar results). All predictors were trained on a

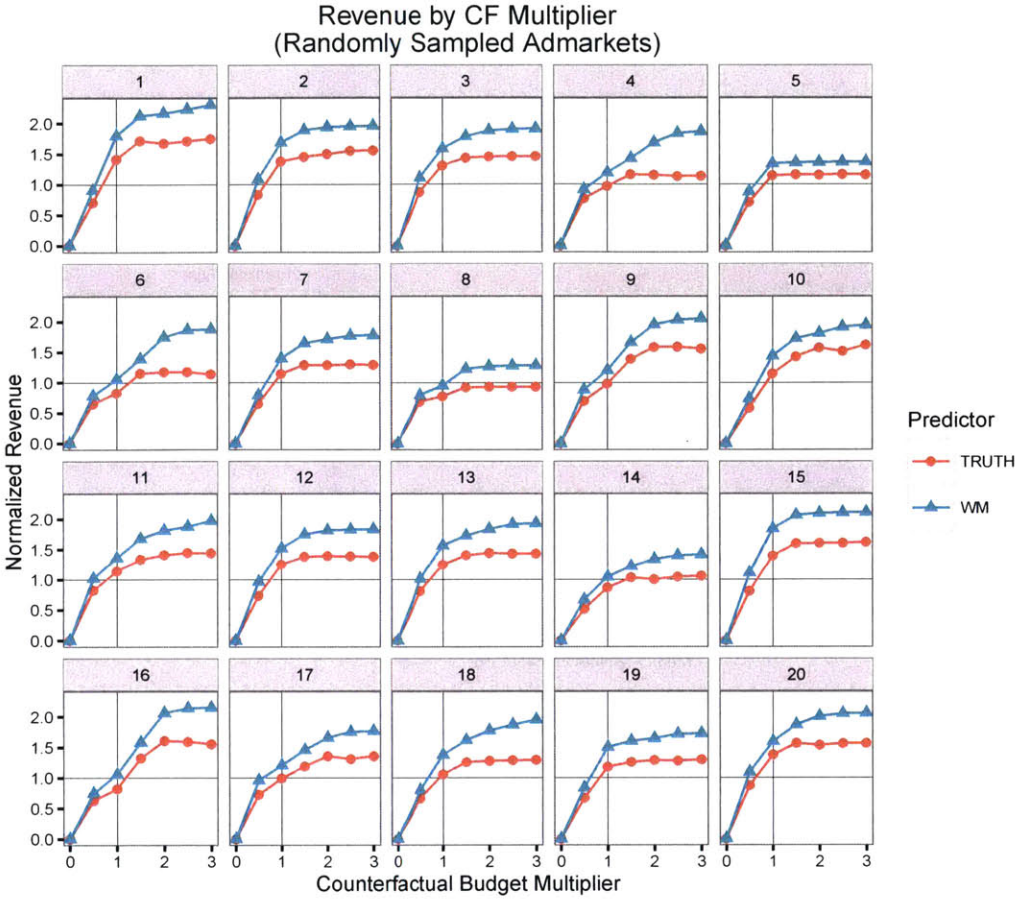


Figure 1-5: Ground-truth versus WM publisher's utility for a random sample of auction market instances for varying budget multipliers. Utility is normalized by the mean ground-truth publisher's utility across all instances with budget multiplier 1.0.

sample of the base instances (i.e., instances with counterfactual budget multiplier 1.0). In the upper plot, test instances were also drawn from the same distribution (i.e., had counterfactual multiplier 1.0). In the lower plot, test instances all had counterfactual multiplier 2.0. Other counterfactual multipliers had qualitatively similar results and are shown in Appendix 1.A.

The main takeaways for the top plot (within-sample predictions) are: (1) WM alone gives the worst predictions because of the bias shown in Figure 1-5, and the fact that it is the only predictor that uses no training data. (2) For non-WM ML predictors, ML4 is the best single-feature predictor, and combining all non-WM features performed better than any single-feature predictor. (3) WM+ML models gave the best results. All WM+ML models perform similarly well. In particular, the simple model WM_ML1

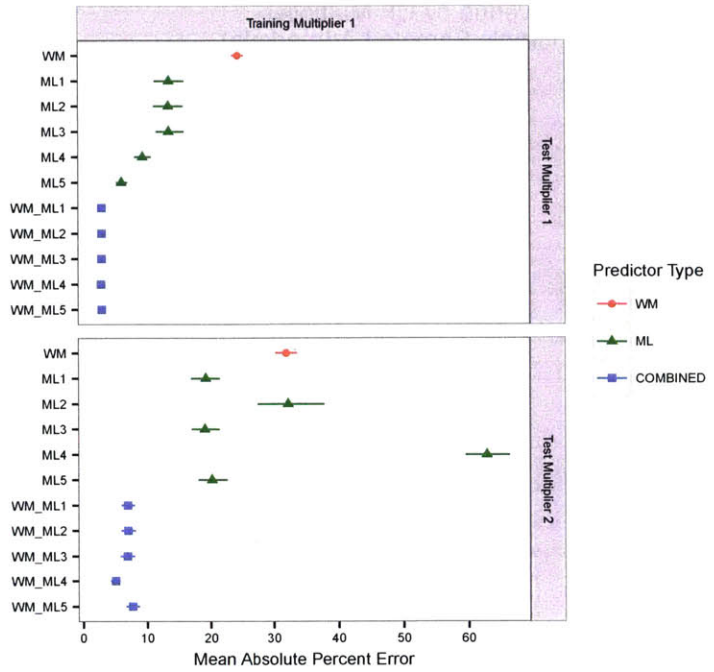


Figure 1-6: Prediction accuracy of WM and/or additional features

returns $\hat{\beta}_0 + \hat{\beta}_1 \phi^{\texttt{WM}}(\boldsymbol{x}^{\texttt{AM}})$ which is simply scaling the WM model. This model can be used to estimate by how much WM overestimates the pacing equilibrium on average.

The main takeaways for the bottom plot (out-of-sample predictions) are: (1) WM alone did not degrade by much when the test distribution differed from the training distribution. (2) The combined predictors continued to be the most accurate. (3) ML4, while the best single-feature ML model for within-sample predictions, performs poorly out of sample. Yet, when combined with WM (to get WM_ML4), it is better than any other predictor. This suggests that the ML4 feature captures a different aspect of the auction market compared to WM.

Base versus counterfactual training data Finally, Figure 1-7 shows the average predicted publisher's utility across all instances for a subset of predictors from Figure 1-6. We restrict to a single ML and WM+ML predictor for readability. We focus on feature set 5 — the combined features — because it had good performance with and without the WM feature and was robust to out-of-sample predictions. The left, center and right plots show predictors trained only on base ($CF = \{1\}$), on base plus an additional counterfactual budget multiplier ($CF = \{1, 2\}$) and on all counterfactual instances, respectively. The main takeaways are as follows. (1) While we saw from

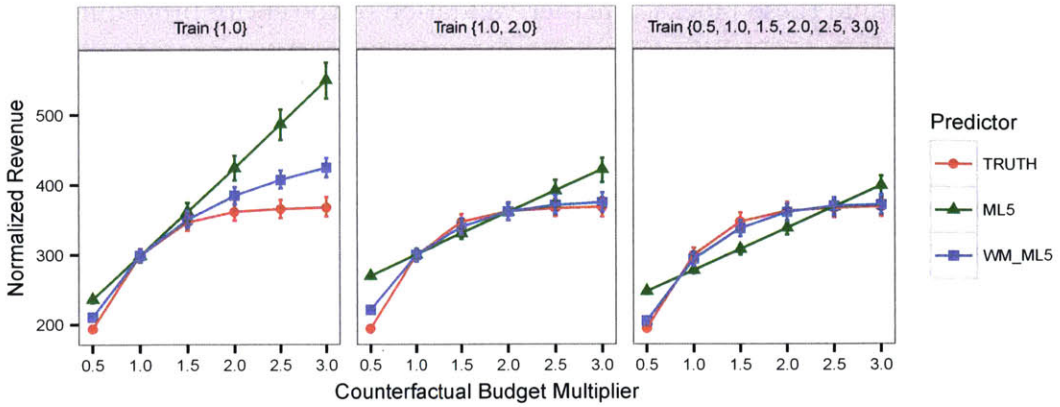


Figure 1-7: Average predicted publisher's utility for different predictors by counterfactual budget multiplier. The plots use different counterfactual training data.

Fig. 1-6 that combining WM and ML was most robust against out-of-sample predictions, predictions get worse as the test instances get further away (in terms of a higher counterfactual budget multiplier) from training instances. (2) Accuracy is improved when WM+ML models are combined with counterfactual training instances. (3) Only one additional training counterfactual budget multiplier was enough to almost match the results for training with all multipliers. This suggests that even a small amount of counterfactual experimentation is sufficient to accurately predict the entire counterfactual publisher's utility curve.

1.8 Conclusion

We have introduced a method for predicting relevant summary statistics of complex online auction markets. Our approach combines techniques of market abstraction, machine learning, and counterfactual analysis. We ran computational experiments establishing the internal consistency of our method and provided evidence that our method accurately predicts relevant market outcome summary statistics, even when predicting on counterfactual inputs different from those previously observed.

There are numerous directions for extending this work. First, there is room for further empirical validation of the WSBM approach, which would require evaluating on larger problem instances than were tractable for our ground truth implementation. Additionally, our current process for generating WSBMs from data assumes that blocks of agents and users are given as input; it is worth considering clustering methods to

dynamically determine these blocks. Further validation of the entire procedure on real-world data would provide additional supporting evidence for the approach and a better understanding of its limitations.

While we found the Walrasian market abstraction to be effective at predicting publisher's utility, we expect that different game abstractions will be more or less effective depending on the summary statistic of interest and structure of the market domain. There is room for further empirical results to understand which abstractions are effective on which types of instances.

There is also room for theoretical results that mirror recent work on game abstractions. That is, given a game and corresponding outcome summary statistic, can we provide bounds on the error of using a given game abstraction for predicting outcome summaries? Is there any structure arising in real-world markets that results in tighter bounds?

Another possible direction for future work combines research in budgeted learning and automated game abstraction. We can think of our market outcome summary statistic problem as being one of function approximation, in which the exact mechanism gives exact output but is computationally expensive to compute, whereas features based on solving different game abstractions are less accurate but also less expensive to extract. How do we decide on a set of game abstractions that optimize for this trade-off between accuracy and computational efficiency?

There also seems to be a connection between automated abstraction in games and the machine learning literature on representation learning. In a sense, we are proposing a type of basis function for representation learning that is evaluated by solving a game. Can we formulate our abstraction approach into existing methods in representation learning? Doing so may enable us to take advantage of developments in both fields.

In practice, we care about accurate predictions of market outcomes because we use them to make decisions. It may be worth investigating case studies for particular decision-making problems in which these predictions are needed. Knowledge of how predictions will be used would allow us to choose abstractions that more directly attempt to optimize the true objective under consideration.

Finally, in real-world markets, agents update their actions over time based on, among other things, their past success on a given platform. It would be interesting to extend the approach to consider longer-term predictions of publisher's utility, which would require accounting for dynamics over time. The question then becomes: what abstractions appropriately capture summary statistics that simultaneously (1) accurately predict

short-term counterfactual publisher's utility (as done in this work) and (2) accurately predict how mechanism inputs will change over time?

Bibliography

Kareem Amin, Michael Kearns, Peter Key, and Anton Schwaighofer. Budget Optimization for Sponsored Search: Censored Learning in MDPs. 2012.

Kenneth J. Arrow, H.D. Block, and Leonid Hurwicz. On the stability of the competitive equilibrium. *Econometrica*, 29:82–109, 1959.

Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149, 2015.

Richard Cole and Lisa Fleischer. Fast-converging tatonnement algorithms for one-time and ongoing market problems. In *ACM Symposium on Theory of Computing*, pages 315–324, 2008.

Ran Duan and Kurt Mehlhorn. A combinatorial polynomial algorithm for the linear arrow-debreu market. In FedorV. Fomin, Rusins Freivalds, Marta Kwiatkowska, and David Peleg, editors, *Automata, Languages, and Programming*, volume 7965 of *Lecture Notes in Computer Science*, pages 425–436. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-39205-4.

M. Florig. Equilibrium correspondence of linear exchange economies. *Journal of Optimization Theory and Applications*, 120(1):97–109, 2004. ISSN 0022-3239.

David Gale. The linear exchange model. *Journal of Mathematical Economics*, 3(2): 205–209, July 1976.

Sam Ganzfried and Tuomas Sandholm. Potential-aware imperfect-recall abstraction with earth mover's distance in imperfect-information games. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

K. Jain. A polynomial time algorithm for computing an arrow-debreu market equilibrium for linear utilities. In *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, pages 286–294, Oct 2004.

Christian Kroer and Tuomas Sandholm. Extensive-form game abstraction with bounds. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 621–638. ACM, 2014.

Andreu Mas-Colell, Michael Whinston, and Jerry Green. *Microeconomic Theory*. Oxford University Press, 1995.

Jean-Francois Mertens. The limit-price mechanism. CORE Discussion Papers 1996050, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), October 1996.

Whitney K. Newey and Daniel McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111 – 2245. Elsevier, 1994.

Tuomas Sandholm. Abstraction for solving large incomplete-information games. In *AAAI Conference on Artificial Intelligence (AAAI). Senior Member Track*, 2015.

Tuomas Sandholm and Satinder Singh. Lossy stochastic game abstraction with bounds. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 880–897. ACM, 2012.

Michael P. Wellman. Methods for Empirical Game-Theoretic Analysis (Extended Abstract). pages 1552–1555, Boston, July 2006.

Michael P Wellman, Daniel M Reeves, Kevin M Lochner, and Yevgeniy Vorobeychik. Price prediction in a trading agent competition. *Journal of Artificial Intelligence Research*, pages 19–36, 2004.

Yinyu Ye. A path to the arrow-debreu competitive market equilibrium. *Mathematical Programming*, 111(1-2):315–348, 2008. ISSN 0025-5610.

Appendix 1.A Additional Results

Figure 1-8, an extension of Figure 1-6, shows prediction accuracy for additional training and test sets. Each column (or row) corresponds to a training (or test) set with in which only multipliers of a particular value were included. The plots on the diagonal have training and test data drawn from the same distribution. The lower-left and upper-right plots show train and test data drawn from drastically different distributions. These results illustrate the robustness of WM+ML models, and particularly WM+ML1, which only adds a scaling multiplier and offset to the raw WM output.

The next set of plots illustrate the prediction accuracy of `WM_SCALED`, whose publisher utility prediction at multiplier m is computed as follows: $\texttt{WM\_SCALED}(m) = \texttt{WM}(m)\frac{\texttt{TRUTH(1)}}{\texttt{WM(1)}}$. Note that this differs from other predictors in that it is given ground truth publisher utility at multiplier 1.0. However, it does not use any training data to make its predictions. Figures 1-9, 1-10, and 1-11 are extensions of Figs. 1-5, 1-6, and 1-7, modified to include results for `WM_SCALED`.

Appendix 1.B CES approximation

Walrasian markets with linear utility have two issues. First, there are mulitple equilibrium price vectors (though they form a convex set, and publisher utility in all equilibria are the same) Gale (1976); Mertens (1996); Florig (2004). Second, linear utility poses a challenge for computation due to its infinite elasticity of substitution, and while polynomial-time algorithms exist for computing equilibria in linear utility Walrasian markets Jain (2004); Ye (2008); Duan and Mehlhorn (2013), they are not scalable for inputs of our size. Thus, for ease of computation, we approximate linear utilities with CES utilities using a sufficiently high elasticity of substitution parameter σ:

$$u_i^{\sigma}(\boldsymbol{a}_i) = \left(\sum_{j=1}^{m^{\texttt{WM}}} a_{i,j}^{(\sigma-1)/\sigma} t_{i,j} \gamma_{i,j} \right)^{\sigma/(\sigma-1)}.$$

Using CES utilities gives us uniqueness of equilibrium Arrow et al. (1959), and also finite elasticity of substitution, which allows us to use the fast and parallelizable algorithm of Cole and Fleischer (2008).

First, we define the transformation that turns an auction market into a CES-utility

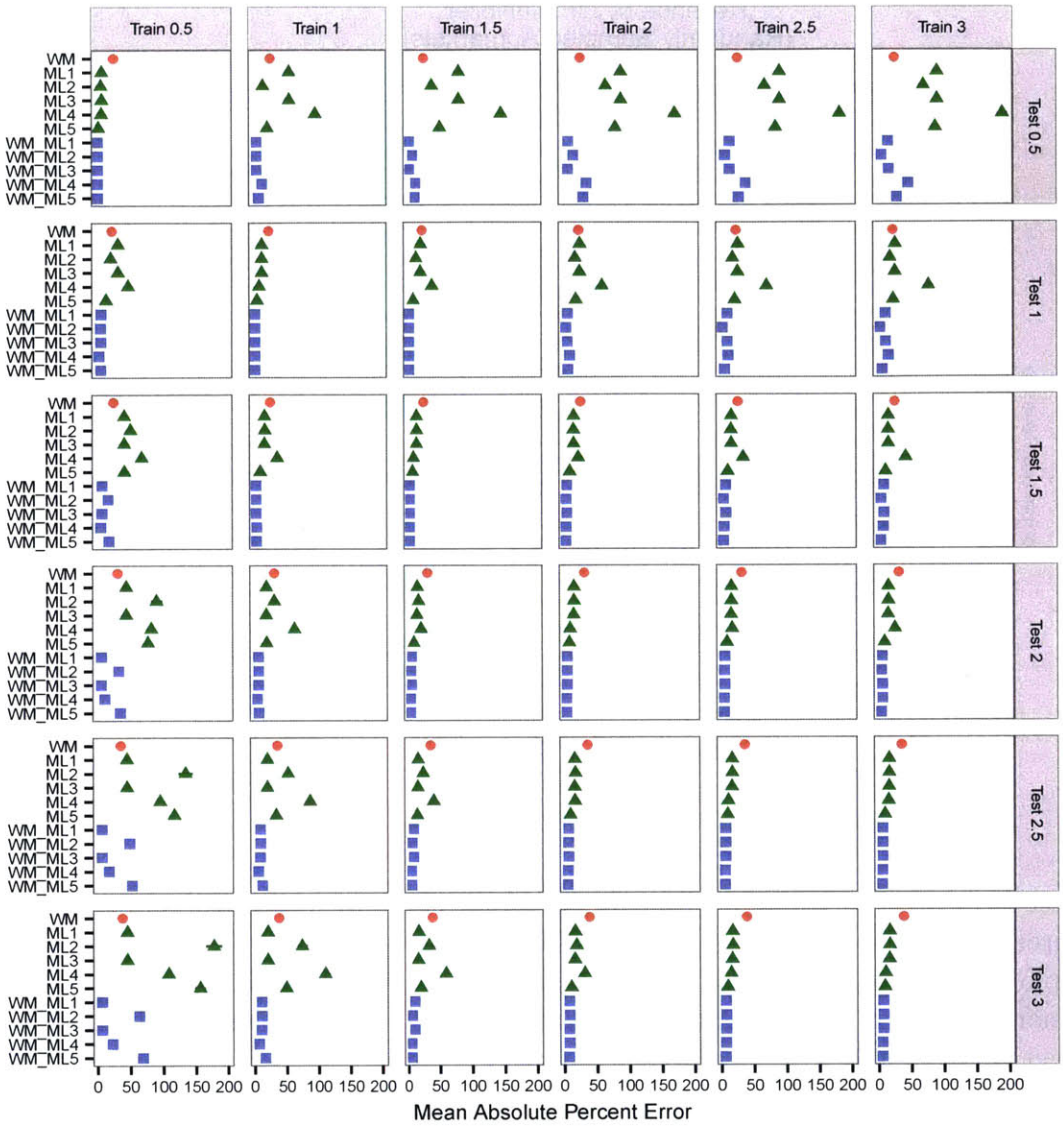


Figure 1-8: Prediction accuracy of WM and/or additional features. These plots are analogous to those in Figure 1-6, but for all combinations of training and test multipliers. We observe that the WM+ML models are robust across a variety of train/test input distributions.

Walrasian market. The procedure is nearly identical to the one in Definition 5, except we endow the buyers with a CES utility function instead of a linear one:

Definition 7 (Walrasian Market Transformation (CES Utilities))**.** *Let* $f^{\texttt{WM-Linear-CES}} : X^{\texttt{AM}} \times \mathbb{R}_+ \to X^{\texttt{WM}}$ *be a function that takes as input an auction market* $\boldsymbol{x}^{\texttt{AM}} = (n^{\texttt{AM}}, m^{\texttt{AM}}, \boldsymbol{v}^{\texttt{AM}}, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{s}, \boldsymbol{\gamma})$ *together with a elasticity-of-substitution parameter* σ*, and outputs a Walrasian market*

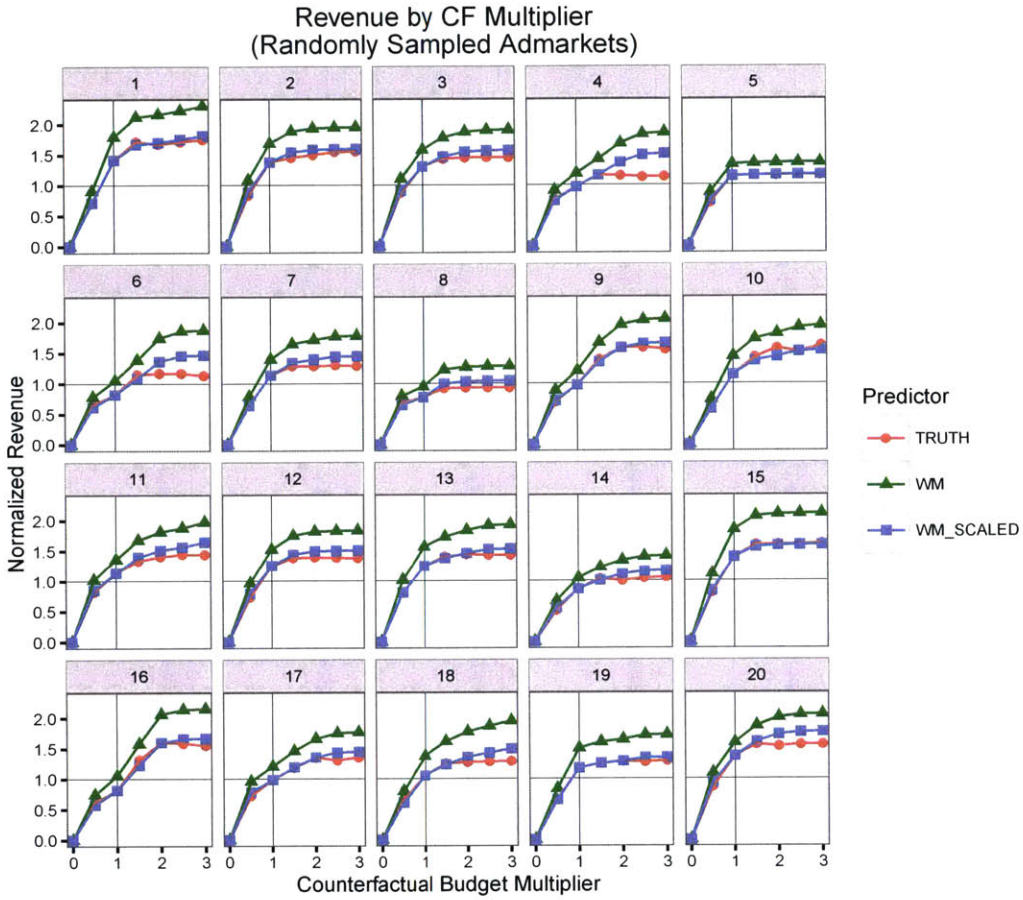


Figure 1-9: Ground truth publisher utility versus WM publisher utility for a random sample of auction market instances for varying budget multipliers. publisher utilities are normalized by the mean ground truth publisher utility across all instances with budget multiplier 1.0.

$\boldsymbol{x}^{WM} = (n^{WM}, m^{WM}, \boldsymbol{e}, \boldsymbol{u})$ *such that* $n^{WM} = n^{AM} + 1$, $m^{WM} = m^{AM} + 1$, $\boldsymbol{e} = (e_{i,j})_{i \in [n^{WM}], j \in [m^{WM}]}$, *where* $e_{i,j} = \{b_i^{AM}$ *if* $i \in [n^{AM}] \wedge j = m^{AM} + 1; 1$ *if* $i = n^{AM} + 1 \wedge j \in m^{AM}; 0$ *otherwise*$\}$, *and* $\boldsymbol{u} = (u_i)_{i \in [n^{WM}]}$, *where* $u_i(\boldsymbol{a}_i) = \left(\sum_{j=1}^{m^{WM}} a_{i,j}^{(\sigma-1)/\sigma} t_{i,j} \gamma_{i,j} \right)^{\sigma/(\sigma-1)}$.

As $\sigma \to \infty$, the CES utility function

$$\left(\sum_{j=1}^{m^{WM}} a_{i,j}^{(\sigma-1)/\sigma} t_{i,j} \gamma_{i,j} \right)^{\sigma/(\sigma-1)}$$

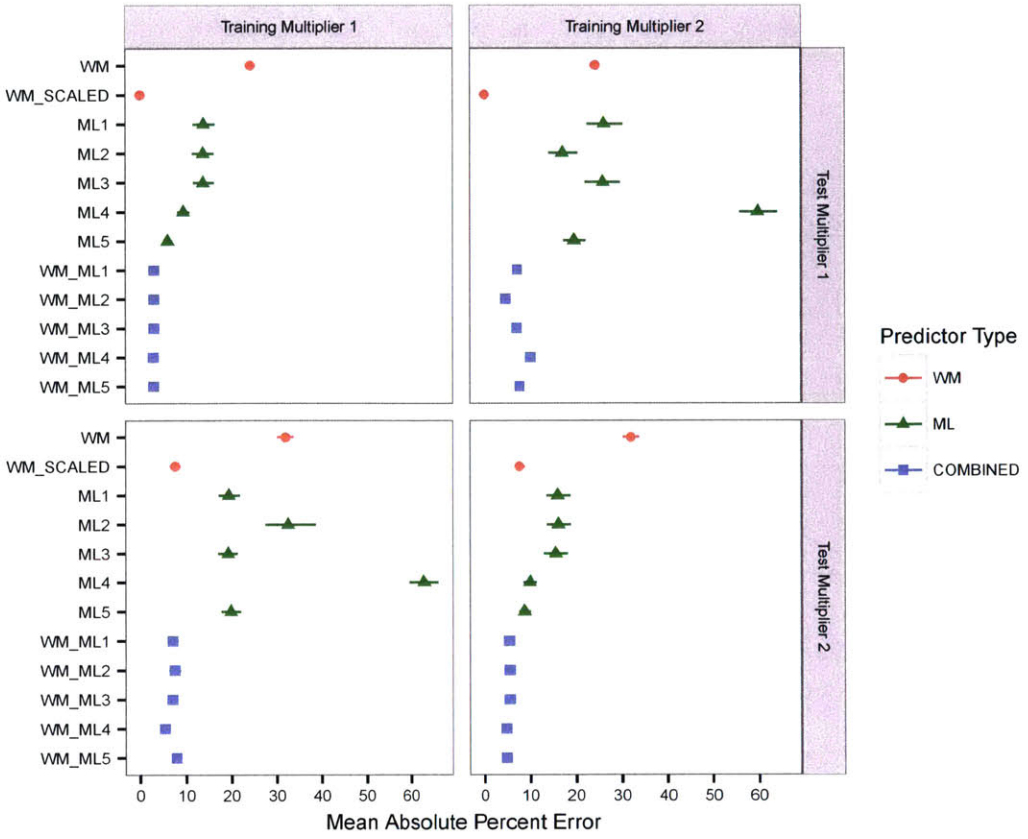


Figure 1-10: Prediction accuracy of WM and/or additional features

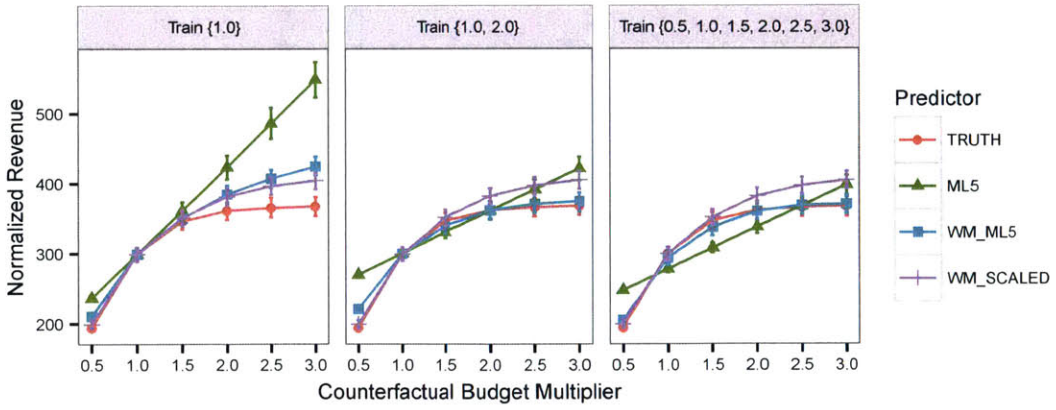


Figure 1-11: Average predicted publisher utility for different predictors by counterfactual budget multiplier. The plots use different counterfactual training data.

converges to the linear utility function

$$v_i \sum_{j=1}^{m^{\text{WM}}} a_{i,j} t_{i,j} \gamma_{i,j}.$$

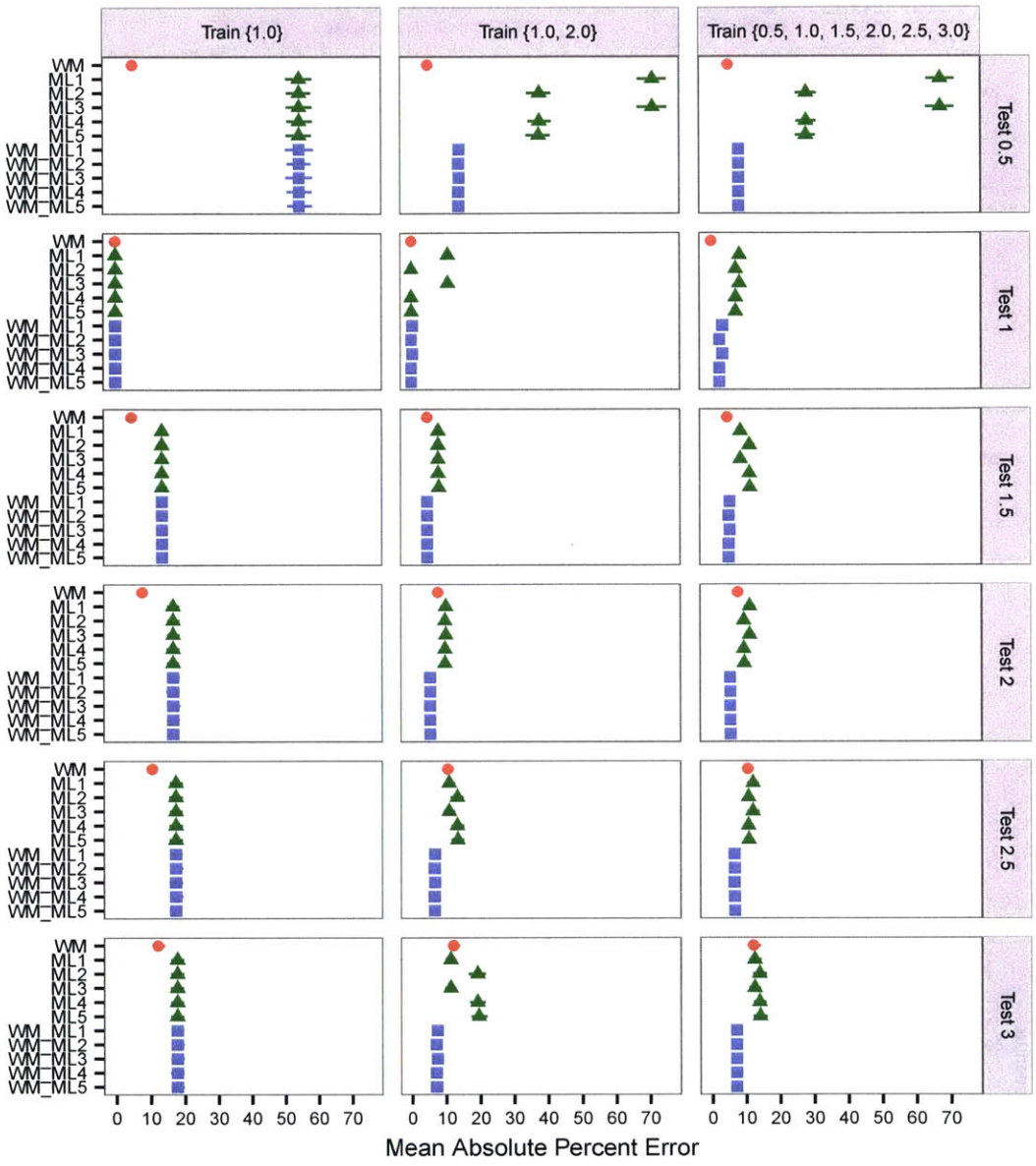


Figure 1-12

Given an auction market $\boldsymbol{x}^{\mathtt{AM}}$, define $\boldsymbol{x}_{\infty}^{\mathtt{WM}}$ be a Walrasian market with linear utilities constructed from $\boldsymbol{x}^{\mathtt{AM}}$ via Definition 5, and define $\boldsymbol{x}_{\sigma}^{\mathtt{WM}}$ be a Walrasian market with CES utilities constructed from $\boldsymbol{x}^{\mathtt{AM}}$ via Definition 7.

We would like to show that, as $\sigma \to \infty$, the Walrasian equilibrium of $\boldsymbol{x}_{\sigma}^{\mathtt{WM}}$ is upper hemicontinuous in σ, and in particular converges to a Walrasian equilibrium of the linear Walrasian market $\boldsymbol{x}_{\infty}^{\mathtt{WM}}$.

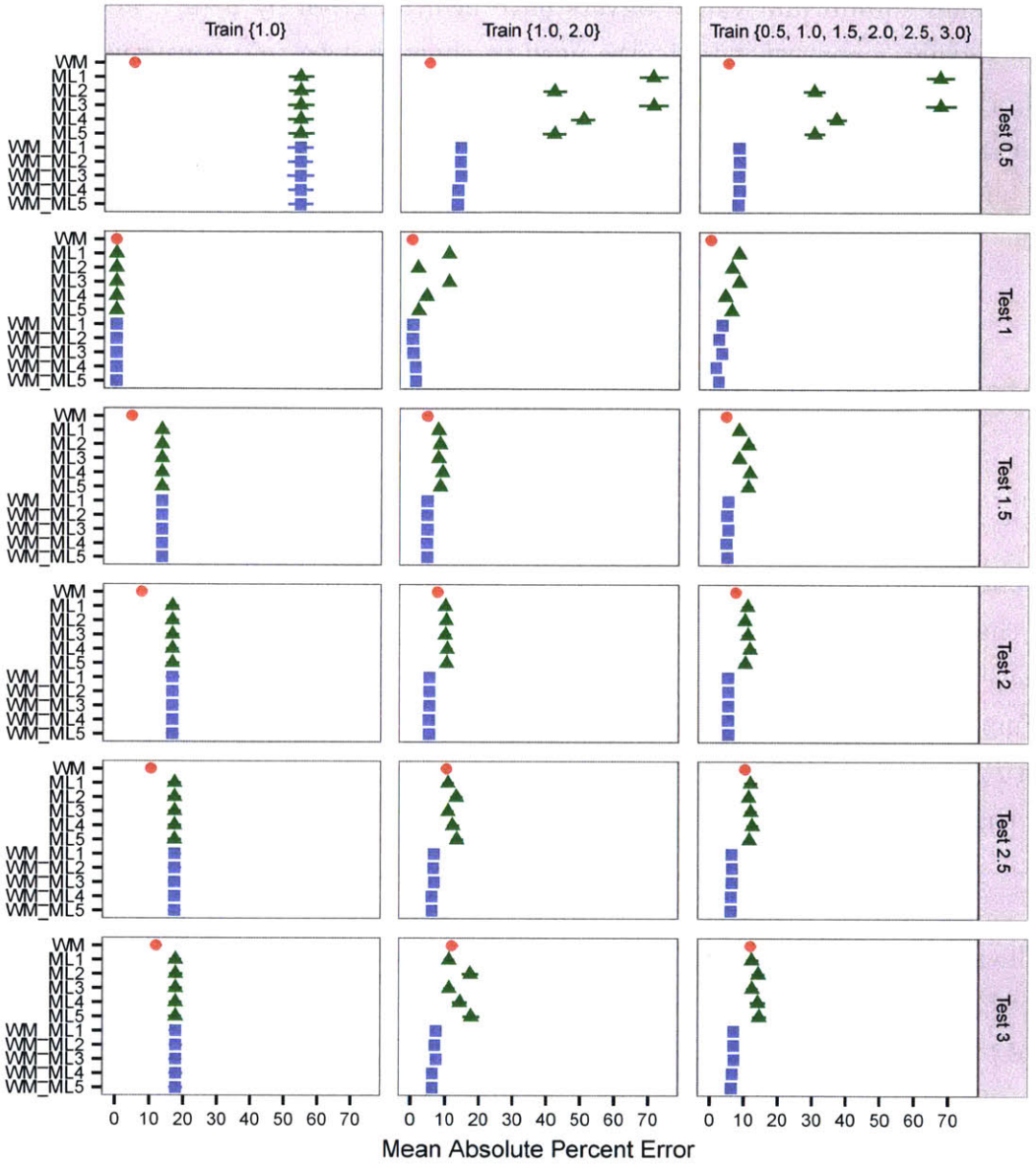


Figure 1-13

First, let us define $\rho = (\sigma - 1)/\sigma$, so that the CES utility function can be written

$$u_i^{\rho}(\boldsymbol{a}_i) = \left(\sum_{j=1}^{m^{\text{WM}}} a_{i,j}^{\rho} t_{i,j} \gamma_{i,j} \right)^{1/\rho} . \tag{1.7}$$

To prove upper hemicontinuity of Walrasian equilibrium, we take a sequence of prices p^t and demands a^t converging towards some price p^* and a^*, where (p^t, x^t) consist of

a Walrasian equilibrium where buyers have utilities given by $u_i^\rho(\boldsymbol{a}_i)$, and σ converges to ∞, so that ρ converges to 1. We need to show that (p^*, a^*) constitutes a Walrsian equilibrium when $\rho = 1$.

First, we show that a_i^* constitutes an optimal and feasible consumption bundle for each buyer i at $\rho = 1$:

$$a_i^* \in \arg\max_y u_i^1(y) \text{ s.t. } \sum_j p_j a_{i,j}^* \leq 1.$$

Assume for the sake of contradiction that there exists some a_i' such that

$$u_1^1(\boldsymbol{a}_i') - u_b^1(\boldsymbol{a}_i^*) =: \epsilon > 0, \quad \sum_j p_j a_{i,j}' \leq 1$$

By continuity of $u_i^\rho(\boldsymbol{a}_i)$ with respect to both ρ and $\boldsymbol{a}_i$, together with $a_i^t \to x^*$, we have that there is some t, δ such that for all $t > T_1$ and for all $\boldsymbol{a}_i''$ such that $||\boldsymbol{a}_i'' - \boldsymbol{a}_i'||_\infty \leq \delta$, we have $u_i^{\rho_i}(\boldsymbol{a}_i'') > u_i^{\rho_i}(\boldsymbol{a}_i^t)$.

Furthermore, as the $\sum_j p_j a_{i,j}$ is continuous with respect to p, we have that there is some T_2 such that for $t > T_2$, we have p^t close enough to p^* such that there is some a^{**} such that

$$||a_i^{**} - a_i'||_\infty \leq \delta, \quad \sum_j p_j^t x_{ij}^{**} \leq 1.$$

Thus, take any $t > \max(T_1, T_2)$, take any i, take the corresponding a_i^{**}, and we have that a_i^{**} both strictly dominates a_i^t in terms of payoffs, and also is feasible at prices p^t. That, is we have a contradiction, and it follows that a^* must produce weakly more utility than any other feasible consumption bundle. Feasibility follows from the continuity of $\sum_j p_j a_{ij}$ wrt p, a and the fact that $\sum_j p_j^i a_{ij}^t \leq 1$ for all t. Thus, a^* is both optimal and feasible.

Finally, we need to check market clearing of a^* and p^*. This follows from the continuity of $\sum_i a_{ij}^t$ and the fact that $\sum_j p_j^t a_{ij}^t = 1 \ \forall j$ for all t.

Appendix 1.C Computing Walrasian Equilibrium

Section 1.4.1 notes that we solve for WME via $f^{\texttt{WM}}$, which is the distributed tatonnement algorithm of Cole and Fleischer (2008). The algorithm initializes a set of positive prices $\boldsymbol{p}^{\texttt{init}}$, one for each (non-money) good, computes excess demand x_j for

each good j, and then updates prices with $p_j \leftarrow p_j(1 + \lambda \min\{1, (x_j - s_j)/s_j\})$. Here, λ is a parameter that affects the magnitude of price adjustments (chosen as $\frac{1}{2\sigma-1}$), and s_j is the total supply of good j, so that

$$z_j := (x_j - s_j)/s_j \tag{1.8}$$

is a normalized measure of excess demand. The price of money is always fixed at 1. This tatonnement process is continued until a tolerance threshold of ϵ is achieved or until κ iterations were completed.

Using the convergence bound obtained in Cole and Fleischer (2008), we have that in our case, the price p^t after t price updates is guaranteed to satisfy

$$\max_j \frac{|p_j^t - p_j^*|}{p_j^*} < \delta \quad \text{for} \quad t \sim O\left(\frac{1}{\lambda} \log\left(\frac{\eta(p^0)}{\delta}\right)\right) \tag{1.9}$$

where p_j^* is the equilibrium price of good j, and p_j^t is the computed price of good j after t rounds of price updates.

It should be noted that there are two places where the size of the sampled market impacts this algorithm:

1. Each iteration of the algorithm requires going through each (buyer, good pair) and computing the demand, and thus requires $O(NM)$ operations where N and M are the numbers of buyers and goods.

2. The $\eta(p^0)$ term (defined in Appendix 1.G), which is a measure of how far away the most extreme price p_i^0 is from the true equilibrium prices. Naturally, as we sample more and more goods, the worst initialized p_i^0 will get worse, though it's unclear the rate at which this happens.

Further investigation of how the number iterations required for convergence increases with sample size is warranted, though in our experience this was less than linear. As a result, we intuitively expect the total complexity of the algorithm to be around $\leq O(NM(N + M)\log(1/\delta))$ with high probability, which compares favorably with the $O((N + M)^4 \log(1/\delta))$ found in e.g. Ye (2008) (to be clear, the comparison is not apples-to-apples, as Ye (2008) computes the exact equilibrium in the linear utilities case whereas we compute an approximation using CES utilities).

The tatonnement algorithm is embarrassingly parallelizable:

1. A central processor disseminates a set of prices to a number of other processors, each of which:
   (a) Computes the demand vectors for a subset of the users.
   (b) Computes a total demand for each good based by summing the demand for that good across subset of users.
   (c) Returns this partially aggregated demand to the central processor.
2. The central processor then sums the partially aggregated demands for each of the peripheral processors into a single total aggregate demand for each good.
3. And then performs the price tatonnement price update.

Appendix 1.D WSBM Pseudocode

Algorithms 2-6 (found at the end of this document) describe the WSBM procedure. We take a high-dimensional Walrasian market as input and transform it into a WSBM which has a lower dimension but similar structural properties. We then proceed similarly to the steps taken for the WM abstraction (i.e., we solve for equilibrium and output the relevant summary statistics for the higher-dimensional problem). The first step is to choose partitions over the sets of bidders and queries. Let c be the number of components in the chosen partition of set $[n]$, and let $h^{buyer}(i) \in [c]$ indicate the component that bidder i falls into. Similarly, let d be the number of components resulting from the chosen partition of set $[m]$, and let $h^{good}(j) \in [d]$ indicate the component that query j falls into.

We consider partitions equal to the market segments used for computing summary statistics. Given these partitions, computing stochastic block model parameters involves computing empirical distributions from the auction market. We let ξ^{buyer} be the empirical distribution of buyer-types and, similarly with ξ^{good}. We let g_l^{budget} be the empirical distribution of budgets for each buyer-type l, and $g_{l,k}^{value}$ be the empirical distribution of $v_i t_{i,j} \gamma_{i,j}$ for each (buyer-type, good-type) pair (l, k).

We use the constructed WSBM to stand-in for the full Walrasian market. To compute equilibria, we repeatedly sample finite Walrasian markets with n buyers and m goods from the WSBM, and find a solution of this finite market. The sampling from a WSBM is straightforward: assign each buyer a buyer-type (good a good-type) according to the categorical distribution defined by ξ^{buyer} (ξ^{good}), assign budgets for each

buyer according to his buyer-type via g^{budget}, and assign a valuation of each buyer for each good via g^{value} as a function of their buyer and good types. To set the sample size, we need to balance the tradeoff between a fine enough representation and having the resulting WM instance be tractable. We sample and solve increasingly larger instances until the outcome summary values converge or the original $(n^{\texttt{AM}}, m^{\texttt{AM}})$ is reached. The equilibrium of the sampled WM is computed via tatonnement. Finally, we summarize the equilibrium by computing the resulting summary statistics. Due to variations arising from the sampling procedure, we repeat the above process many times in order to get a distribution over the summary statistic of interest. We then take the mean over all sampled WMs, and use that as the predicted summary statistics.

Appendix 1.E Evaluation of Internal Consistency

In this section we explore the convergence properties of the algorithm $f^{\texttt{WM}}$ introduced in Sect. 1.4.

1.E.1 Tatonnement Convergence

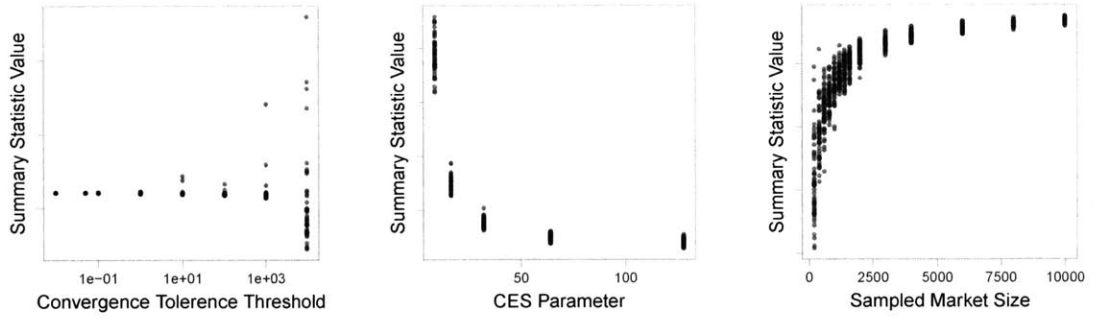
We show how well $f^{\texttt{WM}}$converges with respect to the convergence tolerance threshold ϵ and initial price vector $\boldsymbol{p}^{\texttt{init}}$. We run simulations for a particular Walrasian market $\boldsymbol{x}^{\texttt{WM-Exp1}}$ with other $f^{\texttt{WM}}$parameters fixed. We consider a vector of possible convergence thresholds $\boldsymbol{\epsilon}$ and a vector of possible initial price vectors $\boldsymbol{p}$. For each $(\epsilon, \boldsymbol{p}^{\texttt{init}})$ combination, we compute summary statistics from the resulting market outcome as $z = g^{\texttt{WM-Exp1}}(f^{\texttt{WM}}(\boldsymbol{x}^{\texttt{WM-Exp1}}, \boldsymbol{p}^{\texttt{init}}, \lambda, \epsilon, \kappa))$.

Particular settings of experiment parameters are as follows. The Walrasian market $\boldsymbol{x}^{\texttt{WM}}$ was randomly generated by sampling from a particular WSBM $\boldsymbol{w}^{\texttt{WM-small}}$. Each initial price vector was randomly generated by sampling d i.i.d. values from distribution $\mathcal{U}$, and then we assigned all goods of good type $l \in [d]$ to the same sampled value.[1] The summary statistic was the average price across goods of a particular type $l \in [d]$: $g_l(\boldsymbol{y}^{\texttt{WM}}) = \sum_{j=1}^{m} 1\{d_j = l\} p_j \;/\; |\{j \in [m] : d_j = l\}|$. The parameter values used for the experiment were: $\lambda = 1/255$, $\kappa = 10000$ (but the maximum number of iterations was never reached in our experiments), $\boldsymbol{\epsilon} = (10^{-3}, 10^{-2}, \ldots, 10^{4})$, $c = 20$, $d = 14$,

[1] We also performed the same analysis where we randomly initialized prices i.i.d. for each good, instead of all at the same price for each block. Results were similar.

$n = 1000$, $m = 1000$, $n^{\texttt{iters}} = 50$, and $\mathcal{U} = \texttt{unif}(0, 600)$. Other WSBM parameters $(\xi^{buyer}, \zeta^{good}, \boldsymbol{g}^{budget}, \boldsymbol{g}^{value})$ were calibrated on data from a real auction market.

Figure 1-14a shows a scatter plot of (ϵ, z) values. Each point corresponds to an $(\epsilon, \boldsymbol{p}^{\texttt{init}})$ pair. The x-axis gives the convergence threshold parameter ϵ and the y-axis gives the output summary statistic value z for the particular initial price vector $\boldsymbol{p}^{\texttt{init}}$. When points on the y-axis are far apart for a given x-axis value, it indicates that $f^{\texttt{WM}}$ terminated at different summary statistic values for different initial price vectors. For sufficiently low convergence tolerance thresholds, we find that $f^{\texttt{WM}}$ produces consistent summary statistic values, regardless of initial price vector. These results complement the theoretical convergence guarantees given in Cole and Fleischer (2008). For subsequent experiments, we fix the convergence tolerance to $\epsilon = 0.01$, which corresponds to stopping $f^{\texttt{WM}}$when excess demand for each good j is no more than 1% of that good's supply.



(a) Tatonnement Convergence (b) CES Approximation (c) Sampling Consistency

Figure 1-14: Example of a synthetic output. a The average price of all goods within a representative good type, computed via $f^{\texttt{WM}}$ for different initial prices $\boldsymbol{p}^{\texttt{init}}$ and convergence tolerance thresholds ϵ. b The equilibrium average price of all goods within a representative good type, computed for a variety of different elasticity of substitution parameters σ. c The equilibrium average price of all goods within a representative good type, computed from sampled markets of increasing size.

1.E.2 CES Convergence

As noted in Sect. 1.4.1 and described in more detail in Appendix 1.B, we approximate linear utilities with CES utilities. In this section, we show that in practice the output of $f^{\texttt{WM}}$ converges for sufficiently large values of σ, validating the computational approach.

We consider a vector of possible CES parameters $\boldsymbol{\sigma}$ and a vector of possible initial price vectors $\boldsymbol{p}$. Let $\boldsymbol{x}^{\texttt{WM-Exp1}}(\sigma)$ be a Walrasian market that is identical to $\boldsymbol{x}^{\texttt{WM-Exp1}}$ except that buyers' utility functions are computed using CES parameter σ. For each $\sigma \in \boldsymbol{\sigma}$ and $\boldsymbol{p}^{\texttt{init}} \in \boldsymbol{p}$, we compute the average price $z = g_l^{\texttt{WM-Exp1}}(f^{\texttt{WM}}(\boldsymbol{x}^{\texttt{WM-Exp1}}(\sigma), \boldsymbol{p}^{\texttt{init}}, \lambda, \epsilon, \kappa))$ across goods of type $l \in [d]$. The vector of CES parameters we consider is $\boldsymbol{\sigma} = (8, 16, 32, 64, 128)$. All other parameters take on the same values as in Sect. 1.E.1.

Figure 1-14b shows a scatter plot of (σ, z) values. We find that, for sufficiently high CES parameters σ, the tatonnement algorithm $f^{\texttt{WM}}$ produces consistent summary statistic values, regardless of initial price vector. As a result of these experiments, we let the CES parameter σ equal 128.

1.E.3 WSBM Convergence

In this section, for a given WSBM, we look at the convergence of relevant summary statistics as the number of buyers and goods in the sampled WM increase. It can be shown via standard methods that this sampling procedure converges as the size of the sampled markets tends to ∞, under some restrictive conditions (see Appendix 1.F). The computational experiments in this section provide evidence that converge holds in general.

We consider a vector of possible market sizes $\boldsymbol{k}$ and a vector of $n^{\texttt{iters}}$ possible initial price vectors $\boldsymbol{p}$. Let $\boldsymbol{p}_i^{\texttt{init}}$ be the ith element of $\boldsymbol{p}$. Let $\boldsymbol{w}^{\texttt{WM-Exp1}}(k)$ be a WSBM that is identical to $\boldsymbol{w}^{\texttt{WM-Exp1}}$ except that the number of buyers and goods are both equal to k. Let $\boldsymbol{x}^{\texttt{WM-Exp1}}(k, i)$ be the ith Walrasian market sampled from $\boldsymbol{w}^{\texttt{WM-Exp1}}(k)$, where $i \in [n^{\texttt{iters}}]$. For each market size $k \in \boldsymbol{k}$ and sample $i \in [n^{\texttt{iters}}]$, we compute the average price $z_{l,k,i} = g_l^{\texttt{WM-Exp1}}(f^{\texttt{WM}}(\boldsymbol{x}^{\texttt{WM-Exp1}}(k, i), \boldsymbol{p}_i^{\texttt{init}}, \lambda, \epsilon, \kappa))$ across goods of type $l \in [d]$. We then compute the average across iterations as $z_{l,k} = \sum_{i=1}^{n^{\texttt{iters}}} z_{l,k,i}/n^{\texttt{iters}}$. The vector of market sizes we consider is $\boldsymbol{k} = (100, 200, 400, 600, 800, 1000, 1200, 1400, 2000, 3000, 6000, 8000, 10000)$. All other parameters take on the same values as in the preceding sections.

Figure 1-14c shows a scatter plot of $(k, z_{l,k})$ values for a particular good type $l \in [d]$.[2] As we sample larger markets, the amount of variation in the computed average prices decreases, and these prices appear to converge.

[2]Other good types $l \in [d]$ converged similarly.

Appendix 1.F Sampling Consistency: Finite Goods

This section shows that if goods can be grouped together so that all goods in a group are treated as identical by all buyers (so that we can assign a single price to each cluster), then sampling finite graphs and computing the Walrasian equilibrium on increasingly larger finite graphs leads to convergence of the estimated equilibrium prices. For this result, we assume the following.

- We have a stochastic block model.
- Goods can be grouped into G groups, and we can assign a single price to each group.
- The total population of buyers and goods in the stochastic block model model are N and M.
- Price can take values on some compact $P \subset \mathbb{R}^G$.
- Average demand of a random buyers is some $D : P \mapsto \mathbb{R}^G_{\geq 0}$ satisfying strict gross substitutes.
- Average supply of a random good is vector $S \in \mathbb{R}^G_{\geq 0}$.
- A Walrasian equilibrium is vector of G prices $p^* \in P$ such that $ND(p^*) = MS$, and such a p^* is assumed to exist.

To get the compact set B, we can bound prices from below and above. Bounding prices from below by a small positive value is sensible: this will be the case if some buyer values every good at some small positive amount, e.g. if an auction exchange would rather keep each good instead of selling it if the match value with all current bidders is too low. Bounding prices from above at some large positive value is natural in our case, as we can't imagine any situations where a single item could be unboundedly valuable to a bidder.

Our sampling scheme is as follows:

1. Draw n buyers and m goods from the block model.
2. Define the quantity of supply provided by a good j as $S_m \in \mathbb{R}^G$, $S_m \geq 0$.

3. Define sample average supply as

$$S_m := \frac{M}{m} \sum_{j=1}^{m} S_j.$$

4. We may define demand of each buyer as a function mapping from prices into quantities.

5. Define sample average demand as

$$D_n := \frac{N}{n} \sum_{i=1}^{n} D_i.$$

6. Compute equilibrium prices in the sampled market by finding p such that $S_m = D_n(p)$.

The intuition is as follows. A Walrasian equilibrium in this case is a set of prices, one price for each group of goods, so that markets clear. As we sample more and more goods, the total supply starts to look like the real supply, and likewise for demand as we sample more buyers. Thus, one having sampled enough buyers and users, we can compute the equilibrium on the sampled graph, and equilibrium prices should be the same.

From the law of large numbers, if S is the average supply of the block model, then $S_m \to MS$ as $m \to \infty$.

Due to strict gross substitutes, we have that D is injective, and thus D^{-1} exists. Due to compactness of P and continuity of D on P, we have that D^{-1} is continuous on its domain, which is compact by continuity of D and compactness of P. It follows by uniform law of large numbers (e.g. Newey and McFadden (1994)) that $D_n \to ND$ uniformly for $p \in P$.

In the sampled market, we compute Walrasian equilibrium as the price $p_{n,m}$ such that $D_n(p_{n,m}) = S_m$, so we have that

$$p_{n,m} = D_n^{-1}(S_m).$$

Choose some n such that $\sup_p ||D_n(p) - D(p)||_\infty < \epsilon$ with probability $\leq 1 - \epsilon$. Choose some δ such that for all S' in a δ ball around S, we have $|D^{-1}(S') - D^{-1}(S)| < \delta$. Choose some m such that $|S_m - S| \leq \delta$ with probability $\geq 1 - \delta$.

Then, we have that with probability at least $1 - \delta - \epsilon$,

$$\begin{aligned} |p^* - p_{n,m}| &= |D_n^{-1}(S_m) - D^{-1}(S)| \\ &\leq |D_n^{-1}(S_m) - D^{-1}(S_m)| + |D^{-1}(S_m) - D^{-1}(S)| \\ &\leq \epsilon + \epsilon \end{aligned}$$

Thus, as $n, m \to \infty$, $p_{n,m}$ converges to the equilibrium p^* of the stochastic block model in probability.

Appendix 1.G Convergence Bounds

Cole and Fleischer (2008) proved the following upper bound on convergence:

$$\eta(p) < \delta \quad \text{in} \quad O\left(\frac{1}{\alpha\beta\lambda} \log\left(\frac{\eta(p^0)}{\delta}\right)\right),$$

where

- $\eta(p) := \max_i(\eta_i(p_i))$ is a measure of how far p is from equilibrium prices p^*, with

$$\eta_i(p_i) := \begin{cases} \left|\frac{p_i - p_i^*}{p_i^*}\right| & \text{if } p_i^* < 3p_i \\ \frac{p_i^*}{p_i} & \text{else} \end{cases},$$

  where p^* is the equilibrium price vector of the goods;

- $\delta > 0$ is some positive tolerance;

- $\lambda := \frac{1}{2\sigma - 1}$ where σ is the elasticity of substitution term in the CES utility function in (1.7);

- $\alpha := \min_i \frac{1}{w_i} \sum_l \frac{x_{il}^* B_l}{wealth_l(p)}$, where w_i is the total supply of good i, $wealth_l(p)$ is the total budget of buyer l at prices p; and

- $\beta := \min_{i,l}(\partial x_{il}(p, B_l)/\partial B_l)/(x_{il}/B_l)$, is the minimum overall user, good pairs of the wealth effect.

As we are using the CES utility functions, it is well known that additional wealth increases demand for all goods uniformly, so that $\beta = 1$ (see e.g. (Mas-Colell et al., 1995, p.97)). Furthermore, because all the buyers in our case are endowed purely with

money, $wealth_l(p) = B_l$, so it's trivially the case that $\alpha = 1$. This yields the expression in (1.9).

Appendix 1.H Uniform budgets

In Definition 5, we construct a Walrasian market from a given auction market. This construction is straightforward, but has some issues with high variance when used to compute the summary statistics of interest in our markets. The issue stems from the fact that bidders in an auction market may have drastically different budgets, and thus the distribution of bidder budgets exhibits very heavy tails. This is an issue, as our procedure involves sampling finite Walrasian markets from a Walrasian stochastic block model, and the outcome of such a market can be heavily skewed if the market contains e.g. a single bidder which accounted for 10% of the total budget. This can lead to high variance in the market summary statistics.

To overcome this issue, in practice we define a modified Walrasian market, where instead of treating each bidder as a buyer, we treat each B *dollars of budget* as a buyer. For example, if we set $B = 1$, then an bidder i with a budget of \$100, targeting criterion t_i, value-per-click v_i, and click through rates γ_i would be treated as 100 separate buyers. Each of these 100 buyers would have a budget of \$1, together with the exact same targeting criterion, value-per-click, and click through rates.

Definition 8 (Walrasian Market Transformation (Uniform Budgets))**.** *Let* $f^{\text{WM-Linear-Uniform}} : X^{\text{AM}} \to X^{\text{WM}}$ *be a function that takes as input an auction market* $\boldsymbol{x}^{\text{AM}} = (n^{\text{AM}}, m^{\text{AM}}, \boldsymbol{v}^{\text{AM}}, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{s}, \boldsymbol{\gamma})$ *together with a budget variable* B*, and outputs a Walrasian market* $\boldsymbol{x}^{\text{WM}} = (n^{\text{WM}}, m^{\text{WM}}, \boldsymbol{e}, \boldsymbol{u})$ *such that* $n^{\text{WM}} = \sum_{i=1}^{n^{\text{AM}}} \lfloor b_i/B \rfloor + 1$*,* $m^{\text{WM}} = m^{\text{AM}} + 1$*,* $\boldsymbol{e} = (e_{i,j})_{i \in [n^{\text{WM}}], j \in [m^{\text{WM}}]}$*, where* $e_{i,j} = \{B \text{ if } i \in [n^{\text{AM}}] \wedge j = m^{\text{AM}}+1; 1 \text{ if } i = n^{\text{AM}}+1 \wedge j \in m^{\text{AM}}; 0 \text{ otherwise}\}$*, and* $\boldsymbol{u} = (u_i)_{i \in [n^{\text{WM}}]}$*, where* $u_i(\boldsymbol{a}_i) = \left(\sum_{j=1}^{m^{\text{WM}}} a_{i,j}^{(\sigma-1)/\sigma} t_{i,j} \gamma_{i,j}\right)^{\sigma/(\sigma-1)}$.

That is, Definition 8 breaks each bidder of the auction market into a large number of buyers, each of which has a fraction of the budget. This breaking has no impact on the equilibrium, as both the linear and CES utility functions are separable in the budget (i.e. the optimal consumption of a buyer is a function of prices and budget where the budget enters multiplicatively). In the CES case, the optimal consumption is

$$a_{ij}^* = \frac{(v_i \gamma_{ij} t_{ij}/p_j)^\sigma}{\sum_{j'} (v_i \gamma_{ij'} t_{ij'}/p_{j'})^\sigma p_{j'}} b_i$$

and in the linear case,

$$a_{ij}^* = b_i / p_j \mathbb{I}(v_i \gamma_{ij} t_{ij} / p_j \text{ is maximal})$$

so that modeling a single bidder as a unified buyer or a number of fragmented of individual entities has no impact on the total demand of that bidder.

Similarly, we can define, in a fashion similar to Definition 6, a WSBM where we treat each buyer as some amount B of budget as opposed to a bidder:

Definition 9 (WSBM with Uniform Budgets (WSBMU)). *A* Walrasian stochastic block model with uniform budgets *is a tuple* $\boldsymbol{w}^{WM} = (n, m, c, d, \xi^{buyer}, \zeta^{good}, B, \boldsymbol{g}^{value})$ *where* $n \in \mathbb{N}$ *is the number of buyers,* $m \in \mathbb{N}$ *is the number of goods,* $c \in \mathbb{N}$ *is the number of buyer types,* $d \in \mathbb{N}$ *is the number of good types,* $\xi^{buyer} : [c] \to [0, 1]$ *is an i.i.d. distribution over buyer types,* $\zeta^{good} : [d] \to [0, 1]$ *is an i.i.d. distribution over good types,* $B \in \mathbb{R}_+$ *is the budget of each buyer* $\boldsymbol{g}^{value} = (g_{i,j}^{value} : \mathbb{R}_{\geq 0} \to [0, 1])_{i \in [c], j \in [d]}$ *is a tuple of i.i.d. distributions over values for each (buyer type, good type) pair.*

The key difference between Definitions 6 and 9 is that in the latter, each buyer has a fixed budget B, so we don't keep track of the distribution over budgets. This also necessitates some differences in how we define these quantities; in particular, ξ should now be weighted according to the total budget of each buyer-type instead of the raw count, and similarly for $\boldsymbol{g}^{value}$:

$$\xi^{buyer}(l) := \frac{\sum_{i \in [n]: c_i = l} b_i}{\sum_i b_i},$$

$$\xi^{good}(k) := \frac{\#\{j \in [m] : d_j = k\}}{m},$$

$$g_{l,k}^{value}(x) := \frac{\sum_{(i,j) \in [n] \times [m]: c_i = l \wedge d_j = k \wedge v_i t_{i,j} \gamma_{i,j} \leq x} b_i}{\sum_{(i,j) \in [n] \times [m]: c_i = l \wedge d_j = k} b_i} \quad (\forall l \in [c], k \in [d]).$$

We take a high-dimensional Walrasian market as input and transform it into a WSBM which has a lower dimension but similar structural properties. We then proceed similarly to the steps taken for the WM abstraction (i.e., we solve for equilibrium and output the relevant summary statistics for the higher-dimensional problem). The first step is to choose partitions over the sets of bidders and queries. Let c be the number of components in the chosen partition of set $[n]$, and let $h^{buyer}(i) \in [c]$ indicate the

component that bidder i falls into. Similarly, let d be the number of components resulting from the chosen partition of set $[m]$, and let $h^{good}(j) \in [d]$ indicate the component that query j falls into.

We consider partitions equal to the market segments used for computing summary statistics. Given these partitions, computing stochastic block model parameters involves computing empirical distributions from the auction market. We let ξ^{buyer} be the empirical distribution of buyer-types and, similarly with ξ^{good}. We let g_l^{budget} be the empirical distribution of budgets for each buyer-type l, and $g_{l,k}^{value}$ be the empirical distribution of $v_i t_{i,j} \gamma_{i,j}$ for each (buyer-type, good-type) pair (l, k).

We use the constructed WSBM to stand-in for the full Walrasian market. To compute equilibria, we repeatedly sample finite Walrasian markets with n buyers and m goods from the WSBM, and find a solution of this finite market. The sampling from a WSBM is straightforward: assign each buyer a buyer-type (good a good-type) according to the categorical distribution defined by ξ^{buyer} (ξ^{good}), assign budgets for each buyer according to his buyer-type via g^{budget}, and assign a valuation of each buyer for each good via g^{value} as a function of their buyer and good types. To set the sample size, we need to balance the tradeoff between a fine enough representation and having the resulting WM instance be tractable. We sample and solve increasingly larger instances until the outcome summary values converge or the original $(n^{\mathtt{AM}}, m^{\mathtt{AM}})$ is reached. The equilibrium of the sampled WM is computed via tatonnement. Finally, we summarize the equilibrium by computing the resulting summary statistics. Due to variations arising from the sampling procedure, we repeat the above process many times in order to get a distribution over the summary statistic of interest. We then take the mean over all sampled WMs, and use that as the predicted summary statistics.

Algorithm 2: CalibrateSBM($\boldsymbol{x}^{\mathtt{WM}}_{(1)}, \ldots, \boldsymbol{x}^{\mathtt{WM}}_{(z)}$)

Input: auction market $\boldsymbol{x}^{\mathtt{AM}} = (n, m, \boldsymbol{v}, \boldsymbol{b}, \boldsymbol{t}, \boldsymbol{\gamma})$, buyer-types $1, ..., e$, good-types $1, ..., d$, bidder-to-buyer-type mapping $h^{buyer} : \{1, ..., n\} \mapsto \{1, ..., e\}$, query-to-good-type mapping $h^{good} : \{1, ..., m\} \mapsto \{1, ..., d\}$.

Output: Stochastic block model parameters $\boldsymbol{w}^{\mathtt{WM}}$, where
$\boldsymbol{w}^{\mathtt{WM}} = (N, M, e, d, \xi^{buyer}, \zeta^{good}, (g_i^{budget})_{i\in[c]}, (g_{i,j}^{value})_{i\in[c],j\in[d]})$

$N \leftarrow n$

$M \leftarrow m$

for $l \in [e]$ **do**

$$\xi^{buyer}(l) := \frac{\#\{i \in [n] : h^{buyer}(i) = l\}}{n}$$

end

for $k \in [d]$ **do**

$$\zeta^{good}(k) := \frac{\#\{j \in [m] : h^{good}(j) = k\}}{m}$$

end

for $l \in [e], x \in \mathbb{R}_+$ **do**

$$g_l^{budget}(x) := \frac{\#\{i \in [n] : h^{buyer}(i) = l \wedge b_i \leq x\}}{\#\{i \in [n] : h^{buyer}(i) = l\}} \quad (\forall l \in 1 \ldots e)$$

end

for $l \in [e], k \in [d], x \in \mathbb{R}_+$ **do**

$$g_{l,k}^{value}(x) := \frac{\#\{(i,j) \in [n] \times [m] : h^{buyer}(i) = l \wedge h^{good}(j) = k \wedge v_i t_{i,j} \gamma_{i,j} \leq x\}}{\#\{(i,j) \in [n] \times [m] : h^{buyer}(i) = l \wedge h^{good}(j) = k\}} \quad (\forall l \in [e], k \in [d])$$

end

Algorithm 3: SampleWM ($\boldsymbol{w}^{\mathtt{WM}}$)

Input: WSBM $\boldsymbol{w}^{\mathtt{WM}} = (N, M, c, d, \xi^{buyer}, \zeta^{good}, \boldsymbol{g}^{budget}, \boldsymbol{g}^{value})$, scaling factor α
Output: $\boldsymbol{x}^{\mathtt{WM}} = (n, m, \boldsymbol{e}, \boldsymbol{u})$
$n \leftarrow \lceil \alpha N \rceil$
$m \leftarrow \lceil \alpha M \rceil$
for $i \leftarrow 1 \dots n$ **do**
  $l_i \sim \xi^{buyer}$;
  $e_{i,0} \sim g_{l_i}^{budget}$;
  $v_{i,m^{\mathtt{WM}}} = 1$;
end
for $j \leftarrow 1 \dots m$ **do**
  $k_j \sim \zeta^{good}$;
  for $i \leftarrow 1 \dots n$ **do**
    $v_{i,j} \sim g_{l_i,k_j}^{value}$;
    $e_{i,j} \leftarrow 0$;
  end
end
for $i \leftarrow 1 \dots n$ **do**
  $u_i(\boldsymbol{a}_i) \leftarrow \sum_{j=1}^{m^{\mathtt{WM}}-1} a_{i,j} v_{i,j}$;
end
$\boldsymbol{x}^{\mathtt{WM}} \leftarrow (n, m, \boldsymbol{e}, \boldsymbol{u})$;
return $\boldsymbol{x}^{\mathtt{WM}}$

Algorithm 4: ComputeSummaryStats $(\boldsymbol{g}^{\mathtt{WM}}, q, (\boldsymbol{w}^{\mathtt{WM}}, \sigma), (\boldsymbol{p}^{\mathtt{init}}, \lambda, \epsilon, \kappa))$

Input: $\boldsymbol{g}^{\mathtt{WM}} = (g_j^{\mathtt{WM}} : Y^{\mathtt{WM}} \to \mathbb{R})_{j\in[k]}$ vector of k summary functions
$q \in \mathbb{N}$: number of samples
$\boldsymbol{w}^{\mathtt{WM}}$: stochastic block model parameters
σ : CES parameter
$(\boldsymbol{p}^{\mathtt{init}}, \lambda, \epsilon, \kappa)$: WM solver parameters

Output: $\boldsymbol{z}^{\mathtt{WM}} = (\hat{z}_j^{\mathtt{WM}})_{j\in[k]}$: estimated summary values

for $i \leftarrow 1 \dots q$ **do**
  $\boldsymbol{x}_{(i)}^{\mathtt{WM}} \leftarrow \mathsf{SampleWM}(\boldsymbol{w}^{\mathtt{WM}}, \sigma)$;
  $\boldsymbol{y}_{(i)}^{\mathtt{WM}} \leftarrow f^{\mathtt{WM}}(\boldsymbol{x}_{(i)}^{\mathtt{WM}}, \boldsymbol{p}^{\mathtt{init}}, \lambda, \epsilon, \kappa)$;
  for $j \leftarrow 1 \dots k$ **do**
    $\hat{z}_{(i),j}^{\mathtt{WM}} \leftarrow g_j^{\mathtt{WM}}(\boldsymbol{y}_{(i)}^{\mathtt{WM}})$;
  end
end
for $j \leftarrow 1 \dots k$ **do**
  $\hat{z}_j^{\mathtt{WM}} \leftarrow \frac{1}{q}\sum_{i=1}^{m} \hat{z}_{(i),j}^{\mathtt{WM}}$;
end
$\hat{\boldsymbol{z}}^{\mathtt{WM}} \leftarrow (\hat{z}_j^{\mathtt{WM}})_{j\in[k]}$;
return $\hat{\boldsymbol{z}}^{WM}$

Algorithm 5: ComputeSummaryStatsScaled $(\boldsymbol{g}^{\mathtt{WM}}, q, (\boldsymbol{w}^{\mathtt{WM}}, \sigma), (\boldsymbol{p}^{\mathtt{init}}, \lambda, \epsilon', \kappa'), (\alpha, \epsilon, \kappa))$

Input: $\boldsymbol{g}^{\mathtt{WM}} = (g_j^{\mathtt{WM}} : Y^{\mathtt{WM}} \to \mathbb{R})_{j\in[k]}$ vector of k summary functions
$q \in \mathbb{N}$: number of samples
$\boldsymbol{w}^{\mathtt{WM}} = (n, m, \widetilde{\boldsymbol{w}}^{\mathtt{WM}})$
σ : CES parameter
$(\boldsymbol{p}^{\mathtt{init}}, \lambda, \epsilon', \kappa')$: WM solver parameters
$\alpha \in (0, 1]$: initial graph size scaling factor
$\epsilon \in \mathbb{R}_{\geq 0}$: convergence threshold
$\kappa \in \mathbb{N}$: max iterations

Output: $\boldsymbol{z}^{\mathtt{WM}} = (\hat{z}_j^{\mathtt{WM}} \in \mathbb{R})_{j\in[k]}$: estimated summary values

$\alpha_{(1)} \leftarrow \alpha$; $\delta_{(1)} \leftarrow \infty$; $\delta_{(2)} \leftarrow \infty$; $i \leftarrow 1$;
while $i \leq \kappa$ *and* $\delta_{(i)} > \epsilon$ **do**
  $n_{(i)} \leftarrow \lceil \alpha_{(i)} n \rceil$;
  $m_{(i)} \leftarrow \lceil \alpha_{(i)} m \rceil$;
  $\boldsymbol{w}_{(i)}^{\mathtt{WM}} \leftarrow (n_{(i)}, m_{(i)}, \widetilde{\boldsymbol{w}}^{\mathtt{WM}})$;
  $\hat{\boldsymbol{z}}_{(i),j}^{\mathtt{WM}} \leftarrow \frac{1}{\alpha_{(i)}}$ ComputeSummaryStats$(\boldsymbol{g}^{\mathtt{WM}}, q, (\boldsymbol{w}_{(i)}^{\mathtt{WM}}, \sigma), (\boldsymbol{p}^{\mathtt{init}}, \lambda, \epsilon', \kappa'))$;
  if $i > 1$ **then**
    $\delta_{(i+1)} \leftarrow \max_{j\in[k]}\{|\hat{\boldsymbol{z}}_{(i),j}^{\mathtt{WM}} - \hat{\boldsymbol{z}}_{(i-1),j}^{\mathtt{WM}}|\}$;
  end
  $\alpha_{(i+1)} \leftarrow \alpha_{(i)} + \frac{1-\alpha}{\kappa}$;
  $i \leftarrow i + 1$;
end
$\boldsymbol{z}^{\mathtt{WM}} \leftarrow (\hat{z}_{(i-1),j}^{\mathtt{WM}})_{j\in[k]}$;
return $\boldsymbol{z}^{\mathtt{WM}}$

Algorithm 6: WSBMSD $(\boldsymbol{w}^{\mathtt{WM}})$

Input: WSBM $\boldsymbol{w}^{\mathtt{WM}} = (\tilde{N}, \tilde{M}, \tilde{e}, \tilde{d}, \tilde{\xi}^{buyer}, \tilde{\zeta}^{good}, \tilde{\boldsymbol{g}}^{budget}, \tilde{\boldsymbol{g}}^{value})$, supply multiples $\{\nu^1, ..., \nu^c\}$, demand multiples $\{\eta_1, ..., \eta_d\}$

Output: $\boldsymbol{w}^{\mathtt{WM}} = (N, M, c, d, \xi^{buyer}, \zeta^{good}, \boldsymbol{g}^{budget}, \boldsymbol{g}^{value})$

$N \leftarrow \tilde{N}$
$M \leftarrow \tilde{M} \sum_{k=1}^{d} \tilde{\zeta}^{good}(k)\nu_k$
$e \leftarrow \tilde{e}$
$d \leftarrow \tilde{d}$
for $k \leftarrow 1 \ldots d$ **do**
  $\zeta^{good}(k) \leftarrow \frac{\tilde{\zeta}^{good}(k)\nu_k}{\sum_{k'=1}^{d} \tilde{\zeta}^{good}(k')\nu_{k'}}$
end
for $l \leftarrow 1 \ldots e, x \in \mathbb{R}_+$ **do**
  $\boldsymbol{g}_l^{budget}(x) \leftarrow \tilde{\boldsymbol{g}}_l^{budget}(\eta_l x)$
end
$\boldsymbol{g}^{value} \leftarrow \tilde{\boldsymbol{g}}^{value}$ **return** $\boldsymbol{w}^{WM} = (N, M, c, d, \xi^{buyer}, \zeta^{good}, \boldsymbol{g}^{budget}, \boldsymbol{g}^{value})$

Chapter 2

Cognitive Simplicity and Consumer Choice

2.1 Introduction

How should we model consumer choice behavior in situations when products have large numbers of attributes? Traditional discrete choice models, originally designed for situations in which products only have a few attributes, typically exhibit two characteristics: (1) consumers have utilities defined as some function (e.g. linear combination) of product attributes, and (2) heterogeneity in preferences is modeled by distributions over these utility functions (e.g. a jointly normal distribution over coefficients of the linear combination). Models with this flavor have been highly successful in traditional economics and marketing applications when the set of attributes are small, but encounter some issues in situations where the dimensionality of the attribute space is large.

There are two main issues that traditional methods encounter when the attribute space is large. First, given the large number of attributes, it is cognitively implausible that consumers are taking into account such a large number of attributes when making choices. Motivated by this issue, some recent work have begun incorporating an assumption of cognitively simple/sparse preferences in consumer choice, where consumers only use a small number of attributes to evaluate products when making a choice. Such an assumption, however, leads to the second issue: how to tractably model consumer heterogeneity in preferences with such a model. Traditional discrete choice models typically assume that e.g. consumer utilities are linear combinations of attribute values, and that the coefficients on these attributes are drawn from a e.g. Gaussian distribution

or a mixture of Gaussians (the parameters of which are to be estimated). While such models are highly tractable when there are only a few attributes, they quickly become infeasible to estimate as the dimensionality of the attribute space grows.

In order to overcome these two issues, we propose a model of consumer choice that incorporates two features:

- Consumers have a latent 'true' utility function that is not sparse, but make choices by only taking into account a 'simplified' utility function that is derived from the latent utility function via the post-lasso.
- Heterogeneity in consumer preferences is modeled via a cost-of-cognition parameter that enters into the post-lasso utility simplification process, and thus determines the complexity of the consumers simplified utility functions.

These two features allow us to maintain both the assumption cognitive simplicity / sparsity in high dimensional settings while still allowing for preference heterogeneity in a tractable way.

This paper, in introducing such a model, contributes to a variety of literatures. Most directly, this paper constitutes a contribution to the discrete choice literature by introducing a tractable way of incorporating preference heterogeneity in the high-dimensional settings. As the model and estimation procedure presented in this paper can be directly applied to experimental data, this paper also contributes to the literature on choice-based conjoint analysis. By sufficiently restricting the choice set, a discrete choice model can be viewed as a model of consumer consideration behavior, and thus this paper contributes to that literature by introducing a tractable model of cognitively simple consumer consideration behavior. Finally, this paper contributes to the literature on cognitively constrained models of consumer choice by describing how to make such a model empirically viable and estimating it on experimental consideration data.

The paper is organized as follows: Section 2.2 discusses existing work in the various literatures mentioned above and how it relates to this paper. Section 2.3 introduces the baseline version of the model, and introduces the stochastic gradient descent based procedure for maximum likelihood estimation. Section 2.4 analyzes the ability of the estimation procedure to consistently recover the model parameters in the baseline model, finding that model parameters can be consistently estimated with enough data. Section 2.5 analyses performance of the estimated parameters on a consumer choice prediction task, finding that the estimated parameters out-perform more reduced form prediction

methods and approach the theoretical upper bound of prediction performance, even in situations where the model parameters are not very precisely estimated. Section 2.6 applies the baseline model to experimental data on consumer consideration, and uses the estimated parameters to give a sense of how large consumers' cognitive costs are. Section 2.7 explicitly microfounds the baseline model via a model of bounded rationality. Section 2.8 concludes.

2.2 Background

The present work fits within the framework of discrete choice, where consumers are presented with a set of products and must choose at most one out of that set to consume. Discrete choice models have been applied to a wide variety of problems in economics, marketing, transportation, operations management, and many other fields. Work in analyzing choice behavior goes back to at least to Mcfadden (1974), who introduced the innovation of deriving choice probabilities as a logit function of an agent's utility of consuming each of the products in the choice set, with products being defined via some attributes. To deal with population heterogeneity in a more systematic way, various authors (e.g. Train and Revelt (1998), Allenby and Rossi (1998) among many others) introduced the random coefficients logit, a generalization the standard conditional logit model that allowed for different agents to value different attributes differently by assuming some population distribution for the valuations (typically a multivariate Gaussian). This type of model has seen much use in economics, most notably in Berry et al. (1995) and subsequent work utilizing similar methodology. More recently, some authors have departed from the unimodal distribution typically assumed in random coefficient models by estimating more sophisticated semiparametric distributions on the population heterogeneity (e.g. Burda et al. (2008), Burda et al. (2013)) and thus allowing for much more general distributions over preferences.

With the advent of electronic commerce and the increasing availability of consumer choice data in which consumers choose among products with hundreds to thousands of recorded attributes, increasing attention has been focused on estimating models of consumer behavior where the attribute space is very high dimensional. Gillen et al. (2014) introduces a BLP-type model where the product attribute space is very large, but consumers' preferences only depend on a small subset of these attributes. Crucially, their work relies on the assumption that all consumers have the same sparse preferences

over product attributes, and heterogeneity enters only in the consumers preferences for price. With the aim of incorporating consumer heterogeneity with some element of sparsity, Hoderlein and Spindler (2014) presents a high dimensional random coefficient discrete choice model in which consumer preferences are drawn from an Gaussian distribution, with the central innovation (relative to the standard random coefficients logit) being that the mean vector of the Gaussian is sparse. They also provide an estimation procedure for this model, and show that it works well on synthetic data trials. The style of heterogeneity introduced in this work, by staying very close to the standard mixed logit model, generates non-sparse preferences that consumers use in decision making, while the mean of the underlying distribution is sparse. Our work contributes to this literature by introducing an alternative model of consumer preference heterogeneity that essentially adopts the reverse assumption, so that the underlying population preference vector is dense, while the individual preferences that consumers use for choice are sparse.

The marketing literature has also applied discrete models to the problem of conjoint analysis, where the aim is to estimate consumer preferences over products in an experimental setup where product attributes are explicitly manipulated (as opposed to most work in economics where estimation is done from observation data e.g. Berry et al. (1995)). Initially, conjoint analysis was performed by asking consumers to literally state how much they value a given good (see e.g. Green and Rao (1971) and Green and Srinivasan (1978)). Choice-based conjoint analysis, where instead consumers are merely asked to choose among a set of products, was introduced in Louviere and Woodworth (1983) and has since then become the dominant approach. One persistent problem with standard choice-based conjoint analysis is its inability to deal with large numbers of attributes. This limits the applicability of the method to settings in which there are relatively few attributes. Some ways of dealing with this problem by importing techniques from machine learning exist in the literature. For example, Evgeniou et al. (2005) proposes a method using cross-validated support vector machines to estimate choice-based conjoint. However, a drawback of such methods is their inability to introduce any sort of population structure: in the case of Evgeniou et al. (2005), the SVM must be run separately on each respondent's individual choice data, or some way of aggregating the separate agents' data must be used to produce a single unified data set. This paper's contribution to this literature is to produce a tractable method for choice-based conjoint that can be applied to high-dimensional attribute spaces while also incorporating population structure.

Our work also relates to the literature on consideration behavior, where consumers use some rough rules of thumb to pick a subset of products to more seriously consider (the consideration set) before choosing one out of that subset to purchase. In particular, we may view the consumer's decision of which goods to put into the consideration set as a choice problem in its own right. Models of consideration behavior can be roughly classified into two categories: compensatory and non-compensatory. In compensatory models of consideration behavior (e.g. Gilbride and Allenby (2004)), a consumer considers a product if and only if the utility of the product exceeds a certain threshold level, where utility is determined by an additive function of the part-worths of the attribute (i.e. an additive utility model with only first-order terms). Similarly, one can define probabilistic versions of this consideration set behavior by adding e.g. some random Gumbel or normal distributed shock to the utility of each product. Another compensatory model is the q-compensatory model, which the part-worths (the coefficients on each attribute in the linear utility function) are restricted such that the largest part-worth is not more than q times as large as the smallest (e.g. Yee et al. (2007)). One significant drawback of compensatory consideration rules lies in its implausibility as a literal algorithmic model of human consideration behavior. In a compensatory model, the agent must compute the utility of each product, compare it to the cutoff level, and decide whether to put it into the consideration set or not. If the good has many attributes, then this is process is quite cognitively taxing, and thus intuitively implausible. To explicitly capture the notion of cognitive simplicity in consideration behavior, many researchers have introduced non-compensatory models of consideration. In such models, consideration sets are specified directly, without reference to any underlying utility function. Some examples of such noncompensatory models include: disjunctive consideration rules (e.g. Gilbride and Allenby (2004)), which stipulate that an agent considers a product IFF the product has at least one attribute out of some set of desirable attributes; conjunctive consideration rules which state that an agent considers a product if and only if it has all the attributes out of some set of desirable attributes, disjunction-of-conjunctions consideration rules which allow the consideration set to be any subset expressible in terms of a disjunction of conjunctions (e.g. Hauser et al. (2010)). Other examples include subset conjunctive consideration rules, in which a product is considered IFF it has at least $t \geq 1$ attributes out of some set of $n \geq t$ desirable attributes (e.g. Jedidi and Kohli (2005)), lexicographic rules in which agents iteratively consider products by considering a single attribute at a time and taking only the products with the max value of that attribute. In addition, there are some probab-

ilistic versions of some of these procedures (e.g. Jedidi and Kohli (2005)), which allows agents to probabilistically deviate from the stipulations of the consideration rules (i.e. make mistakes). Common to most noncompensatory models of consideration rules is some notion of cognitive complexity. For example, Hauser et al. (2010) defines complexity via the number of attributes appearing in the DOC form of the rule, and penalizes complexity by restricting the set of acceptable rules to be all DOC rules with 4 or fewer attributes. In this case, the penalization can be seen as a penalty function $P(\text{rule}) = 0$ for rules with ≤ 4 attributes and $= \infty$ on rules with more than 4 attributes. However, an issue with such noncompensatory models is their general intractability when the number of attributes gets large. For example, Hauser et al. (2010) constructs an inference problem over rules can only be solved approximately, and Goodman et al. (2008) must restrict themselves to 4 total attributes as their algorithm perform a random walk in the space of possible rules (which as size $2^{2^{|A|}}$ where A is the set of attributes). This paper contributes to the literature on consideration behavior by producing a model of consideration behavior that captures cognitive simplicity while maintaining computational tractability even as the dimensionality of the product attribute space gets large. The approach presented here may be viewed as a compensatory model of consideration behavior that nonetheless captures the intuition of cognitive simplicity.

The approach used to model cognitive simplicity in this paper is very similar to the formalism introduced by Gabaix (2014). Though the precise details of the model in this paper differs somewhat from the one in Gabaix (2014) in many respects, the basic idea underlying our model, that agents chooses a simple model to stand in for a complex one via a lasso-type procedure is the same one underlying Gabaix (2014). Thus, the present work can also be seen as a contribution to the literature on cognitively simple choice behavior by demonstrating how such a model could be estimated on empirical choice data, and using the model to derive some estimates of how significant consumers' cognitive costs are.

2.3 Baseline Model

This section introduces a baseline version of the model. Extensions are discussed in the appendix. Consider a setting where a number of consumers are faced with some set of decisions, each decision being whether to accept or reject a product. Each product is defined by a set of binary attributes, and we assume that consumers all

have a single identical 'latent' preferences, which is some linear combination of the attribute values. While all consumers have identical latent preferences, their decisions are actually driven by 'simple' preferences derived from the latent preferences, taking into account a cognitive cost term. Formally, our setup is as follows:

Definition 10. *Baseline Model:*

- *Consumers* $i = 1, ..., N$
- *Goods* $j = 1, ..., J$
- *Each good has up to* K *binary attributes, so that a good* j *may be identified with its attribute set* $X_j \in \{0,1\}^K$
- *All consumers share latent preferences* $\beta \in \mathbb{R}^K$
- *Each consumer has an IID exponentially distributed cognitive cost parameter* $\lambda_i \sim \mathcal{E}(1/\mu)$ *so that* $F_{\mathcal{E}}(a) = 1 - \exp\left(-\frac{1}{\mu}a\right)$
- *Each consumer* i *has simple preferences defined as the entries of* β *with absolute value greater than* λ_i*:*

$$\tilde{\beta}_i = \beta \odot s_i, \quad s_i \in \{0,1\}^K, \quad s_{ik} = \mathbb{I}[|\beta_k| > \lambda_i] \tag{2.1}$$

  where $\odot$ *denotes element-wise multiplication* $\mathbb{I}[|\beta_k| > \lambda_i]$ *takes on value 1 if* $|\beta_k| > \lambda_i$ *else 0.*

- *Consumer* i *has utility for good* j *defined by his simple preferences:*

$$u_{ij} := X_j \cdot \tilde{\beta}_i + \varepsilon_{ij} = X_j \cdot \beta \odot s_i + \varepsilon_{ij}$$

  where ε_{ij} *is an IID Gumbel shock.*

- *For each good* j*, consumer* i *chooses to consume good* j *IFF* $u_{ij} > 0$*. Let* Y_{ij} *denote this decision.*
- *The consumption decision* $\{Y_{ij}\}$ *of each consumer for each good, and the attributes* $\{X_j\}$ *of each good is recorded.*
- *The objective of the analyst is to estimate* β *and* μ *given* Y *and* X*.*

2.3.1 Model discussion

This setup in Definition 10 is very similar to the standard logit discrete choice setting, with the only deviation being that choice behavior is generated via simplified preferences, which are sparse vectors derived from a single latent preference (which is a non-sparse vector). This model has two major benefits, the first being (as we will discuss below) that this style of heterogeneity allows for tractable estimation of the structural parameters β and μ, and the second being that most consumers use only sparse subset of the full attribute set in order to make their decisions, with the sparseness of the subset determined by a cognitive cost term.

Intuitively, this preference simplification process can be understood as the consumer using only the 'most important' attributes of a good when making choices, and ignoring the rest, with the cost-of-cognition parameter λ_i controlling how much simplification happens (so that a larger λ_i leads to fewer attributes being nonzero in the simplified preferences $\tilde{\beta}_i$). The particular form of the simple preferences in Equation 2.1, which are generated by just taking the entries of the latent preferences β that are sufficiently large, can be microfounded as the outcome of a more general procedure which we describe in Section 2.7. Formally, this procedure runs a post-lasso least squares regression using λ_i as the regularization factor (see Procedure 4). This procedure generates the particular form of the simplified preferences in Equation 2.1 when the background distribution over goods has an attribute-attribute correlation matrix that is the identity matrix I_K. One could also introduce situations where the attributes were correlated, in which case the simplified preferences would have a somewhat more involved representation that the one described in Equation 2.1. We describe how this can be done in Section 2.7, as well as modifications to the estimation procedure that the more general model would require.

Preference heterogeneity in the model of Definition 10 is induced by the cost of cognition parameter λ_i; the larger λ_i, the smaller the set of attributes that the consumer considers. This heterogeneity is of a different form than typically seen in economics and marketing research, where the standard assumption is that individual preferences are e.g. normally distributed around some baseline. The form of heterogeneity in Definition 10 can be motivated as 'quality', i.e. there is a single true measure of the quality of a good (given by the latent preference vector β), while consumers, due to cognitive effort being costly, use a simplified models of this true quality measure when making choices. To model situations in which there is not a single unified measure of quality, but rather

multiple heterogeneous preferences, we can instead allow there to be multiple 'types', each with a different latent preference β, and then proceed as in Definition 10. Section 2.B discusses this extension, along with how the modified model could be estimated.

The setting of Definition 10 features consumers accepting and rejecting each product based purely on the utility of that product as compared to a baseline of 0. In the discrete choice setting this can be understood as the choice set of each decision consisting of only a single good, and the consumer's choice as whether or not to consume that good. This assumption is not crucial, and the estimation procedure we detail below can be easily modified to work in situations where the choice sets contain multiple goods. The model in Definition 10 can also be understood as the consideration phase of a consider-then-choose model, where consumers first decide which goods to consider before picking a good out of their consideration set. In this consideration setting, the model in Definition 10 can be interpreted as consumers looking at each good j, and choosing to put it in their consideration set if the goods exceeds a certain utility threshold; that is, a compensatory model of consideration behavior (compensatory in that there is an additive utility function driving the consideration behavior) that incorporates both cognitive simplicity (simple preferences $\tilde{\beta}_i$ are sparse) and population structure (simple preferences are derived from a latent preference β).

2.3.2 Identification

Random coefficient discrete choice models in which the form of the utility function is linear are nonparametrically identified (see Fox and Gandhi (2016)). Given this results, it's thus possible to identify the distribution of simple preferences in the setup of Definition 10: the model may be viewed a random coefficient model with a particular structure for the coefficients, so that the distribution over simple preferences is nonparametrically identified. Given the distribution over simple preferences, and the direct correspondence between a consumer's cognitive cost and his simple preferences, it follows that with a sufficiently structured distribution for the cost-of-cognition parameter λ_i (exponential in our case), the whole model is identified.

It should be noted that while random coefficient mixture models are identified, they are typically very hard to estimate in high dimensions. For example, it has been shown that the parameter estimates from BLP-type model exhibit dependency on the initialization of the parameter estimates (see Dubé et al. (2012)), and that the nonparametric estimators of random coefficients models exhibit slow convergence, particularly in high

dimensional settings Hoderlein et al. (2010). Thus, the main technical hurdle for random coefficient models in high dimensions is not identification, but rather whether parameter values can be well estimated given finite data.

2.3.3 Estimation

The model in Definition 10 can be estimated via maximum likelihood. Due to the particular structure of the model, the gradient can be analytically derived, so that computing the max-likelihood parameter values via stochastic gradient descent with large numbers of consumers is computationally feasible.

Denote the set of all potential active attribute sets (i.e. the set of all possible s_i in 2.1) as S. As the attribute sets of each consumer depend on the cognitive cost parameter λ_i in a straightforward way, namely $s_{ik} = \mathbb{I}[|\beta_k| > \lambda_i]$, S is easy to characterize. Observe that for λ_i larger than the max absolute value of β_k, the active set will be the null set, so that the consumer's s_i will be a vector of zeros. Similarly, if λ_i is between the largest absolute value of β_k and the second-largest, then the active set of this consumer will consist of a single attribute, namely the attribute with largest absolute value, corresponding to an s_i for which $s_{ik} = 1$ IFF k is the attribute with the largest absolute value of β_k. Continuing in this fashion, it follows that the various admissible active attribute sets involve just listing the β_ks in terms of absolute value from largest to smallest, and taking the largest 0 attributes, or 1 attribute, or $, 2, 3, ..., K$ attributes.

Thus, the probability of a particular s occurring is just the probability that the value of λ_i is somewhere between the smallest value of $|\beta_k|$ for $k \in s$ and the largest value of $|\beta_k|$ for $k \notin s$. Thus, we can define

$$\overline{k}_s := \arg\min_{k \in s} |\beta_k|, \quad \underline{k}_s := \arg\max_{k \notin s} |\beta_k|$$

By convention, if s is empty, we define $\beta_{\overline{k}_s} = \infty$, and if the complement of s is empty, we define $\beta_{\underline{k}_s} = 0$. It then follows that

$$\begin{aligned} P(s_i = s|\beta, \mu) &= F_{\mathcal{E}}(|\beta_{\overline{k}_s}|) - F_{\mathcal{E}}(|\beta_{\underline{k}_s}|) \\ &= 1 - \exp\left(-\frac{1}{\mu}|\beta_{\overline{k}_s}|\right) - 1 + \exp\left(-\frac{1}{\mu}|\beta_{\underline{k}_s}|\right) \\ &= \exp\left(-\frac{1}{\mu}|\beta_{\underline{k}_s}|\right) - \exp\left(-\frac{1}{\mu}|\beta_{\overline{k}_s}|\right) \end{aligned}$$

Now, due to the error structure in the utility specification of each consumer, it follows that the probability of a consumer choosing a good is just the standard logit form, albeit with the simple preferences $\tilde{\beta}_i = \beta \odot s_i$:

$$Y_{ij} \sim \text{ Bernoulli } \left(\frac{\exp(X_j \cdot \beta \odot s)}{\exp(X_j \cdot \beta \odot s) + 1} \right)$$

Then, we can write

$$\begin{aligned}
P(X, Y|\beta, \mu) &= \prod_i \sum_{s \in S} P(s_i = s|\beta, \mu) \prod_j \left(\frac{\exp(X_j \cdot \beta \odot s)}{1 + \exp(X_j \cdot \beta \odot s)} \right)^{Y_{ij}} \left(\frac{1}{1 + \exp(X_j \cdot \beta \odot s)} \right)^{1 - Y_{ij}} \\
&= \prod_i \sum_{s \in S} P(s_i = s|\beta, \mu) \prod_j \left(\frac{\exp(X_j \cdot \beta \odot s Y_{ij})}{1 + \exp(X_j \cdot \beta \odot s)} \right) \\
\Rightarrow l(\beta, \mu|X, Y) &= \sum_i \log \left(\sum_{s \in S} P(s_i = s|\beta, \mu) \prod_j \left(\frac{\exp(X_j \cdot \beta \odot s Y_{ij})}{1 + \exp(X_j \cdot \beta \odot s)} \right) \right) \\
&= \sum_i \log \left(\sum_{s \in S} \rho_s \prod_j l_{ij}^s \right) \\
&= \sum_i \log(L_i) \qquad (2.2)
\end{aligned}$$

$$L_i = \sum_{s \in S} \rho_s \prod_j l_{ij}^s, \quad \rho_s = \exp\left(-\frac{1}{\mu} |\beta_{\underline{k}_s}| \right) - \exp\left(-\frac{1}{\mu} |\beta_{\overline{k}_s}| \right), \quad l_{ij}^s = \frac{\exp(X_j \cdot \beta \odot s Y_{ij})}{1 + \exp(X_j \cdot \beta \odot s)} \qquad (2.3)$$

Now, computing the gradients is straightforward: note that ρ_s depends on both β and μ, whereas and l_{ij}^s depends only on β. Thus, we have

$$\begin{aligned}
\frac{d\rho_s}{d\mu} &= \frac{1}{\mu^2} |\beta_{\underline{k}_s}| \exp\left(-\frac{1}{\mu} |\beta_{\underline{k}_s}| \right) - \frac{1}{\mu^2} |\beta_{\overline{k}_s}| \exp\left(-\frac{1}{\mu} |\beta_{\overline{k}_s}| \right) \\
\frac{d\rho_s}{d\beta_k} &= -\frac{1}{\mu} \exp\left(-\frac{1}{\mu} |\beta_{\underline{k}_s}| \right) \operatorname{sgn}(\beta_{\underline{k}_s}) \mathbb{I}[k = \underline{k}_s] + \frac{1}{\mu} \exp\left(-\frac{1}{\mu} |\beta_{\overline{k}_s}| \right) \operatorname{sgn}(\beta_{\overline{k}_s}) \mathbb{I}[k = \overline{k}_s]
\end{aligned}$$

and

$$\begin{aligned}\frac{dl_{ij}^s}{d\beta_k} &= \frac{X_{jk}Y_{ij}\exp(X_j\cdot\beta\odot sY_{ij})(1+\exp(X_j\cdot\beta\odot s))-X_{jk}\exp(X_j\cdot\beta\odot s)\exp(X_j\cdot\beta\odot sY_{ij})}{(1+\exp(X_j\cdot\beta\odot s))^2}s_k \\ &= l_{ij}^s X_{jk}\left(\frac{Y_{ij}(1+\exp(X_j\cdot\beta\odot s))-\exp(X_j\cdot\beta\odot s)}{1+\exp(X_j\cdot\beta\odot s)}\right)s_k \\ &= l_{ij}^s X_{jk}\left(\frac{Y_{ij}+(Y_{ij}-1)\exp(X_j\cdot\beta\odot s)}{1+\exp(X_j\cdot\beta\odot s)}\right)s_k\end{aligned}$$

The derivative of the complete log-likelihood is then:

$$\frac{dl(\beta,\mu|X,Y)}{d\mu} = \sum_i \frac{1}{L_i}\sum_{s\in S}\left(\frac{d\rho_s}{d\mu}\prod_j l_{ij}^s\right) \tag{2.4}$$

$$\begin{aligned}\frac{dl(\beta,\mu|X,Y)}{d\beta_k} &= \sum_i \frac{1}{L_i}\sum_{s\in S}\left(\frac{d\rho_s}{d\beta_k}\prod_j l_{ij}^s + \rho_s\sum_j \frac{dl_{ij}^s}{d\beta_k}\prod_{j'\neq j} l_{ij'}^s\right) \\ &= \sum_i \frac{1}{L_i}\sum_{s\in S}\left(\rho_s\prod_j l_{ij}^s\right)\left(\frac{1}{\rho_s}\frac{d\rho_s}{d\beta_k}+\sum_j\frac{1}{l_{ij}^s}\frac{dl_{ij}^s}{d\beta_k}\right)\end{aligned} \tag{2.5}$$

Given this, we may construct a stochastic gradient descent based algorithm for finding the β, μ that optimizes the log likelihood. The algorithm first initializes the estimate $\hat{\beta}$ by running a standard logistic regression for each individual separately, and then taking the per-attribute mean. This initialization is motivated by the fact that each consumer's simplified preferences is equal to the true latent preferences for the attributes in that consumer's active attribute set, and zero elsewhere. This suggests that in spite of the noise inherent in estimating the each consumer's simple preferences on just that consumer's choice data, the mean of the coefficient estimates, aggregated over all consumers, should point roughly in the same direction as the value of the latent β_k. The estimate of μ is then initialized at the mean of $|\hat{\beta})_k|$ over all values of k in order to ensure that each of the various active attribute sets has a sufficiently nonzero probability. After initialization, the log likelihood is maximized via stochastic gradient ascent with updates performed via Adagrad (see Duchi et al. (2011)).

Performing stochastic gradient descent instead of standard gradient descent has two advantages. First, given the large size of datasets that are likely to be encountered in practice, fitting the full gradient aggregated over all consumers is likely to be infeasible. Second, due to the nonconvexity in the objective function, stochastic gradient descent introduces noise in the gradients, and thus helps push the parameter estimates out of

local minima and towards more globally optimal solutions.

One additional caveat applies: the likelihood function exhibits many saddle points when the various $|\hat{\beta}_k|$ are zero (so that the corresponding gradients are zero), causing a standard stochastic gradient descent algorithm to get stuck. In order to circumvent this particular issue, a modification to the algorithm is made: when $\hat{\beta}_k$ falls below some level in absolute value, it is provided with some random shock to drive it away from zero. The size of this shock decreases with the iteration number, so that this modification has no effect as the algorithm is iterated over a long time frame. In practice, it appears to significantly improve the quality of the final estimates (in terms of log likelihood). The details of this estimation procedure can be found in Algorithm 7.

2.4 Estimation performance

In the previous section, the model and estimation procedure were defined, and it was noted that the model is identified. However, it remains an empirical question whether this identification translates into the estimation procedure being able to consistently estimated the model parameters. To that end, this section analyzes under which situations the true parameters of the model may be accurately recovered, and under which situations estimation is more problematic.

2.4.1 Identification

As a first step towards understanding the quality of the estimation procedure detailed in Algorithm 7, it is useful to plot the log likelihood function given finite data and look at its shape. If the likelihood appears to be maximized near the true parameter values, this would suggest that it in such cases it would be easier to consistently estimate the model parameters. On the other hand, if the likelihood at the true value isn't too much greater than the likelihood at other parameter values, that would suggest that estimation might be more difficult. In order to evaluate this claim, this subsection provides some synthetic experiments to plot the shape of the log likelihood function around the true parameter value. The experiments are constructed as follows:

Procedure 1. *Identification experiments:*

1. *Initialize N consumers, J goods (observations per consumer), K attributes.*
2. *Set the mean of the cost-of-cognition parameter $\mu = 1$.*

Algorithm 7: Stochastic gradient descent for maximizing the log likelihood

Input: T epochs, B batch size, r learning rate
for $i = 1, ..., N$ **do**
  $\hat{\beta}^i \leftarrow$ logistic regression of $\{Y_{ij}\}_{j=1,...,J}$ on $\{X_j\}_{j=1,...,J}$
end
for $k = 1, ..., K$ **do**
  $\hat{\beta}_k \leftarrow (1/N) \sum_{i=1}^{N} \hat{\beta}_k^i$
end
$\hat{\mu} \leftarrow (1/N) \sum_{i=1}^{N} |\hat{\beta}_k|$
$ll^0 \leftarrow l(\hat{\beta}, \hat{\mu}|X, Y)$ via Equations 2.2, 2.3
$\hat{\beta}^t \leftarrow \hat{\beta}$; // store estimated parameters
$\hat{\mu}^t \leftarrow \hat{\mu}$
$ss\_\mu \leftarrow 0$; // initialize adagrad terms
for $k = 1, ..., K$ **do**
  $ss\_\beta_k \leftarrow 0$
end
for $t = 1, ..., T$ **do**
  $\{i_1, ..., i_N\} \leftarrow$ permutation of $\{1, ..., N\}$; // permute consumers
  for $l = 1, ..., \lfloor N/B \rfloor$ **do**
    $\tilde{Y} \leftarrow \{Y_{ij}\}_{i \in i_{(l-1)B+1}, ..., i_{lB}}$; // take a batch of consumers
    $d\mu \leftarrow \frac{dl(\hat{\beta}, \hat{\mu}|X, \tilde{Y})}{d\mu}$ as in Equation 2.4 ; // compute derivatives
    for $k = 1, ..., K$ **do**
      $d\beta_k \leftarrow \frac{dl(\hat{\beta}, \hat{\mu}|X, \tilde{Y})}{d\beta_k}$ as in Equation 2.5
    end
    $ss\_\mu \leftarrow ss\_\mu + (d\mu)^2$; // update adagrad terms
    for $k = 1, ..., K$ **do**
      $ss\_\beta_k \leftarrow ss\_\beta_k + (d\beta_k)^2$
    end
    $\hat{\mu} \leftarrow \hat{\mu} + (r)(d\mu)(ss\_\mu)^{-1/2}$; // gradient updates
    for $k = 1, ..., K$ **do**
      $\hat{\beta}_k \leftarrow \hat{\beta}_k + (r)(d\beta_k)(ss\_\beta_k)^{-1/2}$
    end
    for $k = 1, ..., K$ **do**
      if $|\hat{\beta}_k| < .01\hat{\mu}$ **then**
        $\sigma \leftarrow$ uniform distribution on $[-.1\hat{\mu}, .1\hat{\mu}]$ $\hat{\beta}_k \leftarrow \hat{\beta}_k + (r)(\sigma)(ss\_\beta_k)^{-1/2}$; // push away from 0
      end
    end
  end
  $ll^t \leftarrow l(\hat{\beta}, \hat{\mu}|X, Y)$ via Equations 2.2, 2.3 ; // store likelihood
  $\hat{\beta}^t \leftarrow \hat{\beta}$; // store estimated parameters
  $\hat{\mu}^t \leftarrow \hat{\mu}$
end
$tmin \leftarrow \arg\max_{(t \in 1, ..., T)} ll^t$; // return max likelihood estimates
return $(\hat{\beta}^{tmin}, \hat{\mu}^{tmin})$

3. *Generate latent preferences* $\beta \sim N(0, I_K)$.

4. *Generate a good-attribute matrix* X *of size* $J \times K$ *by IID setting each attribute of each good to 1 or 0 with equal probability.*

5. *Generate a consumer choice matrix of size* $N \times J$ *as in Definition 10. Due to the choice of* β *and* X*, a consumer in expectation consumes half of the* J *goods and rejects the rest, though there is significant variance.*

6. *For each of 300 iterations, randomly draw a perturbation* ε *uniformly on* $[0, 1]$.

   (a) *Generate perturbed parameters by randomly increasing or decreasing each entry of* β *by* ε*, and randomly increasing or decreasing* μ *by* ε*:*

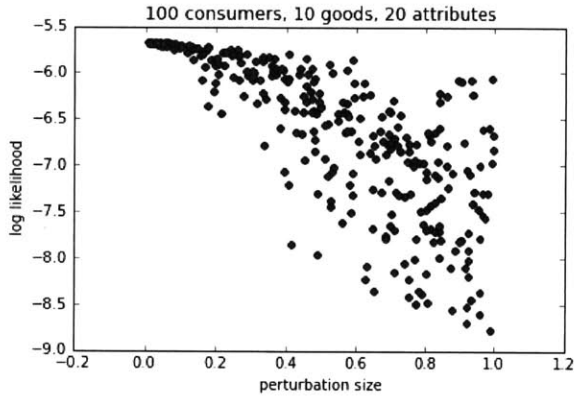
   $$\hat{\beta}_k = \beta_k + \varepsilon\sigma_k, \quad \hat{\mu} = \mu + \varepsilon\sigma_e, \quad \sigma_k, \sigma_e \sim \{-1, 1\} \text{ with equal probability, IID}$$

   (b) *Compute log likelihood of* $\hat{\beta}, \hat{\mu}$.

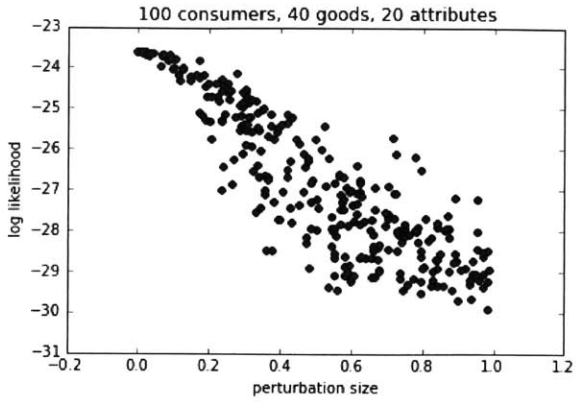
These log likelihoods, for various levels of N, J, K, z are presented in Figure 2-1. The model appears to be identified, with likelihood greatest when the parameters are close to the truth, though the various levels of N, J, K does seem to influence the shape of the likelihood.

The likelihood appears to be steeper when there are more observations per consumer (as in, larger J) as opposed to fewer (compare Figure 2-1a to Figure 2-1b, or Figure 2-1c to Figure 2-1d). In particular, when there are few observations per consumer relative to the the number of attributes (Figures 2-1a and 2-1c), there appear to be points far away from the origin that nonetheless have log likelihood values not too much worse than that of the true parameter values. This suggests that in such cases, estimation may be more difficult due to the presence of parameter values far from the true parameters that nonetheless have high likelihood. On the other hand, when the number of observations per consumer is large relative to the number of attributes (Figures 2-1b and 2-1d), the problem of far-away local parameter values with high likelihood seems to be somewhat less severe.

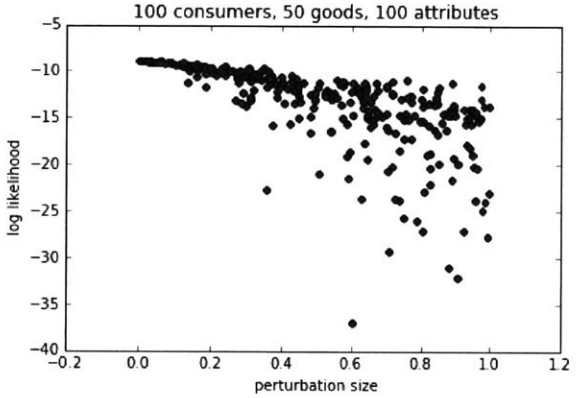
Increasing the number of consumers (N) without increasing the number of observations per consumer doesn't appear to have a particularly pronounced effect on the likelihood shape; comparing Figure 2-1e to 2-1a, or Figure 2-1f to 2-1c, the likelihoods appear to be very similar, in spite of the 10-fold difference in number of consumers.



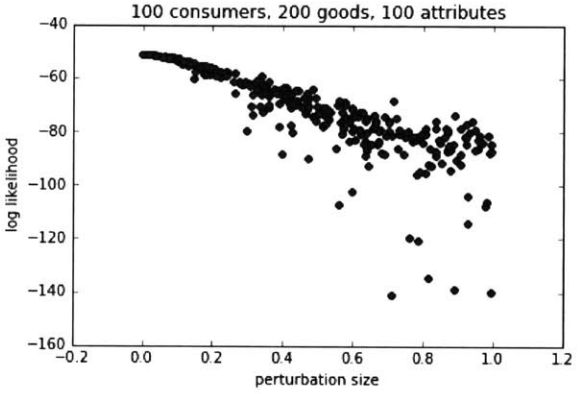
(a) With few attributes, and few observations per consumer



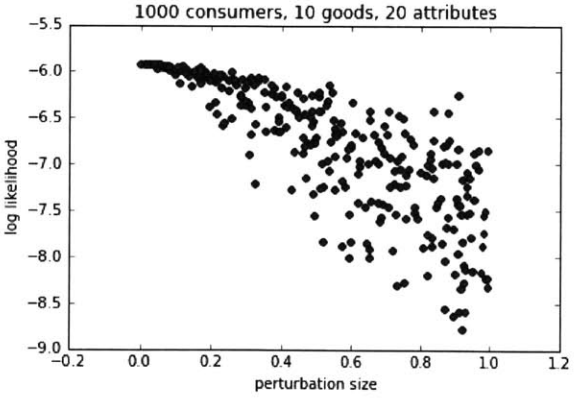
(b) With few attributes, and many observations per consumer



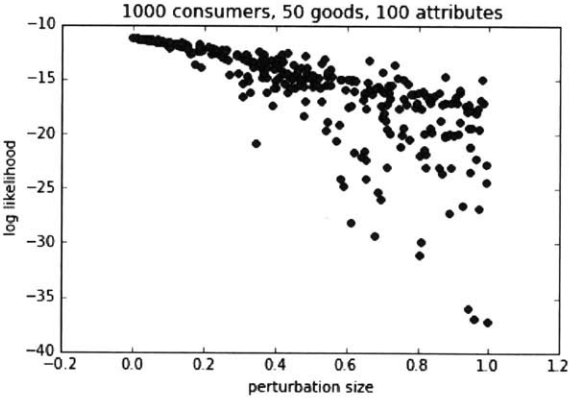
(c) With many attributes, and few observations per consumer



(d) With many attributes, and many observations per consumer



(e) With few attributes, few observations per consumer, but many consumers



(f) With many attributes, few observations per consumer, but many consumers

Figure 2-1: Log likelihood of various parameter values, plotted vs deviation from true parameter values. Each graph corresponds to a single (β, μ). Each point on a graph corresponds to $(\hat{\beta}, \hat{\mu})$ generated by perturbing (β, μ). Figures 2-1a, 2-1b, and 2-1e share the same β and μ parameters; likewise for Figures 2-1c, 2-1d, and 2-1f.

This is a bit of a surprise; in particular, as it suggests that observing the choice behavior of additional consumers is not particularly helpful; in our case, 100 consumers appears to be enough. Figure 2-1 seems to indicate that identification depends primarily on having many observations per consumer, as opposed to having many consumers with few observations each.

2.4.2 Estimation consistency

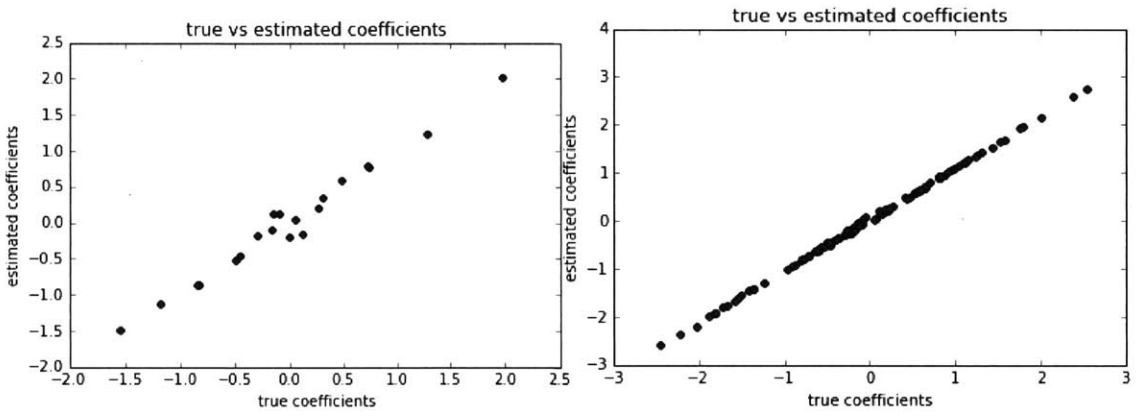
The above results regarding identification suggests that the model of Definition 10 is identified, though the identification is stronger when the number of observations per consumer is high. However, given the high dimensionality of the attribute space and the general nonconvexity of the log likelihood, being able to find the parameter values that maximize the log likelihood isn't necessarily trivial. This subsection presents some synthetic experiments on the ability of the estimation procedure described in Algorithm 7 to recover the true parameter values, and how its performance varies with various parameters of the problem. The results here largely coincide with those from the identification analysis above: Algorithm 7 estimates the model parameters very well when the number of observations per consumer is high, and less well when the number of observations per consumer is low. The synthetic experiments here are performed as follows:

Procedure 2. *Estimation consistency experiments:*

1. *Initialize N consumers, J goods, K attributes.*
2. *Set the mean of the cost-of-cognition parameter $\mu = 1$.*
3. *Generate latent preferences $\beta \sim N(0, I_K)$.*
4. *Generate a good-attribute matrix X of size $J \times K$ by IID setting each attribute of each good to 1 or 0 with equal probability.*
5. *Generate a choice matrix of size $N \times J$ as in Definition 10. Due to the choice of β and X, a consumer in expectation consumes half of the J goods and rejects the rest, though there is significant variance.*
6. *Run Algorithm 7, store computed parameter values.*

The parameters used in Algorithm 7 were hand tuned for convergence in a variety of settings. Precisely, they were $T = 50$ (SGD was repeatedly run until each consumer's data had been used 50 times before termination) $B = 5$ (each batch of the SGD used data from 5 consumers), and $r = .01$ (the scaling factor on the gradient update).

The maximum likelihood procedure in Algorithm 7 is able to recover the true parameters well in situations where the number of observations per consumers is larger than the number of attributes. As shown in 2-2, irrespective of the total number of attributes, when the number of observed choices per consumer is twice the number of attributes, the true latent preference parameter values β can be recovered very convincingly. Similarly, the mean of the cost-of-cognition parameter can also be estimated accurately.



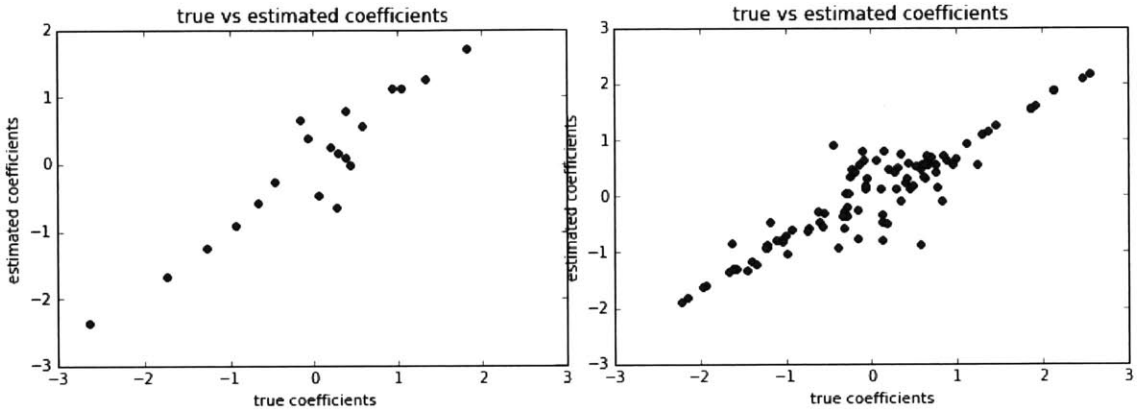
(a) 20 attributes, 40 observations per consumer. In the example here, estimated μ is 1.165 compared with the true value of 1. The log likelihood of the estimated parameters was -11.971, relative to the true parameters which stood at -11.953

(b) 100 attributes, 200 observations per consumer. In the example here, estimate μ is 1.071 compared with the true value of 1. The log likelihood of the estimated parameters was -57.679. compared to -57.571 for the true parameters.

Figure 2-2: Estimation when the amount of data per consumer is twice the number of attributes. In both of these representative examples, the estimated values of β line up very closely with the true parameter values. Each point on the graph corresponds to a single attribute. 1000 consumers.

The performance of the estimation procedure is less stellar when the amount of data is decreased. In situations where each consumer is observed making about as many choices as the total number of attributes, the maximum likelihood procedure of Algorithm 7 is still able to recover the values of large β_k, but has a more difficult time recovering β_k when that value is close to 0. This holds both when the number

of attributes is small, as well as when the number of attributes is large, as shown in Figure 2-3, where the true and estimated β_k line up very well when β_k is large and less well when β_k is small.



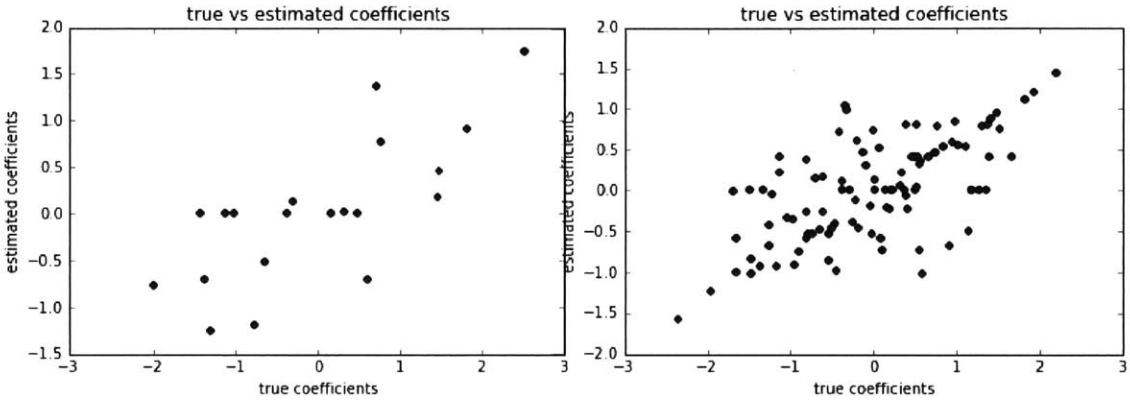
(a) 20 attributes, 20 observations per consumer. In the example here, estimated μ is 0.961 compared with the true value of 1. Log likelihood of the estimated parameters was -8.870, relative to the log likelihood of the true parameters which was -8.884.

(b) 100 attributes, 100 observations per consumer. In the example here, estimate μ is 0.838 compared with the true value of 1. Log likelihood of the estimated parameters was -32.187, relative to the log likelihood of the true parameters of -31.186.

Figure 2-3: Estimation when the amount of data per consumer is the same the number of attributes. In both of these representative examples, the estimated values of β_k line up very closely with the true parameter values when the true parameter values are large, and much less closely when the parameter values are close to 0. Each point on the graph corresponds to a single attribute. 1000 consumers.

In situations where data is even more scarce, recovering the true parameter values becomes even more difficult. As figure 2-4 shows, when the number of observations per consumer is half the number of attributes, it becomes difficult to accurately recover any parameter values. There is still a strong positive correlation between the estimated parameters and the true parameters, and for the case with 100 attributes the largest attributes are still estimated accurately (Figure 2-4b), but overall these results are a far cry from the results in Figure 2-2.

It is notable that the poor estimation performance when the number of observations per consumer is small relative to the number of attributes doesn't seem to be much alleviated with data on additional consumers. Figure 2-5, compares estimation performance using 1000 vs 100000 consumers, in a situation with 20 attributes and 10 observations per consumer. It doesn't appear to be a significant difference in the ability

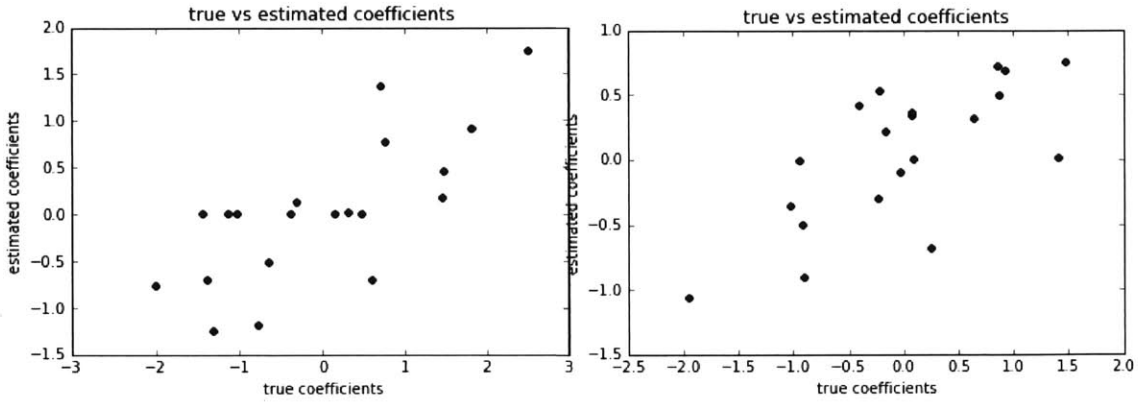


(a) 20 attributes, 10 observations per consumer. In the example here, estimated μ is 0.536 compared with the true value of 1. Log likelihood of estimated parameters was -4.383, relative to the -4.353 of the true parameters.

(b) 100 attributes, 50 observations per consumer. In the example here, estimate μ is 0.653 compared with the true value of 1. Log likelihood of the estimated parameters was -18.500, relative to -17.997 for the true parameters.

Figure 2-4: Estimation when the amount of data per consumer is the same the number of attributes. In both of these representative examples, the estimated values of β_k do not line up very closely with the true parameter values when the true parameter except when the values are quite large. Each point on the graph corresponds to a single attribute. 1000 consumers.

of the algorithm to recover the true parameter values in the two cases, in spite of the hundred-fold difference in the number of observations per consumer.



(a) 1000 consumers

(b) 100000 consumers

Figure 2-5: Effect of additional consumer data on estimation. 20 attributes, 10 observations per consumer.

2.4.3 Likelihood nonconvexity

As shown above, the estimation procedure detailed in Algorithm 7 appears to degrade in performance when the number of observations per consumer is small. This degradation seems to be driven at least partly by the fact that with fewer observations per consumer, the likelihood function becomes increasingly nonconvex, with many local optima, so that finding the maximum likelihood parameter estimates becomes increasing difficult. Furthermore, as the number of observations per consumer decreases, it appears that local optima far away from true parameter values tend to take on higher likelihood values, making it more likely for the estimation procedure to end up stuck at a local optimum far away from the true parameters. Synthetic experiments based on the following procedure are constructed to illustrate this issue of nonconvexity:

Procedure 3. *Estimation convergence experiments:*

1. *Initialize N consumers, J goods, K attributes.*

2. *Set the mean of the cost-of-cognition parameter $\mu = 1$.*

3. *Generate latent preferences $\beta \sim N(0, I_K)$.*

4. *Generate a good-attribute matrix X of size $J \times K$ by IID setting each attribute of each good to 1 or 0 with equal probability.*

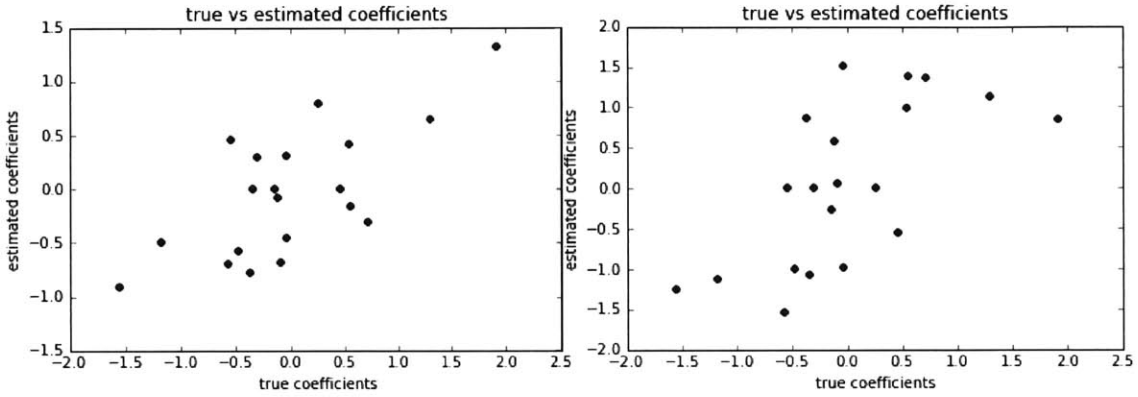
5. *Generate a choice matrix of size $N \times J$ as in Definition 10. Due to the choice of β and X, a consumer in expectation consumes half of the J goods and rejects the rest, though there is significant variance.*

6. *Run Algorithm 7, store the estimated parameter values.*

7. *Run Algorithm 7, except initialize the estimates of $\hat{\beta}$ randomly (instead of at the per-attribute means of the individual-level logit regressions). Store the estimated parameter values. Potentially, do this many times for multiple random initializations.*

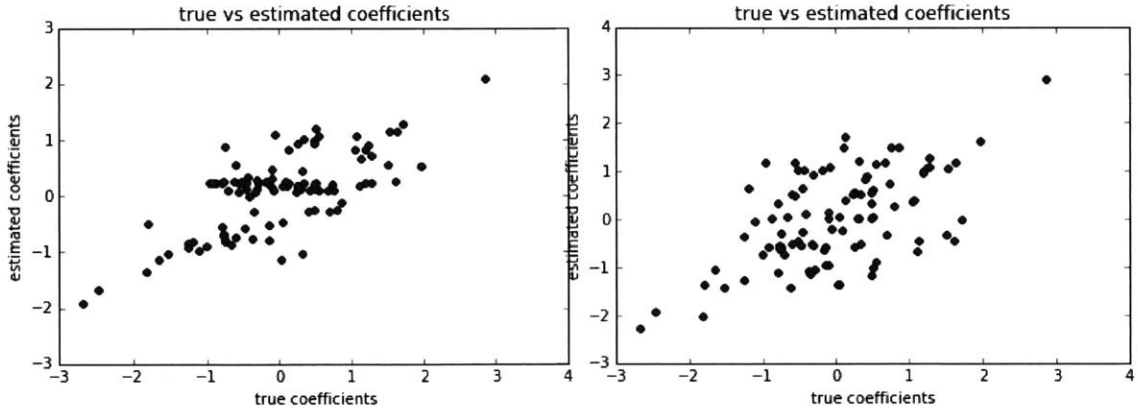
If the estimated $\hat{\beta}$ appear to be about the same, irrespective of initialization, then this would constitute evidence that there is a single optimum to which the estimation procedure is converging irrespective of initialization, or that the local optima of the likelihood function are sufficiently shallow that the stochastic gradient descent algorithm is able to jump out of them and approach the global optimum. On the other hand, if the estimated $\hat{\beta}$ appear to converge to different values depending on the initialization, then this would indicate that the presence of many local optima, some far away from the true value of β.

In Figure 2-6, the experimental procedure above is applied in situations where the number of observations per consumer J is half of the number of attributes K. As is apparent, the final converged-to values exhibits dependence on the choice of the initialization, and this holds for both when using 20 attributes with 10 observations per consumer, and 100 attributes with 50 observations per consumer. This indicates that the initialization of the estimate matters for the final estimated values, indicating nonconvexity of the log likelihood. To illustrate this point further, we perform the same procedure as above, except with many random initializations instead of one, and plot the estimated $\hat{\beta}$ in Figure 2-7. Again, as the multiple subgraphs of that figure show, the final estimated $\hat{\beta}$ varies significantly across initializations, again indicating that when the number of observations per consumer is small relative to the number of attributes, the total log likelihood function is highly nonconvex. I such cases, local optima appear to exist in areas of the parameter space far away from the true parameter vector β, making it difficult to estimate β well.



(a) 20 attributes, 10 observations per consumer, initializing at mean

(b) 20 attributes, 10 observations per consumer, initializing at random



(c) 100 attributes, 50 observations per consumer, initializing at mean

(d) 100 attributes, 50 observations per consumer, initializing at random

Figure 2-6: β vs estimated $\hat{\beta}$, computed with mean vs random initializations. 2-6a and 2-6b share the same β, μ, and synthetic data; similarly for 2-6c and 2-6d. In 2-6a and 2-6c, the initialization of $\hat{\beta}$ is as in Algorithm 7, whereas in 2-6b, 2-6d the $\hat{\beta}$ were initialized randomly. The computed $\hat{\beta}$ exhibits dependence on the choice of initialization.

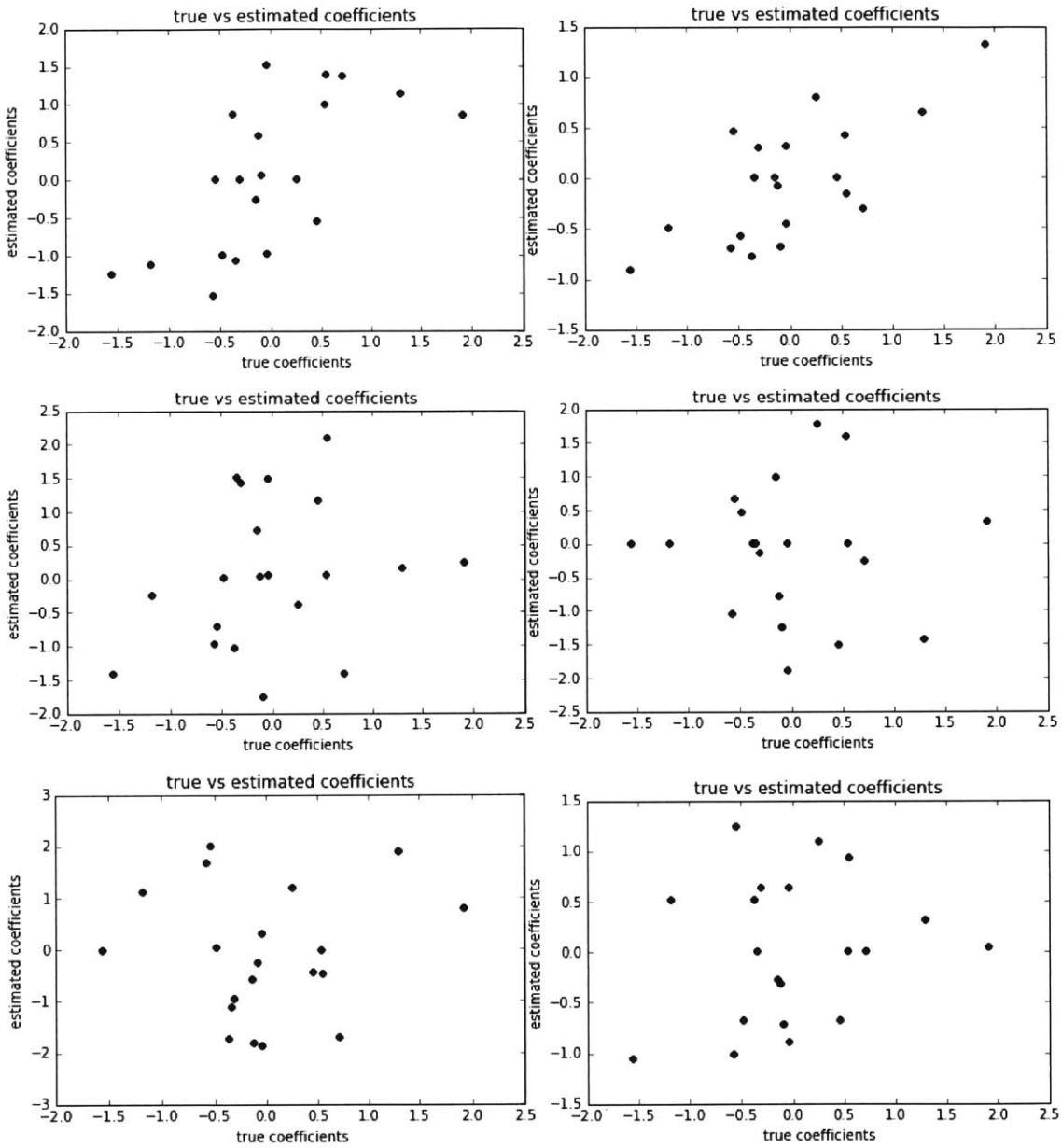


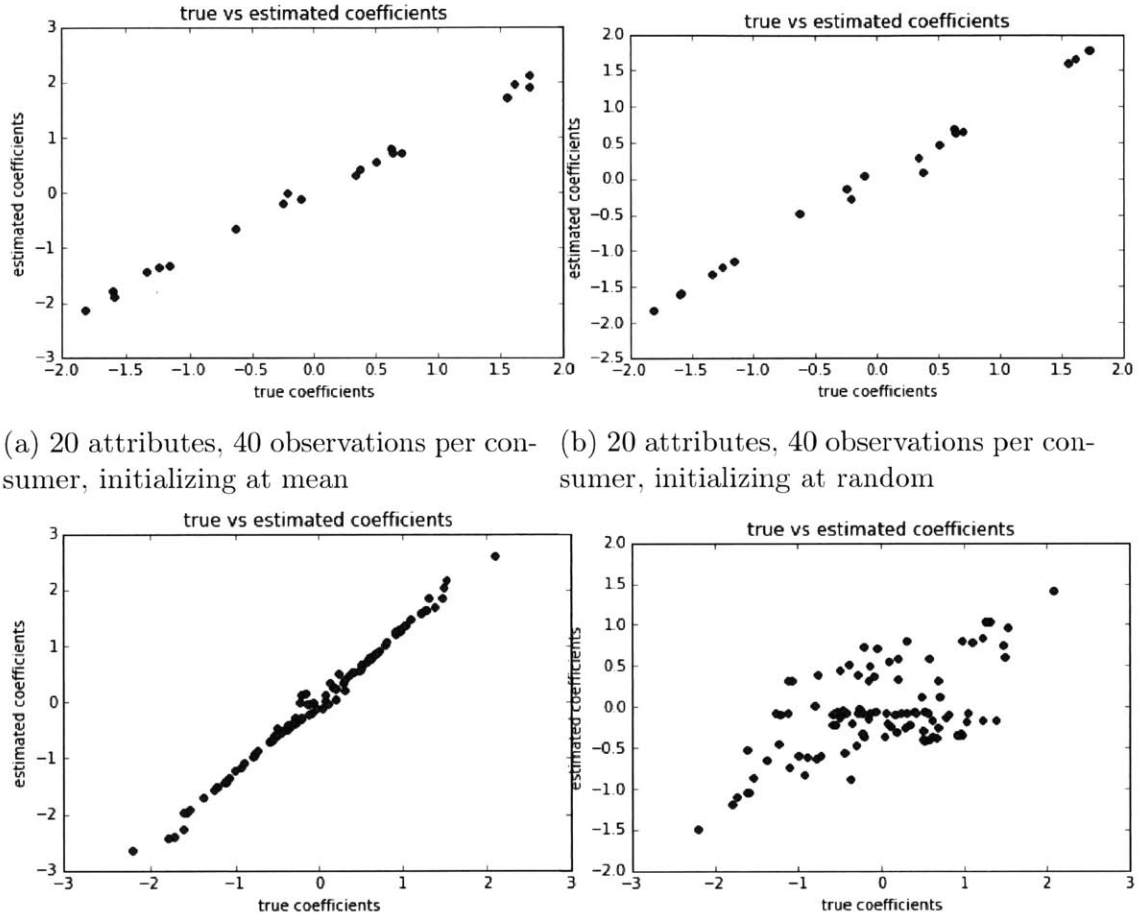
Figure 2-7: β vs estimated $\hat{\beta}$, computed with multiple random initializations. 20 attributes, 10 observations per consumers. β, μ, and the synthetic data were the same in all random initializations.

Now, the experimental procedure in Procedure 3 is also applied to situations where there is an abundance of observations per consumer compared to the number of attributes. Figures 2-8 and 2-9 contain the relevant output. The outlook in this case is mixed. When the number of attribute is relatively small (Figures 2-8a, 2-8b, and 2-9), performance is good in that irrespective of initialization, the estimated $\hat{\beta}$ is close to the truth. This suggests that local optima are much less severe of a problem, in that for the most part, they appear to be clustered very closely around the true parameters. There are, however, situations in which local minima do appear further away from the true β, e.g. the bottom right sub-figure of Figure 2-9, suggesting that the likelihood landscape may still exhibit a few local optima far away from the true β. However, when the number of attributes is large (Figures 2-8c, 2-8d), initialization still seems to matter. This feature seems to suggest that shape of the likelihood has significant nonconvexities, and that these nonconvexities become increasingly difficult to escape from as the dimensionality of the space gets large. In spite of this, it appears that initializing the estimates at the mean causes the final estimated values of $\hat{\beta}$ to reliably approximate to the true β in such cases.

2.5 Prediction performance

The previous section established that when the amount of data per consumer is sufficiently large, it's possible to precisely estimate the model parameters for the model in Definition 10, while if the amount of data per consumer is less abundant, the estimated parameters are much more loosely correlated with the true ones. While consistently estimating model parameters is important, this section considers another relevant metric: prediction performance.

Prediction performance is of independent interest, as in practical applications (e.g. an online retailer attempting to optimize which products to show which users) the purpose of estimating preferences is often not so much for its own sake, but rather for optimizing decision making. These tasks typically are closely related to how well a model predicts behavior out-of-sample (e.g. what matters is if a consumer ultimately clicks on or purchases a product the retailers shows to him). In addition, prediction serves as a useful test of the quality of the estimated parameters when these parameters do not exactly match the true ones (as in our case if the number of observations per consumer is low). In such situations, the estimated preference, though not identical



(a) 20 attributes, 40 observations per consumer, initializing at mean

(b) 20 attributes, 40 observations per consumer, initializing at random

(c) 100 attributes, 200 observations per consumer, initializing at mean

(d) 100 attributes, 200 observations per consumer, initializing at random

Figure 2-8: β vs estimated $\hat{\beta}$, computed with mean vs random initializations. 2-8a and 2-8b share the same β, μ, and synthetic data; similarly for 2-8c and 2-8d. In 2-8a and 2-8c, the initialization of $\hat{\beta}$ is as in Algorithm 7, whereas in 2-8b, 2-8d the $\hat{\beta}$ were initialized randomly. The computed $\hat{\beta}$ exhibits little dependence on the choice of initialization.

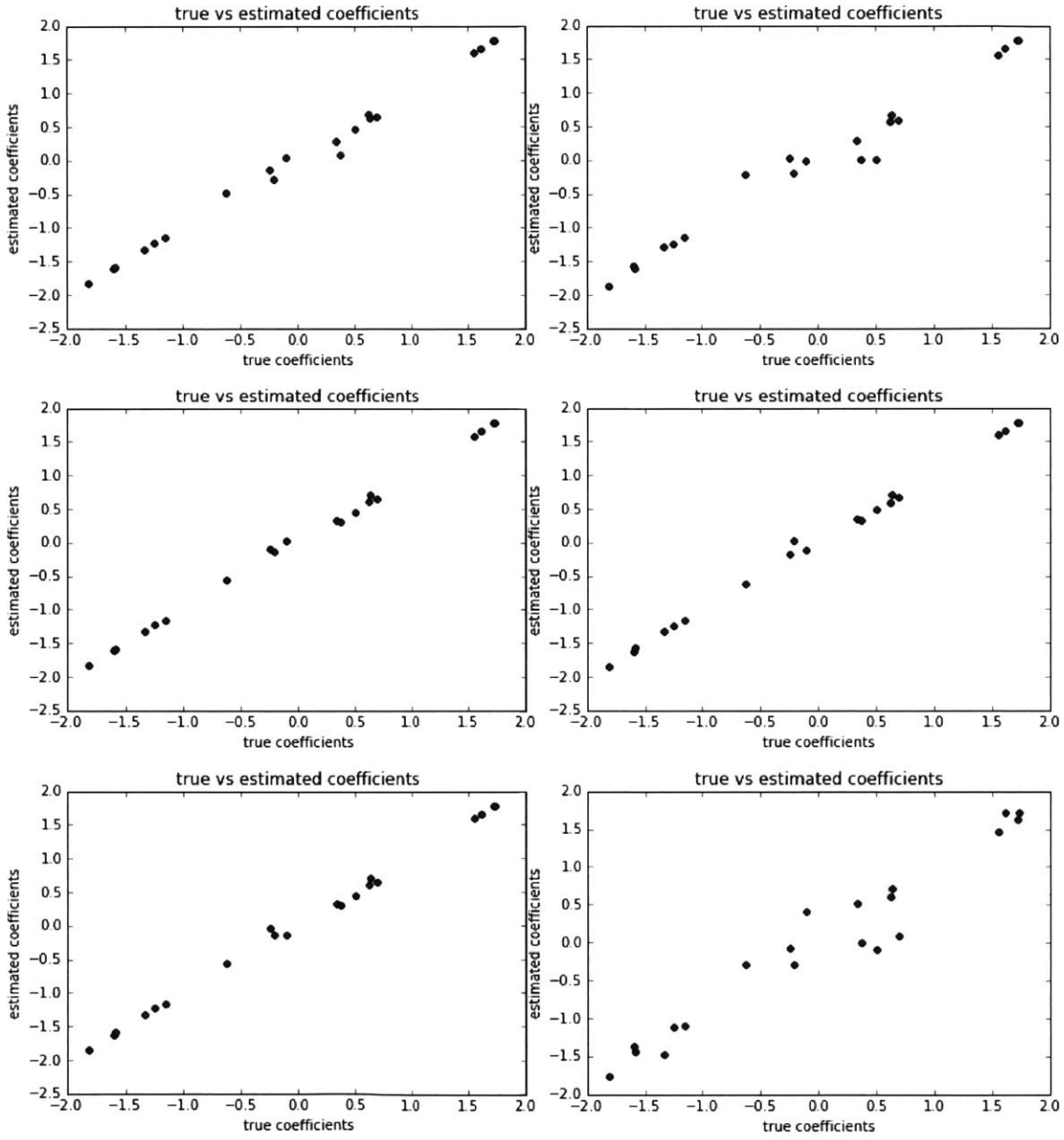


Figure 2-9: β vs estimated $\hat{\beta}$, computed with multiple random initializations. 20 attributes, 40 observations per consumers. β, μ, and the synthetic data were the same in all random initializations.

to the true ones, may still be predictively useful, and the gap in predictive performance between the estimated model and the true model gives a sense of how 'close' the estimated model is to the truth. Thus, evaluating predictive performance can as another measure of how well the estimation procedure works when the direct measures in Section 2.4 are not so easy to interpret.

Thus, this section focuses on a synthetic prediction task. The usefulness of the preferences estimated via Algorithm 7 for predicting consumer choice behavior is evaluated by comparison to the theoretical upper bound, as well as by comparison with more reduced-form methods. The experimental setup is described in Section 2.5.1, the set of estimation procedures used is described in Section 2.5.2, and the results in Section 2.5.3. The results indicate that the estimated parameter values are highly useful for prediction, coming close to the theoretical upper bound and significantly outperforming the tested reduced-form methods in all experimental trials run here. These results hold even in the cases with few observations per consumer, where consistently estimating the true model parameters is hard.

2.5.1 Experimental setup

The synthetic experiments were performed as follows:

1. Initialize number of consumers N, number of observations per consumer J, number of attributes K, and number of trials T.

2. For each trial:

   (a) Randomly initialize true preferences $\beta \sim \mathcal{N}(0, I_K)$, and mean cognitive cost parameter $\mu \sim \mathcal{U}[.8, 1.2]$.

   (b) Given β and mu, initialize a market with N consumers, $J + 100$ goods, and K attributes as in Definition 10.

   (c) Use the choice data of each consumer for each of the first J goods train a variety of prediction methods (which are described in Section 2.5.2), keeping the remaining 100 good as a hold-out test set.

   (d) Use each of the prediction methods to predict whether each consumer will choose to consume each of these 100 test-set goods.

   (e) Compare the predictions of each algorithm with the generated choice data on the test set, average this accuracy across all consumers and all goods to

generate a single number in $[0, 1]$, capturing the average accuracy of each algorithm on the test set.

3. This generates a list of T accuracy values for each algorithm.

For the experiments in this section, T was uniformly set to 40, N was uniformly set to 100, and J, K took on a variety of values listed below. The values of K and J here were chosen to reflect situations with few observations per consumer relative to attributes vs situations with more observations per consumer relative to attributes. As the previous section demonstrated that consistent estimation of preferences depends crucially on the number of observations per consumer, these choices of of K, J allow comparison of predictive performance in situations where preferences can be consistently estimated vs situations in which they cannot.

- $K = 20, J = 10$,: few attributes, few observations per consumer
- $K = 20, J = 20$: few attributes, moderate observations per consumer
- $K = 20, J = 40$: few attributes, many observations per consumer
- $K = 100, J = 50$: many attributes, few observations per consumer
- $K = 100, J = 100$: many attributes, moderate observations per consumer
- $K = 100, J = 200$: many attributes, many observations per consumer

2.5.2 Prediction methods

The experimental setup detailed above is applied to ten predictive methods, described in detail below. Method 9, the 'Estimated bayesian' method is the predictive model based on the estimation procedure described in Algorithm 7, and as such it is the main method of interest in this section. This method involves estimating $\hat{\beta}$, $\hat{\mu}$ via Algorithm 7, and then using these estimated parameters to predict consumers' choice behavior on the test data.

Method 10, the 'Optimal bayesian' method constitutes a theoretical upper bound on prediction performance. This method utilizes the same basic procedure as the Estimated bayesian method, but instead of using the estimated parameters $\hat{\beta}$ and $\hat{\mu}$, it uses the true latent parameters β and μ. Thus, this method is infeasible in practice, as it requires knowledge of the true parameter values. As it incorporates all relevant

information about the true structural model, it serves as an upper-bound comparison for all other prediction methods.

Methods 1-8 are reduced form methods that serve as a lower-bound comparisons for method 9. These models are standard methods often applied to prediction problems in practice. If any of the methods 1-8 consistently outperform method 9 in prediction, then there wouldn't be any purpose for trying to estimate the true parameter values for prediction; one could much more easily just apply one of these reduced form prediction methods and get better predictive performance. In addition, given that method 9 incorporates structural assumptions that are not incorporated in any of the methods 1-8, it utilizes information not available to the other methods, and thus should also be expected to perform better. Methods 1-8 were implemented in Python from the SKLearn package Buitinck et al. (2013).

In the description of the methods below, $X_j \in \mathbb{R}^K$ refers to a vector of attributes for good j, $Y_{ij} \in \{0, 1\}$ refers to the consumption choice of consumer i for good j, and $j = 1, ..., J$ refers to the observations that were used for training each of the prediction algorithms. The prediction methods:

1. Aggregate majority: extremely simple prediction method, where data from all consumers is combined into a single data set, and if there are more instances of consumers consuming a good as opposed to not consuming it, predict that all consumers will consume all goods. If not, predict that all consumers will consume no goods:

$$\text{consumer } i \text{ consumes good } j \iff \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} Y_{ij} > .5$$

2. Aggregate logit: data from all consumers is combined into a single data set, and a single logistic regression relating the attributes of each good to the choice outcome is run, with estimated parameters being used for prediction:

$$\text{consumer } i \text{ consumes good } j \iff u_j > .5, \quad u_j = \hat{\beta} \cdot X_j$$

$$\hat{\beta} = \arg\max_{\beta} \prod_{i=1}^{N} \prod_{j=1}^{J} \left(\frac{\exp(\beta \cdot X_j)}{1 + \exp(\beta \cdot X_j)} \right)^{Y_{ij}} \left(\frac{1}{1 + \exp(\beta \cdot X_j)} \right)^{1 - Y_{ij}}$$

3. Aggregate lasso logit: same as aggregate logit, except with an L_1 regularization

term used to induce sparsity in estimating $\hat{\beta}$. The regularization term η is chosen via 5-fold cross-validation.

$$\text{consumer } i \text{ consumes good } j \iff u_j > .5, \quad u_j = \hat{\beta} \cdot X_j$$

$$\hat{\beta} = \arg\min_{\beta} \left[-\prod_{i=1}^{N} \prod_{j=1}^{J} \left(\frac{\exp(\beta \cdot X_j)}{1 + \exp(\beta \cdot X_j)} \right)^{Y_{ij}} \left(\frac{1}{1 + \exp(\beta \cdot X_j)} \right)^{1-Y_{ij}} + \eta ||\beta||_1 \right]$$

4. Aggregate ridge logit: same as aggregate logit, except with an L_2 regularization term used to account for overfitting in estimating $\hat{\beta}$. The regularization term η is chosen via 5-fold cross-validation.

$$\text{consumer } i \text{ consumes good } j \iff u_j > .5, \quad u_j = \hat{\beta} \cdot X_j$$

$$\hat{\beta} = \arg\min_{\beta} \left[-\prod_{i=1}^{N} \prod_{j=1}^{J} \left(\frac{\exp(\beta \cdot X_j)}{1 + \exp(\beta \cdot X_j)} \right)^{Y_{ij}} \left(\frac{1}{1 + \exp(\beta \cdot X_j)} \right)^{1-Y_{ij}} + \eta ||\beta||_2 \right]$$

5. Individual majority: like the aggregate majority, except the majority vote is taken on an individual level now. For each consumer, if on the training set he consumes more often than note, then predict that he always consumes, and vice versa.

$$\text{consumer } i \text{ consumes good } j \iff \frac{1}{J} \sum_{j=1}^{J} Y_{ij} > .5$$

6. Individual logit: like the aggregate logit, but without aggregating the data across consumers. A different logistic regression relating the attributes of each good to the choice outcome is run for each consumer, with estimated parameters being used for predicting choice behavior for that user only:

$$\text{consumer } i \text{ consumes good } j \iff u_{ij} > .5, \quad u_{ij} = \hat{\beta}_i \cdot X_j$$

$$\hat{\beta}_i = \arg\max_{\beta} \prod_{j=1}^{J} \left(\frac{\exp(\beta \cdot X_j)}{1 + \exp(\beta \cdot X_j)} \right)^{Y_{ij}} \left(\frac{1}{1 + \exp(\beta \cdot X_j)} \right)^{1-Y_{ij}}$$

7. Individual lasso logit: same as individual logit, except with an L_1 regularization term used to induce sparsity in estimating $\hat{\beta}_i$. The regularization term η is chosen

via 5-fold cross-validation.

$$\text{consumer } i \text{ consumes good } j \iff u_j > .5, \quad u_j = \hat{\beta} \cdot X_j$$

$$\hat{\beta}_i = \arg\min_{\beta} \left[-\prod_{j=1}^{J} \left(\frac{\exp(\beta \cdot X_j)}{1 + \exp(\beta \cdot X_j)} \right)^{Y_{ij}} \left(\frac{1}{1 + \exp(\beta \cdot X_j)} \right)^{1 - Y_{ij}} + \eta ||\beta||_1 \right]$$

8. Individual ridge logit: same as individual logit, except with an L_2 regularization term used control overfitting in estimating $\hat{\beta}_i$. The regularization term η is chosen via 5-fold cross-validation.

$$\text{consumer } i \text{ consumes good } j \iff u_j > .5, \quad u_j = \hat{\beta} \cdot X_j$$

$$\hat{\beta}_i = \arg\min_{\beta} \left[-\prod_{j=1}^{J} \left(\frac{\exp(\beta \cdot X_j)}{1 + \exp(\beta \cdot X_j)} \right)^{Y_{ij}} \left(\frac{1}{1 + \exp(\beta \cdot X_j)} \right)^{1 - Y_{ij}} + \eta ||\beta||_2 \right]$$

9. Estimated bayesian: this is the main prediction method of interest. It involves running the maximum likelihood estimation procedure in Algorithm 7 to get the estimated $\hat{\beta}$ and $\hat{\mu}$, and then using this to compute the bayesian posterior probability that a consumer will choose to consume a given good. Formally, this is done as follows:

   - For each potential active attribute set s (defined as in 2.1), compute its probability given $\hat{\beta}$ and $\hat{\mu}$. Denote this ρ_s (see Equation 2.3 for the precise formula).
   - Compute the likelihood, given the data that consumer i would have made his choices $(Y_{ij})_{j=1}^{J}$ if his active set were s. Denote this by $l_i^s = \prod_{j=1}^{J} l_{ij}^s$ (see Equation 2.3 for the precise formula).
   - Compute the posterior probability, given the observed choice data, that each consumer has each active set s:

$$\frac{\rho_s l_i^s}{\sum_{s'} \rho_{s'} l_i^{s'}}$$

   - For each good j in the test set, compute the probability that a consumer

with active attribute set s will choose to consume j:

$$\frac{\exp(\hat{\beta} \cdot X_j \odot s)}{1 + \exp(\hat{\beta} \cdot X_j \odot s)}$$

- For each good j in the test set, compute the expected probability that consumer i will choose j by taking a weighted sum of the product of the two quantities computed above (probability that a consumer has an active set, probability of choice given active set). Then, predict that the consumer will choose to consumer the good IFF this probability exceeds 1/2:

$$\text{consumer } i \text{ consumes good } j \iff \sum_s \left(\frac{\rho_s l_i^s}{\sum_{s'} \rho_{s'} l_i^{s'}} \right) \left(\frac{\exp(\hat{\beta} \cdot X_j \odot s)}{1 + \exp(\hat{\beta} \cdot X_j \odot s)} \right) > .5$$

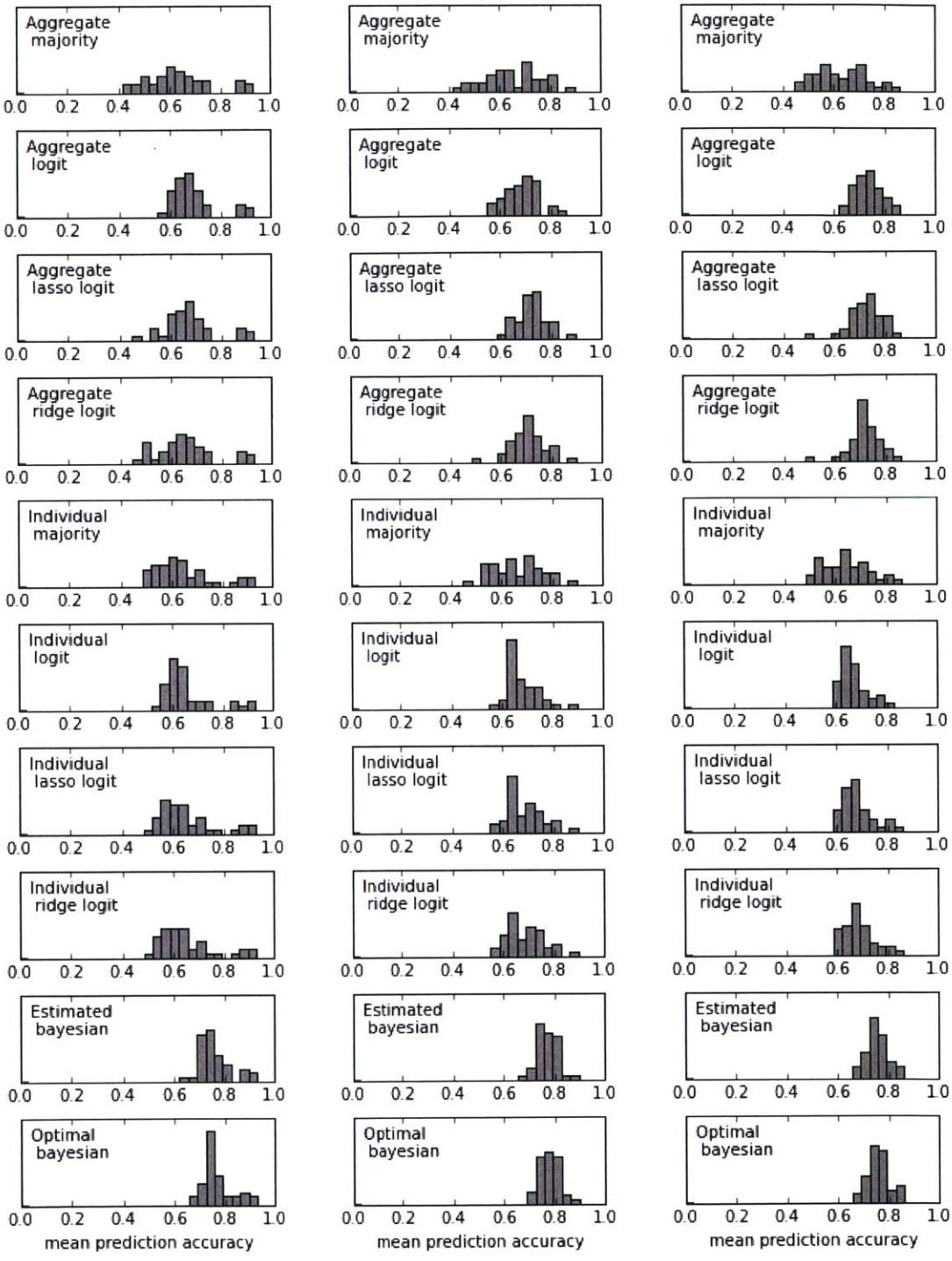
10. Optimal bayesian: identical to the Estimated bayesian method, but with the true latent parameters β, μ replacing the estimated ones $\hat{\beta}, \hat{\mu}$. Note that ρ_s and l_i^s depend on β, μ.

2.5.3 Results

The prediction results indicate that the estimated parameters are quite useful for prediction: the 'Estimated bayesian' method (method 9) outperforms the reduced form methods 1-8, and achieves performance close to that of the theoretical upper bound given by the 'Optimal bayesian' method (method 10).

Figure 2-10 contains histograms of prediction performance for simulations with 20 attributes and various numbers of observations per consumer. Prediction performance is between 0 and 1, with higher values corresponding to more accurate predictions. Each observation corresponds to a single simulation run, with its own randomly generated true model parameters β and μ.

As indicated in the second to last row of Figure 2-10, the Estimated bayesian method has predictive performance very close to the Optimal bayesian method, indicating that the estimated parameters do a nearly-optimal job of predicting consumer choice behavior. Somewhat surprisingly, this is the case irrespective of the number of observations per consumer. When the number of observations per consumer is large (e.g. in the rightmost panel of the figure, with 40 observations per consumer and 20 attributes), this is quite sensible; it was noted in the previous section that in this case, where ob-



(a) 10 observations per consumer

(b) 20 observations per consumer

(c) 40 observations per consumer

Figure 2-10: Histograms of prediction performance of various methods on hold-out data. 20 attributes, 100 consumers. 40 observation, each being a separate simulation run.

servations per consumer are plentiful, it is possible to very precisely estimate β and μ, so that $\hat{\beta}$ and $\hat{\mu}$ are very close to the true β and μ (see e.g.Figure 2-9). As a result, it follows that prediction performance between the Estimated bayesian and Optimal bayesian methods should be quite similar.

It's somewhat more surprising that even in situations where the number of observations per consumer is low or medium that prediction performance of the Estimated bayesian method remains close to that of the Optimal bayesian method. In such cases, it was shown in the previous sections that the estimated $\hat{\beta}$ and $\hat{\mu}$ do not very closely approximate the true β and μ (see e.g. Figure 2-7). However, in spite of this fact, there doesn't appear to be any noticeable drop in prediction performance. Thus, it appears that the discrepancy between the estimated and real parameters doesn't particularly matter for prediction; this suggests that, when the number of observations per consumer is small, the likelihood function exhibits multiple optima of quality comparable to the true parameter values.

Comparing the prediction performance of the Estimated bayesian method (row 9) with the reduced form prediction methods (rows 1-8), we see that across the board, the Estimated bayesian method substantially out-performs the reduced form methods. Since the synthetic data used in these experiments is generated via a structural model, and the Estimated bayesian method incorporates some knowledge of this during estimation whereas the reduced form prediction methods do not, the better performance of the Estimated bayesian method is not entirely surprising. However, it is notable that the gap between the reduced-form methods and the Estimated bayesian method remains even when the number of observations per consumer is large (see Figure 2-10c). In such cases, with many observations per consumer, the individual-level reduced-form methods gain some advantage, as more observations allows better estimation of each consumer's preferences on an individual level, even without using population information. As can be seen in the figure, the gap between e.g. Individual ridge logit and Estimated bayesian (rows 8 and 9) decreases from left to right in the figure. However, in the rightmost panel, this gap still persists. This indicates that even when the number of observations per consumer is fairly large relative to the number of attributes, incorporating knowledge of population structure into the estimation procedure (as is done in the Estimated bayesian method) still significantly improves prediction performance.

Figure 2-11 contains histograms of prediction performance for simulations with 100 attributes and various numbers of observations per consumer. The results are broadly similar to the results in 2-10, indicating that even with a much larger number of attrib-

utes, the Estimated bayesian procedure produces results that are close to optimal, and significantly better than the reduced-form prediction methods.

2.6 Empirical application

In order to complement the synthetic experiments performed in the previous sections, this section estimates the model in Definition 10 using experimental data on consumer consideration behavior. The estimated parameters give a measure of the distribution of cognitive costs within the population of consumers; in particular, the estimated values indicate that the average consumer uses simple preferences that only utilize 6 out of the 16 available attributes, ignoring the rest.

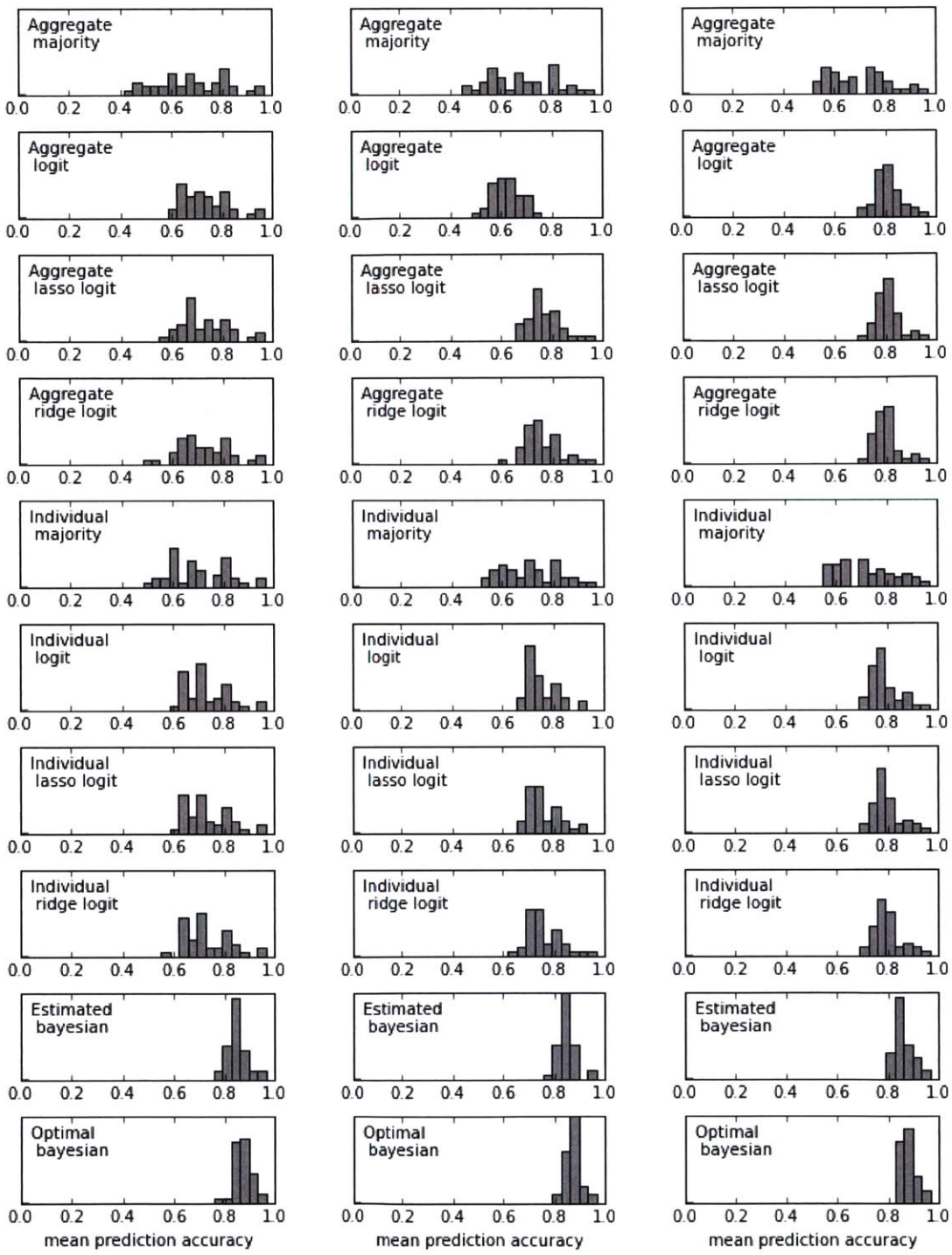
Furthermore, the fact that consumers make choices using simple preferences instead of the true latent preferences implies that the choices might be suboptimal. Under some assumptions, the distribution of cognitive costs can thus be translated into a distribution of utility loss from using simple preferences instead of the true latent preferences. Performing this analysis yields an an average value of around $7.8 for the utility loss by consumers.

2.6.1 Data

The data analyzed here comes from Hauser et al. (2010), in which the authors conduct an experimental study of consumer consideration behavior. The experiment, which was conducted through a website, involved respondents being presented with a number of global positioning system (GPS) units, and then being tasked with choosing which ones to put into a consideration set. Each respondent has a chance of randomly winning either cash prize or one of the GPS units in their consideration set, with the rewards structured in such a way as to induce consumers to answer truthfully. A more thorough description of this experimental study may be found in Hauser et al. (2010). Some key details are described below in Definition 11.

Definition 11. *Summary of experimental setting of Hauser et al. (2010):*

1. *A respondent is repeatedly presented with products, in random order. The products are presented on a computer screen, along with information on the various attributes of the product.*



(a) 50 observations per consumer

(b) 100 observations per consumer

(c) 200 observations per consumer

Figure 2-11: Histograms of prediction performance of various methods on hold-out data. 100 attributes, 100 consumers. 40 observation, each being a separate simulation run.

2. *For each product, the respondent has the option of putting it into their consideration set or not.*

3. *Once all products have been presented once, the respondent is then shown their entire consideration set, as well as all of products not in the consideration set.*

4. *At this point, the respondent is free to revisit any of these products and move any of them into or out of their consideration set.*

5. *The respondent finalizes and submits their consideration set.*

6. *The consumers now have some chance to win a prize, which is a random lottery over cash and/or a GPS unit. The particular structure of the lottery is constructed to make truth-telling incentive compatible.*

In total, each consumer faced 64 different products (GPS units) across two separate experimental trials. That is, the procedure of Definition 11 was repeated twice for each respondent, each time with 32 distinct products. The same set of 32 products was used for all respondents in each of the two trials, and no products were repeated across the two repetitions (thus yielding 64 distinct products). The products in each of the two repetitions were constructed in such a way so that all the attributes (excepting price) were orthogonal to each other. Some summary statistics are presented in Table 2.1.

Each product was defined by 16 attributes: price, which took on four possible values ($\$249, \$299, \$349, \$399,$), and the remaining 15 attributes, which were binary. These binary attributes are mostly self-explanatory, e.g. large size=1 indicates that the GPS is large, whereas large size=0 implies that the GPS is small. For the most part, the 'better' value of each attribute is assigned to the 1 value.

The data from this experimental study is well-suited for the model described in Definition 10, as almost all the attributes here appear to have unambiguous sign (excepting brand). For example, it's clear that a higher price should be worse, whereas having higher battery life or a larger display should be better, and so forth. Thus, ex-ante, this setting looks quite close to one in which there is a single 'true' measure of quality that serve as a basis for the consumers' simplified preferences that drive consideration behavior.

Experimental consideration data from a total of 645 respondents was collected by Hauser et al. (2010) via the procedure detailed above. This data includes, for each of the 645×64 (respondent, product) pairs, whether that respondent considered that

Table 2.1: Summary statistics for the 64 GPS units. All attributes except price are equally likely to be 0 or 1, and are all mutually orthogonal.

Statistic	N	Mean	St. Dev.	Min	Max
price	64	324.000	57.044	249	399
Magellan brand	64	0.500	0.504	0	1
large size	64	0.500	0.504	0	1
heavy weight	64	0.500	0.504	0	1
color display	64	0.500	0.504	0	1
bright display	64	0.500	0.504	0	1
large display	64	0.500	0.504	0	1
high-resolution display	64	0.500	0.504	0	1
fast acquisition time	64	0.500	0.504	0	1
long battery life	64	0.500	0.504	0	1
high sensitivity	64	0.500	0.504	0	1
high accuracy	64	0.500	0.504	0	1
track log	64	0.500	0.504	0	1
mini-usb port	64	0.500	0.504	0	1
backlit	64	0.500	0.504	0	1
floats on water	64	0.500	0.504	0	1

product. On average, a consumer considered 14.5 products out of the 64 possible, with a standard deviation of 10.2. The empirical distribution of the number of products considered across all 645 respondents is plotted in Figure 2-12.

2.6.2 Analysis and results

The consideration data described above almost be directly plugged into the model of Definition 10. To do so, two objects need to be defined: the matrix X containing information on the goods and attributes, and the matrix Y containing information on the respondent' consideration behavior.

- To construct X: each of the 16 attributes is taken, and divided by its standard deviation, so that each of the resulting adjusted attributes has standard deviation of 1. This results in a 64×16 matrix indicating the value of each attribute of each good.

- To construct Y: all the consideration data from both repetitions of the experimental study is compiled into a single data set, giving a 645×64 matrix taking

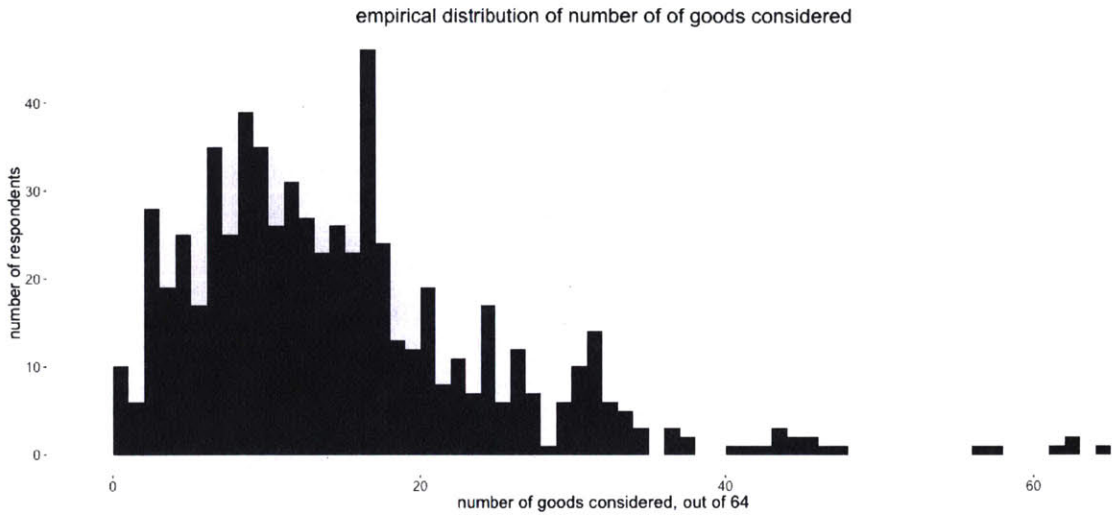


Figure 2-12: Histogram of the sizes of consideration sets of the 645 respondents. Maximum possible size is 64.

values in $\{0, 1\}$ that indicates whether or not each respondent considered each product.

Given this, it is straightforward to produce estimates of the latent preference parameters β and the mean cognitive cost parameter μ via Algorithm 7. Confidence intervals are produced via bootstrap, where 645 random respondents are chosen from the existing pool with replacement, and the estimation procedure run with the bootstrapped sample. 400 bootstrap draws were used. The results are listed in Table 2.2.

The estimates of the latent preference parameters β are largely as expected. Note that, as the attributes all have their standard deviations normalized to one, the sizes of the coefficients are comparable. Price is the most important attribute, exerting a large negative impact on the latent utility of a product. The signs of the remaining attributes also largely conform to expectations, e.g. color displays are better than black and white ones, larger displays are better than smaller ones, etc. It is notable, however, that consumers appear to prefer heavier GPS units over light ones, which might not have been obvious ex-ante. Most of these coefficients are also significantly nonzero, excepting the coefficients on brand and backlight.

The estimate of the mean cognitive cost parameter μ comes out to around 0.3, which is larger than the coefficients on 10 out of the 16 attributes. The preferences simplification procedure described Equation 2.1 involves a consumer simply ignoring all attributes except those with coefficients exceeding their individual cognitive cost

Table 2.2: Estimates of the latent preference parameters β and the mean cognitive cost parameter μ. Estimates were computed via Algorithm 7. Confidence intervals computed via bootstrap. Most attributes are significant and positive, excepting price. The mean cognitive cost parameter is larger than 10 of the 16 coefficients, so that the average respondent uses a simplified model with only 6 nonzero coefficients.

VARIABLE	ESTIMATE	95% CI
price	-0.766*	(-0.831, -0.689)
bright display	0.559*	(0.498, 0.618)
color display	0.446*	(0.364, 0.509)
mini-usb port	0.415*	(0.358, 0.483)
large display	0.392*	(0.35, 0.442)
long battery life	0.315*	(0.259, 0.386)
heavy weight	0.238*	(0.159, 0.276)
large size	0.202*	(0.126, 0.246)
high accuracy	0.191*	(0.124, 0.235)
track log	0.163*	(0.125, 0.208)
high sensitivity	0.151*	(0.122, 0.212)
fast acquisition time	0.118*	(0.093, 0.157)
high-resolution display	0.103*	(0.084, 0.161)
floats on water	0.088*	(0.064, 0.133)
Magellan brand	0.052	(-0.098, 0.081)
backlit	0.034	(-0.106, 0.065)
mean cognitive cost	0.299*	(0.271, 0.334)
	* $p<0.05$	

parameter. This implies that the average respondent, which has cognitive cost parameter around 0.3, would have simple preferences in which only the 6 top attributes enter (namely: price, bright display, color display, min-usb port, large display, long battery life). The full distribution of the sizes of simplified preferences is plotted in Figure 2-13. At the estimated values of β and μ, this distribution appears to be spread out across a wide range of model sizes, so that there are respondents whose simple preferences use zero out of the 16 attributes, and also respondents whose simple preferences use all 16 attributes. However, these probabilities are not very precisely estimated.

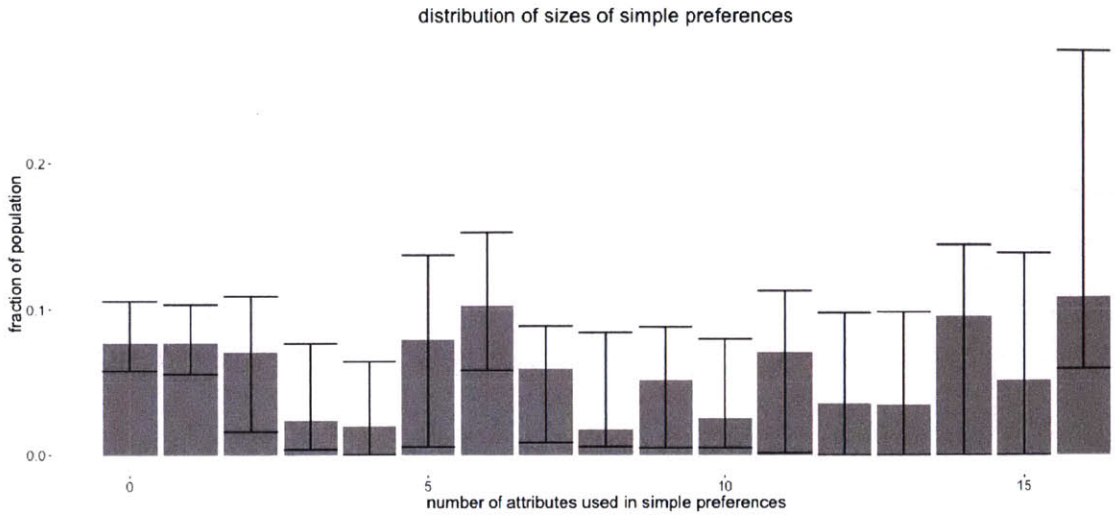


Figure 2-13: Distribution of simple preference sizes. Error bars are bootstrap 95% confidence intervals. There is a significant amount of variance in how much consumers simplify their preferences. About 7% of respondents have simple preferences with no attributes, and about 10% have simple preferences that were no different than the full latent preferences.

2.6.3 Monetary cost of simple preferences

In order to give a more intuitive sense of how significant the cognitive costs estimated above are, this section attempts to convert these cognitive costs into dollar terms. This can be done by noticing that, if consumers use simple preferences instead of the true latent preferences for choosing products, then the chosen products might be suboptimal, which thus implies some amount of utility loss, which can then be translated into dollar amounts.

As the amount of utility loss from using simple preferences depends on the particular situation that the consumers are faced with, it is therefore necessary to make some assumptions about this. Thus, we stipulate a particular choice setting that both mirrors the structure of the experimental study from which the data was collected, as well as being a simple and quite natural setting which a consumer might encounter in real life. This assumption is described below in Definition 12.

Definition 12. *Counterfactual choice setting for evaluating cost of simple preferences:*

1. *A consumer is faced with the 64 products used in the experimental study.*
2. *Consumer i has 'true' latent utility for good j given by $\beta \cdot X_j + \varepsilon_j$, where X_j is*

the vector of attributes of good j, and ε_j is a Gumbel shock. The outside good has utility ε_0, which is again a Gumbel shock.

3. *Consumer i uses some simple preferences, denoted by $\tilde{\beta}$, to make consideration decisions. In particular, consumer i perceives the utility of good j to be $\tilde{\beta} \cdot X_j + \varepsilon_j$ instead of the true latent utility $\beta \cdot X_j + \varepsilon_j$.*

4. *For each product j, consumer i puts good j into their consideration set IFF $\tilde{\beta} \cdot X_j + \varepsilon_j > \varepsilon_0$. However, the true utility of good j for consumer i is still $\beta \cdot X_j + \varepsilon_j$.*

5. *After finalizing their consideration set, one product out of the 64 is randomly selected with equal probability. If that product is in the consumer's consideration set, then the consumer must purchase the good. If the product is not in the consideration set, then the consumer must not purchase the good.*

This setting has several advantages that make it well suited for a first-pass attempt in counterfactually assessing the extent of utility loss from simple preferences:

- It is structured in a similar way as the experimental study from which the data used here is taken, so that it may be plausibly argued that in this setting, consumers make consideration decisions by use the same simple preferences as in the experimental setting.

- The reward structure is extremely simple, while also being plausibly close to a setting that a consumer might encounter in daily life. Thus, it is possible in this setting to get a plausible numerical value for the utility loss without having to make too many modeling choices.

Given the setting of Definition 12, it follows that a consumer i with consideration set C will receive true expected utility given by

$$u(C) := (1/64) \sum_{j=1}^{64} (\mathbb{I}[j \in C](\beta \cdot X_j + \varepsilon_j) + (1 - \mathbb{I}[j \in C])(\varepsilon_0)) \tag{2.6}$$

It is thus apparent that, absent any cognitive costs, it is optimal for consumer i to put a product j into their consideration set IFF $u_{ij} > u_{i0}$. This is exactly what a consumer whose simple preferences $\tilde{\beta}$ coincide exactly with the true latent preferences β would do. Thus, absent any cognitive constraints, it is optimal for consumers to use the full latent preferences for decision making. For consumers whose simple preferences

$\tilde{\beta}$ involve some nontrivial amount of simplification (i.e. some coefficients are set to zero), the simple utilities of products will not coincide with the true latent ones, and therefore the consideration set of such a consumer will be suboptimal. The extent of this suboptimality can then be measured and then put into money terms. Given the estimates of β and μ above, we can estimate a distribution function for the amount of monetary loss. The procedure is as follows:

1. Compute the consideration set C^* generated by the latent preferences $\hat{\beta}$
2. Compute the utility generated by C^* as defined by Equation 2.6, call this U^*.
3. Given $\hat{\beta}$ and $\hat{\mu}$, compute all the various possible simple preferences ($\tilde{\beta}$), and the probability for a consumer's simple preference to taken on each of these various possibilities. As each product has 16 attributes, this leads to 17 different possible simple preferences (ranging from simple preferences that are uniformly zero to simple preferences identical to the latent preferences).
4. For each of these 17 simple preferences $\tilde{\beta}$:
   (a) Compute the consideration set $C_{\tilde{\beta}}$ generated by the simple preferences.
   (b) Compute the expected utility from using $C_{\tilde{\beta}}$, as defined by Equation 2.6. Call this $U_{\tilde{\beta}}$.
   (c) Quantify the utility loss as $U_{\tilde{\beta}} - U^*$.
   (d) Put this in dollar amounts by computing

$$\text{utility loss in USD} := \frac{U_{\tilde{\beta}} - U^*}{\hat{\beta}_{\$}}$$

   where $\hat{\beta}_{\$}$ is the coefficient on the price term of the estimated latent preferences.
5. Taking the utility loss in USD associated with each possible realization of simple preferences, together with the probabilities, it is then possible to construct a distribution of utility losses across the population of consumers.

Figure 2-14 plots this distribution of USD utility losses, along with bootstrap confidence intervals. The mean and several quantiles of this distribution are listed in Table

2.3, again with bootstrap confidence intervals. The mean utility loss from simple preferences comes to about \$7.8. That is, in the particular choice scenario described in Definition 12, if consumers behaved according to the model estimated estimated here, then on average the utility loss experienced by consumers would be around 7.8 dollars.

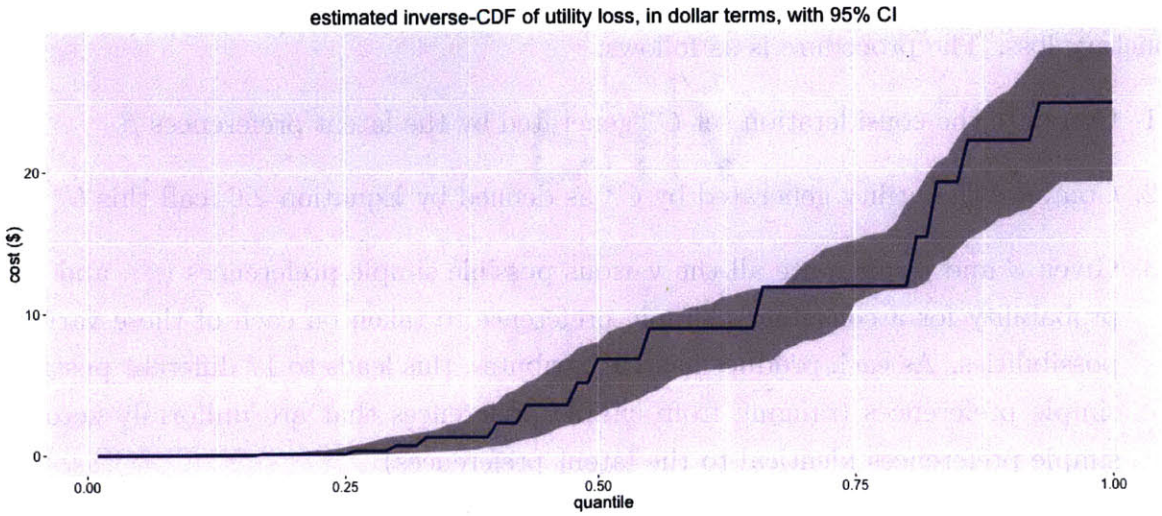


Figure 2-14: Inverse-cdf of the distribution of utility loss from using simple preferences to construct consideration sets. There are many consumers who essentially suffer no loss and some who suffer quite a significant loss. For reference, the prices of the GPS units ranged from \$249 to \$399.

Table 2.3: Some statistics of the distribution of utility loss from using simple preferences. Confidence intervals computed via bootstrap. For example, on average, consumers using simplified preferences for constructing consideration sets results in a loss of about \$7.8.

VARIABLE	ESTIMATE	95% CI
mean of utility loss in USD	7.834	(6.165 , 8.812)
.25 quantile of utility loss in USD	0.084	(0 , 0.582)
.5 quantile of utility loss in USD	6.717	(2.758 , 7.95)
.75 quantile of utility loss in USD	11.826	(10.649 , 15.033)

The estimates of utility loss are not particularly large, around $2-3\%$ relative to the price of the GPS units, which were priced between \$249 and \$399. However, it should be noted that the data used to generate these numbers were collected via a pre-screening process in which respondents were vetted for interest in purchasing GPS units. Thus,

the estimates here are plausible given that the respondents were ex-ante expected to be somewhat sophisticated. In light of this, these numbers seem fairly believable.

Two caveats should be made about the external validity of these dollar figures:

1. The dollar values of the utility loss apply to the the particular counterfactual setting (in Definition 12), which may or may not be a good model for the consumers' choice problem in real life.

2. It was assumed that in such a counterfactual setting, the respondents would use the same models estimated from experimental data to generate consideration sets.

The first assumption can be relaxed by simply considering different choice settings. For example, an alternative setting might suppose that instead of randomly choosing a single product to serve as a prize, a random subset of the 64 products is 'available' for purchase, and then a consumer may choose among the goods that are both available and in their consideration set. Such a specification might look closer to a real-life choice scenario, and could provide interesting starting points for future work. We don't perform such analyses here due to the additional complexity of specifying the probability of availability.

The second assumption, that consumers use the same simple preferences in both the experimental setting as well as in the counterfactual setting, makes it so that it's valid to simply take the estimates of β and μ from experimental data and use them to perform the analysis of utility loss. This assumption might not hold in practice, as it could be the case that consumers construct different simple preferences depending on the stakes at hand. In the context of the baseline model of Definition 10, this may be interpreted as saying that a consumer i might have different values of λ^i in different situation. Thus, if the stakes were high, the λ^i might be low, and vice versa. If this were the case, then the values of the utility loss here are likely overstated: in the experimental setting, there was only a small chance that a respondent would win a prize, and thus a small chance that the constructed consideration set would matter, so that the payoffs from having fully fleshed out preferences would be small, so that λ^i would be large; conversely, if the respondent knew they were going to go out and purchase a GPS unit, constructing a better consideration set would be more important, and thus λ^i might be lower. Thus, the monetary values of the utility losses computed above might be best understood as an upper bound. However, due to the fact that the respondents here were pre-screened for interest in purchasing GPS units, it is likely that

the respondents have already constructed some simple preferences for use in evaluating GPS units. Thus, it is plausible in this setting that they would use the same simple preferences in both experimental and real-life choice settings, which suggests that the dollar figures computed above might not be too high.

2.7 General preference simplification

The preference simplification procedure described in Equation 2.1 may seem a bit ad hoc, but it can actually be viewed as a particular form of a more general procedure in which a consumer simplifies his true latent preferences into some sparse preferences that is somehow 'closest' to his latent preferences given some cognitive limitations. We describe that more general procedure and provide some motivation for it in this section.

2.7.1 Procedure

Consider a consumer choosing products to purchase out of a set of products. Formally, let the products be defined by their attributes, which we take to be a K-dimensional space, so that a good j can be identified with a vector $X_j \in \mathbb{R}^K$. Then, define consumer i as having 'true' latent preferences $\beta_i \in \mathbb{R}^K$. This latent utility is then turned into a set of sparse preferences $\tilde{\beta}_i \in \mathbb{R}^K$ that are actually used in deciding which good to choose via a post-lasso. Formally, the simplification procedure is as follows:

Procedure 4. *Preference simplification: given latent preferences β_i, cost-of-cognition parameter $\lambda_i \in \mathbb{R}_+$, and background probability distribution over products $\nu \in \Delta(\mathbb{R}^K)$, generate simplified preferences $\tilde{\beta}_i$ via a post-lasso procedure:*

1. *Generate a set of active attributes via a lasso least squares regression:*

$$s := \{k \in 1, ..., K \ \ s.t. \ \ \gamma_k^* \neq 0\}, \quad \gamma^* := \arg\min_{\gamma} \int_{\mathbb{R}^K} (x \cdot \beta_i - x \cdot \gamma)^2 d\nu(x) + \lambda_i ||\gamma||_1 \tag{2.7}$$

   where $||\gamma||_1 = \sum_{k=1}^{K} |\gamma_k|$ is the L^1 norm of γ.

2. *Generate simplified preferences $\tilde{\beta}_i$ by ordinary least squares using only the set of attributes contained in s:*

$$\tilde{\beta}_i := \arg\min_{\gamma} \frac{1}{2} \int_{\mathbb{R}_s^K} (x \cdot \beta_i - x \cdot \gamma)^2 d\nu(x) \tag{2.8}$$

where $\mathbb{R}_s^K$ refers to the subspace of $\mathbb{R}^K$ where the dimensions $k \notin s$ are set to zero.

Intuitively, the background distribution over products ν represents the distribution over products in the market, and the cost-of-cognition parameter λ_i controls the size of the set of active attributes. The consumer then optimizes his simple preferences $\tilde{\beta}_i$ to be in some sense the 'closest' approximation to his true preferences β_i given the background product distribution and his cost-of-cognition parameter.

These simplified preferences are then used by the consumer to choose which good (if any) to consume out of a set via the standard logit choice formalism:

Procedure 5. *Consumer choice: given simple preferences $\tilde{\beta}_i$, and a finite set of products $C \subset \mathbb{R}^K$, then consumer i chooses product j in C with probability*

$$P(i \text{ chooses } j \text{ from } C) = \frac{\exp(X_j \cdot \tilde{\beta}_i)}{1 + \sum_{j' \in C} \exp(X_{j'} \cdot \tilde{\beta}_i)}$$

$$P(i \text{ chooses nothing from } C) = \frac{1}{1 + \sum_{j' \in C} \exp(X_{j'} \cdot \tilde{\beta}_i)}$$

The procedure described in Equation 2.1 corresponds to a case where the background distribution over products ν is generated by orthonormal attributes so that $\int_{\mathbb{R}^K} xx' d\nu(x) = I_K$. In this case, the post-lasso variable selection procedure selects only the attributes with $|\beta_k| > \lambda_i$, exactly as in Equation 2.1 (this result is standard, see e.g. Page 71 of Friedman et al. (2008)).

2.7.2 Motivation

The model simplification procedure described above constitutes a departure from the standard discrete choice model. This departure is motivated some shortcomings of the standard discrete choice model in situations where the dimensionality of the attribute is high. This subsection discusses in detail the motivation for this particular modeling choice.

In the standard discrete choice framework, the consumer has some utility for each product, and the probability of choice is determined via some monotonic transform of these utilities. Furthermore, one way characterize the set of products by embedding them in some attribute space, so that each product may be represented by a set of attributes. In such a setting, the basic logit model of discrete choice characterizes the

consumer's preferences over each good as a linear function of these attributes, and then performs a logit transform of these utilities to generate the probabilities of choice.

Formally, if the products are embedded in a K-dimensional attribute space, so that each product j may be described by a vector $X_j \in \mathbb{R}^K$. Then, the standard additive logit model characterizes the subjective utility of a product j for a consumer i as $u_{ij} = X_j \cdot \tilde{\beta}_i$ for some $\tilde{\beta}_i \in \mathbb{R}^+$. Then, given a set of C of potential goods that the consumer can choose at most 1 out of, the standard logit model stipulates that the probability that consumer i chooses good j out of a set of alternatives C is exactly as given in Procedure 5.

In settings where the dimensionality of the attribute space is high, this leads to two issues, one practical and one related to interpretation. The practical issue is that it's hard to identify the preference vector $\tilde{\beta}_i$ from limited amounts of observed consumer choice data. The interpretation issue is that it's implausible that consumers are actually explicitly taking into account the large number of attributes and computing a utility, due to the cognitive cost of doing so given the large number of attributes, thus making it implausible to interpret $\tilde{\beta}_i$ as deep structural parameters.

To deal with these two issues, some authors (e.g. Gillen et al. (2014) among others) have adopted an assumption of sparsity, where only a small fraction of the entries in $\tilde{\beta}_i$ are nonzero. This solves the estimation problem by allowing estimation via some sort of regularized regression (e.g. Tibshirani (1996)). This sparsity assumption also deals with the interpretation issue, as if only a few entries of $\tilde{\beta}_i$ are nonzero, then this can be interpreted as the consumer deciding to take into account only a few attributes when making a choice.

However, this assumption of sparsity has some intuitive plausibility issues of its own. Most saliently, it's unclear why most attributes would be expected to have exactly zero impact on a consumer's utilities. It's very plausible that some attributes might have a small impact, but the precise zero is less plausible. Motivated by this issue, some authors (e.g. Gabaix (2014)) have interpreted the sparse $\tilde{\beta}_i$ as a simplified set of utility parameters derived from a non-sparse set of 'true' preferences β_i via some sort of optimization process.

Thus, a starting point for this modeling convention might be to assume that cognition is costly, with each nonzero attribute incurring some cognitive cost λ_i for the consumer. This can be interpreted as the literal psychological cost of remembering the coefficient of each nonzero attribute, assessing the value of that attribute for each good, and doing the additional multiplication when evaluation the total utility of the good.

Thus, a first attempt at modeling this cognitive cost might look something like this:

$$\tilde{\beta} = \arg\min_{\gamma} \int_{\mathbb{R}^K} (x \cdot \beta_i - x \cdot \gamma)^2 d\nu(x) + \lambda_i ||\gamma||_0$$

This has a clear interpretation: the consumer cares about how accurately his simple preferences mirror his real preferences, but also cares about the cognitive cost of having of having to remember many preference parameters, and as a results chooses the optimal set of attributes to remember to balance this trade off. The cost is simply the number of nonzero coefficients he has to remember.

However, this formulation has one major issue when interpreted as an assumption about the literal cognitive process the consumer performs: the optimization problem described is NP-hard, and thus not actually computationally feasible for any neural mechanism in the consumer's brain to literally implement this procedure. To circumvent this problem of computational tractability / cognitively plausibility while still maintain sparsity of $\tilde{\beta}$, Gabaix (2014) changed the L_0 regularization term to an L_1 term, and produced the following procedure:

$$\tilde{\beta} = \arg\min_{\gamma} \int_{\mathbb{R}^K} (x \cdot \beta_i - x \cdot \gamma)^2 d\nu(x) + \lambda_i ||\gamma||_1$$

That is, the consumer generates simple preference by solving a Lasso least square regression, which produces a sparse $\tilde{\beta}$ while still being computationally tractable. The post-lasso simplification procedure described in Procedure 4 is a direct extension to this lasso simplification formalism, where instead of performing just a lasso, the consumer selects a set of attributes to focus on via the Lasso, and then performs OLS using that set of attributes. This post-lasso procedure entails a trivial additional computational cost over the lasso (just the OLS procedure), while also being the best least-squares approximation for the true preferences β given the set of nonzero attributes.

A crucial aspect of the simplification process in Procedure 4 is the background product distribution $\nu \in \Delta\mathbb{R}^K$. This distribution is meant to capture the set of products that a consumer sees in the market. It forms the basis for which the constrained optimization occurs, as the consumer cares about how well his simple preferences match up with his real preferences on the goods that he encounters in the market. As such, this distribution should be taken as the literal distribution over products in the market.

2.8 Conclusion

This paper introduced a model of high-dimensional discrete choice that tractably incorporates both sparsity and population structure. In this model, each consumer has a set of 'true' latent high-dimensional preferences, but uses a lower-dimensional reduction of the high-dimensional latent preferences for decision making. This model can be microfounded as the outcome of an optimization procedure where a cognitively constrained consumer chooses an optimal simple model to use in decision making. Such a model can be applied to a wide variety of problems in economics and marketing, including high dimensional discrete choice, conjoint analysis, consideration behavior, and behavioral choice. A stochastic gradient descent based algorithm was introduced to estimate this model, and it was shown that this algorithm is able to consistently estimate the model parameters in situations where the number of observations per consumer is not too small. When the number of observations per consumer is too low, the nonconvexity of the likelihood function makes consistent estimation difficult. However, even when consistently estimating model parameters is infeasible, it was shown that the estimated parameters perform very well in predicting consumer choice on hold-out data, nearly matching the performance of the practically infeasible optimal prediction method (which utilizes knowledge of true model parameters), and significantly out-performing reduced-form prediction methods that do not include information on population structure. To complement the synthetic experiments, the model and estimation procedure are applied to an experimental data set on consumer consideration behavior from Hauser et al. (2010). The estimated parameters indicate that consumers on average have simple preferences that utilize the top 6 most important attributes out of the 16 attributes in total, and that the usage of these simple preferences for consideration behavior averages out to about a $7.8 loss in utility for consumers, under some further assumptions.

The baseline model presented in this paper has a few limitations. Two major ones include the assumption of uncorrelated attributes in the background product distribution, and the assumption that there is a single set of true latent preferences for all consumers. Thus, the two extensions discussed in the final section of this paper, which propose ways to generalize the model beyond these two limitations, should form the basis for future work. In addition, the work here assumes a parametric distribution for the cognitive cost parameter. This assumption could also be relaxed, for example, by using a mixture of Gaussians instead of a single exponential distribution.

The empirical application of the model to experimental consideration data could also be improved in several ways. One shortcoming of the current results relates to the distribution of cognitive costs; the model estimated here assumes that the cognitive cost parameter is exponential distributed in the population, though it's unclear if this is a good assumption in practice. A natural next step could be to modify the model to allow for more general distributions over cognitive costs, and to then estimate such a model on the data used in this paper. Another assumption made in the empirical work here is that all consumers share a single 'true' latent preference. While this assumption is not extravagant in this context, due to most of the attributes being unambiguously good or bad (e.g. high price is bad, longer battery life good), there could still be the case that different consumers have different views about what constitutes objective quality, so that modeling multiple latent preference might be a useful extension. Finally, the data used in the empirical application here comes from an experimental study. While such experimental data provides a great starting point for assessing the validity and usefulness of the model, it's somewhat less desirable for describing consumer behavior in non-experimental contexts. Thus, in order to derive estimates of cognitive cost that might be more useful to practitioners, it would be useful to estimate the model proposed in this paper on data collected from a real-life consideration context, e.g. click data from an online retailer.

Bibliography

Greg M. Allenby and Peter E. Rossi. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89:57–78, 1998.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

Martin Burda, Matthew Harding, and Jerry A. Hausman. A bayesian mixed logit-probit model for multinomial choice. *Journal of Econometrics*, 147(2):232–246, 2008.

Martin Burda, Matthew Harding, and Jerry A. Hausman. A bayesian semi-parametric competing risk model with unobserved heterogeneity. *Journal of Applied Econometrics*, forthcoming, 2013.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Jean-Pierre Dubé, Jeremy T Fox, and Che-Lin Su. Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica*, 80(5):2231–2267, 2012.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online

learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Theodoros Evgeniou, Constantinos Boussios, and Giorgos Zacharia. Generalized robust conjoint estimation. *Marketing Science*, 24(3):415–429, 2005.

Jeremy T Fox and Amit Gandhi. Nonparametric identification and estimation of random coefficients in multinomial choice models. *The RAND Journal of Economics*, 47 (1):118–139, 2016.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2nd edition, 2008.

Xavier Gabaix. A sparsity-based model of bounded rationality. *Working Paper*, 2014.

Timothy J. Gilbride and Greg M. Allenby. A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science*, 23(3):391–406, 2004.

Benjamin J Gillen, Hyungsik Roger Moon, and Matthew Shum. Demand estimation with high-dimensional product characteristics. *Bayesian Model Comparison*, pages 301–24, 2014.

Noah D. Goodman, Joshua B. Tenenbaum, Jacob Feldman, and Thomas L. Griffiths. A rational analysis of rule-based concept learning. *Cognitive Science*, 32:108–154, 2008.

Paul E Green and Vithala R Rao. Conjoint measurement for quantifying judgmental data. *Journal of Marketing research*, pages 355–363, 1971.

Paul E Green and Venkataraman Srinivasan. Conjoint analysis in consumer research: issues and outlook. *Journal of consumer research*, pages 103–123, 1978.

John R. Hauser, Olivier Toubia, Theodoros Evgeniou, Rene Befurt, and Daria Dzyabura. Disjunctions of conjunctions, cognitive simplicity, and consideration sets. *Journal of Marketing Research*, 47:485–496, 2010.

Stefan Hoderlein and Martin Spindler. High-dimensional random coefficient models with an application to consumer demand. *mimeo*, 2014.

Stefan Hoderlein, Jussi Klemelä, and Enno Mammen. Analyzing the random coefficient model nonparametrically. *Econometric Theory*, 26(03):804–837, 2010.

Kamel Jedidi and Rajeev Kohli. Probabilistic subset-conjunctive models for heterogeneous consumers. *Journal of Marketing Research*, 42(4):483–494, 2005.

Jordan J Louviere and George Woodworth. Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research (JMR)*, 20(4), 1983.

Daniel Mcfadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York, 1974.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

Kenneth Train and David Revelt. Mixed logit with repeated choices: Households' choices of appliance efficiency level. *The Review of Economics and Statistics*, 80(4): 647–657, 1998.

Michael Yee, Ely Dahan, John R. Hauser, and James Orlin. Greedoid-based noncompensatory inference. *Marketing Science*, 26:532–549, 2007.

Appendix 2.A General background distributions

In the baseline model of Definition 10, the background product distribution is taken to be such that the attribute-attribute correlation matrix is the identity. This is a substantive restriction, as there's no reason to expect that attributes of goods in the background product distribution will be perfectly uncorrelated. Furthermore, the stochastic gradient descent procedure for maximum likelihood estimation in Algorithm 7 depended on having analytical expressions for $\frac{d\rho_s}{d\beta}$, which isn't available in general. This section generalizes the mode in Definition 10 to allow for more general background product distribution, and discusses how the maximum likelihood estimation might be carried out.

The generalized model is very similar to the initial model: the only difference is that a the generalized preferences simplification of Procedure 4 is used instead of the simple simplification procedure in Equation 2.1. The model is formally stated in Definition 13 below:

Definition 13. *Model with general background product distributions:*

- *Consumers* $i = 1, ..., N$
- *Goods* $j = 1, ..., J$
- *Each good has up to* K *binary attributes, so that a good* j *may be identified with its attribute set* $X_j \in \{0,1\}^K$
- *Background distribution over good* $\nu \in \Delta(\mathbb{R}^K)$
- *All consumers share latent preferences* $\beta \in \mathbb{R}^K$
- *Each consumer has an IID exponentially distributed cognitive cost parameter* $\lambda_i \sim \mathcal{E}(1/\mu)$ *so that* $F_{\mathcal{E}}(a) = 1 - \exp\left(-\frac{1}{\mu}a\right)$
- *Each consumer* i *has simple preferences* $\tilde{\beta}_i$ *from* β *via Procedure 4*
- *For each good* j*, consumer* i *chooses to consume good* j *with probability as defined in Procedure 5, with the set choice set* C *consisting of only the good* j*.*
- *The consumption decision* $\{Y_{ij}\}$ *of each consumer for each good, and the attributes* $\{X_j\}$ *of each good is recorded.*

- *The objective of the analyst is to estimate β and μ given Y and X.*

The main complication with this model is that the relationship between the active attribute sets and the λ_i is no longer quite so simple as in the orthonormal case; the formula in Equation 2.1 no longer applies. Given that the gradient-based algorithm for finding the MLE of the baseline mode of Definition 10 relies on computing analytical gradients, a new estimation procedure must be used.

Instead, the MLE can be computed via an EM algorithm, where the M-step involves SGD-based optimization of $\hat{\beta}$ and $\hat{\mu}$, and the E-step involves updating the posterior probability that each consumer has a given values.

1. Initialize $\hat{\beta}, \hat{\mu}$
2. Iterate the following E and M steps
   (a) E step: compute the posterior probability of each consumer having a certain active attribute set, given current parameter estimates and choice data:

$$p_i^s := P(s_i = s|X_i, Y_i, \hat{\beta}, \hat{\mu}) = \frac{P(s_i = s|\hat{\beta}, \hat{\mu}) l_i^s}{\sum_{s'} P(s_i = s'|\hat{\beta}, \hat{\mu}) l_i^{s'}}$$

   The two quantities appearing in the expression may be computed as follows:

   - $P(s_i = s|\hat{\beta}, \hat{\mu})$ is the prior probability that a consumer has a given active attribute set s. While this does not have an analytical expression, it can be computed on a case by case basis by simply computing the lasso path corresponding to Equation 2.7. This lasso path indicates which attributes are active at each value of λ_i. We can then invert this mapping to figure out, for each active attribute set appearing in the lasso set, the range of λ_i values that generate that active attribute set. Then, using the distribution on λ_i given by $\mathcal{E}(\hat{\mu})$, the desired prior probabilities can be computed.
   - l_i^s is the probability of observing the choice data of consumer i if con-

sumer i has active attribute set s. This quantity can be written

$$l_i^s = \prod_j \left(\frac{\exp(X_j \cdot \tilde{\beta}^s)}{1 + \sum_{j' \in C} \exp(X_{j'} \cdot \tilde{\beta}^s)} \right)^{Y_{ij}} \left(\frac{1}{1 + \sum_{j' \in C} \exp(X_{j'} \cdot \tilde{\beta}^s)} \right)^{1 - Y_{ij}}$$
$$= \prod_j \left(\frac{\exp(X_j \cdot \tilde{\beta}^s Y_{ij})}{1 + \sum_{j' \in C} \exp(X_{j'} \cdot \tilde{\beta}^s)} \right)$$

where $\tilde{\beta}^s$ refers to least square approximation to $\hat{\beta}$ using only the active attribute set s, as in Equation 2.8.

(b) M step: SGD-based optimization of the likelihood, fixing the probabilities that each consumer has each active attribute set. Given the p_i^s computed the E-step, as well as the current parameter estimates $\hat{\beta}, \hat{\mu}$, this step updates the $\hat{\beta}$ and $\hat{\mu}$ to maximize the total likelihood. This likelihood is:

$$L(\beta', \mu' | X, Y, \hat{\beta}, \hat{\mu}) = \prod_i \sum_s p_i^s \prod_j \left(\frac{\exp(X_j \cdot \tilde{\beta'}^s Y_{ij})}{1 + \sum_{j' \in C} \exp(X_{j'} \cdot \tilde{\beta'}^s)} \right)$$

where $\tilde{\beta'}^s$ refers to simplified preference parameters computed by the OLS procedure in Equation 2.8, using β' and μ' as the latent parameters, with the set of active attributes being s.

- Optimizing $\hat{\mu}$: to find the new optimal $\hat{\mu}$, one can simply find the value of μ' that induces the given $p^s := \sum_i p_i^s$. That is, given $\hat{\beta}$, one can compute the lasso path in order to find the range of λ_i values that correspond to each active attribute set s. Then, given μ', one may analytically derive the formula for $\rho_s := P(s_i = s | \hat{\beta}, \mu')$, and optimize the μ' via stochastic gradient descent, with the gradient being:

$$\frac{d\rho_s}{d\mu} = \frac{1}{\mu^2} |\underline{\lambda}^s| \exp\left(-\frac{1}{\mu} |\underline{\lambda}^s| \right) - \frac{1}{\mu^2} |\overline{\lambda}^s| \exp\left(-\frac{1}{\mu} |\overline{\lambda}^s| \right)$$

where $\underline{\lambda}^s$ and $\overline{\lambda}^s$ refer to the lower and upper bounds of the range of cognitive cost parameters λ_i that generate active set s.

- Optimizing $\hat{\beta}$: updating $\hat{\beta}$ can be done via stochastic gradient descent on the $l := L(\beta', \mu' | X, Y, \hat{\beta}, \hat{\mu})$ expression above. The derivative can be

written:

$$\frac{dl}{d\beta'_k} = \sum_i \frac{1}{L_i} \sum_s p_i^s \prod_j \left(\frac{\exp(X_j \cdot \tilde{\beta'}^s Y_{ij})}{1 + \sum_{j' \in C} \exp(X_{j'} \cdot \tilde{\beta'}^s)} \right)$$
$$= \sum_i \frac{1}{L_i} \sum_s p_i^s \prod_j L_{ij}^s \left(\frac{Y_{ij} + (Y_{ij} - 1)\exp(X_j \cdot \tilde{\beta'}^s)}{1 + \exp(X_j \cdot \tilde{\beta'}^s)} \right) \sum_{k'} X_{jk'} \frac{d\tilde{\beta'}_{k'}^s}{d\beta'_k}$$

where

$$L_i := \sum_s p_i^s \prod_j L_{ij}^s, \quad L_{ij}^s = \left(\frac{\exp(X_j \cdot \tilde{\beta'}^s Y_{ij})}{1 + \sum_{j' \in C} \exp(X_{j'} \cdot \tilde{\beta'}^s)} \right)$$

and $\frac{d\tilde{\beta'}_{k'}^s}{d\beta'_k}$ is the appropriate entry of the matrix

$$\frac{d\tilde{\beta'}^s}{d\beta'} = \Sigma_{s,s}^{-1} \Sigma_{s,:}$$

where $\Sigma_{s,s} = \int_{\mathbb{R}^K} (x'_s x_s) d\nu(x)$ is the attribute-attribute correlation vector generated between attributes in s, and $\Sigma_{s,:} = \int_{\mathbb{R}^K} (x'_s x) d\nu(x)$ is the attribute-attribute correlation vector generated between attributes in s and the set of all attributes. This derivative follows simply from the formula of the OLS regression coefficients when only using a subset s of the attributes, namely $\tilde{\beta'}^s = \Sigma_{s,s}^{-1} \Sigma_{s,:} \beta'$

Appendix 2.B Multiple latent preferences

The model in Definition 10 has only a single set of latent preferences, with consumer preference heterogeneity induced by the cognitive simplicity procedure. While this model is suitable for situations where there is a single global measure of 'quality', it's less appropriate in situations where consumers can have fundamentally different preferences. This section generalizes the model in Definition 10 to account for this heterogeneity in latent preferences.

The extension can be done in a straightforward manner: simply define multiple types of consumers, each with its own latent preferences, and then proceed according

to Definition 10. Concretely:

Definition 14. *Model with multiple latent preferences:*

- *Consumer types* $t = 1, ..., T$
- *Consumers* $i = 1, ..., N$
- *Each consumer* i *is assigned to a type* $t_i \in 1, ..., T$ *according to an IID categorical distribution with* $\eta_t := P(t_i = t)$, $\sum_t \eta_t = 1$
- *Goods* $j = 1, ..., J$
- *Each good has up to* K *binary attributes, so that a good* j *may be identified with its attribute set* $X_j \in \{0, 1\}^K$
- *A type* t *has latent preferences* $\beta^t \in \mathbb{R}^K$, *as well as a mean cognitive cost* μ^t
- *Each consumer has an IID exponentially distributed cognitive cost parameter governed by the preferences and cognitive cost distribution:* $\lambda_i \sim \mathcal{E}(1/\mu^{t_i})$ *so that* $F_{\mathcal{E}}(a) = 1 - \exp\left(-\frac{1}{\mu^{t_i}} a\right)$
- *Each consumer* i *has simple preferences define as the entries of* β *with absolute value greater than* λ_i*:*

$$\tilde{\beta}_i = \beta^{t_i} \odot s_i, \quad s_i \in \{0, 1\}^K, \quad s_{ik} = \mathbb{I}[|\beta_k^{t_i}| > \lambda_i]$$

  where $\odot$ *denotes element-wise multiplication* $\mathbb{I}[|\beta_k^{t_i}| > \lambda_i]$ *takes on value 1 if* $|\beta_k^{t_i}| > \lambda_i$ *else 0.*

- *Consumer* i *has utility for good* j *defined by his simple preferences:*

$$u_{ij} := X_j \cdot \tilde{\beta}_i + \varepsilon_{ij} = X_j \cdot \beta^{t_i} \odot s_i + \varepsilon_{ij}$$

  where ε_{ij} *is an IID Gumbel shock.*

- *For each good* j, *consumer* i *chooses to consume good* j *IFF* $u_{ij} > 0$. *Let* Y_{ij} *denote this decision.*
- *The consumption decision* $\{Y_{ij}\}$ *of each consumer for each good, and the attributes* $\{X_j\}$ *of each good is recorded.*

- *The objective of the analyst is to estimate β^t, μ^t, η^t for each $t \in 1, ..., T$ given Y and X.*

The estimation will be done via maximum likelihood. Let $\theta := (\beta^t, \mu^t, \eta^t)_{t=1}^T$ be the vector of all parameters to be estimated. Treating the types t_i of each consumer as missing variables, the likelihood may be written

$$l(\theta|X, Y) = \sum_i \log(\mathbb{E}[L_i^t]) \tag{2.9}$$

where

$$L_i^t = \sum_{s \in S} \rho_s^t \prod_j l_{ij}^{st}, \quad \rho_s^t = \exp\left(-\frac{1}{\mu^t}|\beta_{\underline{k}_s}^t|\right) - \exp\left(-\frac{1}{\mu^t}|\beta_{k_s}^t|\right), \quad l_{ij}^{st} = \frac{\exp(X_j \cdot \beta^t \odot sY_{ij})}{1 + \exp(X_j \cdot \beta^t \odot s)} \tag{2.10}$$

and the inner expectation of Equation 2.9 is taken over all types t.

Estimation of this model may then proceed via an EM-type algorithm (see e.g. Dempster et al. (1977)), where the E-step involves computing $\nu_i^t = P(t_i = t|X_i, Y_i)$ for every consumer i and consumer type t given current parameter estimates, and the M-step is a stochastic gradient ascent procedure similar to 7 that optimizes the β^t and μ^t for each type t. Informally, the process is:

1. Initialize $\theta = (\beta^t, \mu^t, \eta^t)_{t=1}^T$
2. Iterate the following E and M steps
   (a) E step: compute probability of each consumer being a given type

$$\nu_i^t \leftarrow P(t_i = t|X_i, Y_i) = \frac{\eta_t L_i^t}{\sum_{t'=1}^T \eta_{t'} L_i^{t'}}$$

   (b) M step: maximize the likelihood over β^t, μ^t, taking ν_i^t as given:

$$l(\theta|X, Y, \nu) = \sum_i \log(\sum_t \nu_i^t L_i^t) \tag{2.11}$$

   This generates updated values β^t, μ^t. η^t is updated by simply summing ν_i^t over all i.

The maximization step can be done via gradient. The gradients are:

$$\begin{aligned}
\frac{dl(\theta|X,Y,\nu)}{d\mu^t} &= \sum_i \left(\frac{1}{\sum_{t'} \nu_i^{t'} L_i^{t'}}\right) \sum_{t'} \nu_i^{t'} \frac{dL_i^{t'}}{d\mu^t} \\
&= \sum_i \left(\frac{1}{\sum_{t'} \nu_i^{t'} L_i^{t'}}\right) \nu_i^t \frac{dL_i^t}{d\mu^t} \\
&= \sum_i \left(\frac{1}{\sum_{t'} \nu_i^{t'} L_i^{t'}}\right) \nu_i^t \sum_{s \in S^t} \left(\frac{d\rho_s^t}{d\mu^t} \prod_j l_{ij}^{st}\right) \qquad (2.12) \\
\frac{dl(\theta|X,Y,\nu)}{d\beta_k^t} &= \sum_i \left(\frac{1}{\sum_{t'} \nu_i^{t'} L_i^{t'}}\right) \sum_{t'} \nu_i^{t'} \frac{dL_i^{t'}}{d\beta_k^t} \\
&= \sum_i \left(\frac{1}{\sum_{t'} \nu_i^{t'} L_i^{t'}}\right) \nu_i^t \frac{dL_i^t}{d\beta_k^t} \\
&= \sum_i \left(\frac{1}{\sum_{t'} \nu_i^{t'} L_i^{t'}}\right) \nu_i^t \sum_{s \in S^t} \left(\rho_s \prod_j l_{ij}^s\right) \left(\frac{1}{\rho_s} \frac{d\rho_s^t}{d\beta_k^t} + \sum_j \frac{1}{l_{ij}^{st}} \frac{dl_{ij}^{st}}{d\beta_k^t}\right) \qquad (2.13)
\end{aligned}$$

Where the expressions for S^t, $\frac{d\rho_s^t}{d\mu^t}$, $\frac{dl_{ij}^{st}}{d\beta_k^t}$ are identical to their non-superscripted counterparts in Section 2.3.3, with the μ and β replaced with μ^t and β^t.

Chapter 3

Academic Performance and Entrance Exam Rank in the University of Bologna Medical School

3.1 Introduction and context

Each year, applicants to the University of Bologna's medical school take an entrance exam, and are then ranked based on their performance, with all the scores and rankings subsequently made public by the university. The ranking itself is used only for determining admission, so that it has no direct effect on a student's academic performance conditional on matriculation. In addition, the ranking contains almost zero information that isn't already included in the full list of scores of all exam takers published by the university. Finally, the rank serves as a competitively-framed measure of a student's academic ability, as rank is a direct numerical comparison of a student's performance relative to their peers.

This setting allows us to gain insight into how this particular measure of student rank, which is practically irrelevant (conditional on admission), but also highly salient and competitively-framed, impacts students' academic performance further down the road. As rank plays no role in students' academic careers beyond admission to the medical school, any impact must operate through some sort of psychological channel. We thus give this impact a psychological interpretation: does receiving a worse rank

on the entrance exam cause a student to feel demotivated and perform worse academically, or does it motivate a student to exert more effort and perform better? To that end, we characterize the average effect of rank on academic performance, as well as heterogeneity in this effect for students of different genders and academic aptitudes. We also analyze how students' academic performance responds on both the intensive (GPA) and extensive margins (courseload) to rank. Our main findings are as follow:

1. Receiving a higher (i.e. worse) rank appears to have a demotivational effect on students. That is, rank negatively impacts academic performance on average.
2. The impact is about equal for male and female students students.
3. The impact is largely concentrated on students who are higher (i.e. worse) ranked.
4. The impact operates mostly through extensive (courseload) rather than intensive (GPA) margin.

Our work here relates to the literature on student motivation and how it is impacted by competition. A large literature in psychology and education-studies has focused on how students' academic performance respond to competition focused vs personal mastery-focused teaching styles (see e.g. Meece et al. (2006) for an overview). These results largely suggest that competition-focused teaching styles (which compare students against each other) are generally worse for student learning than more individual mastery-focused teaching styles. In a longitudinal study of 8th graders, Roeser and Eccles (1998) found that more competitive teaching styles were correlated with students feeling less motivated, leading to lower subject mastery as well as other negative effects such as truancy and depression. Similarly, in a study of junior high and high school students, Ames and Archer (1988) found that more competitive environments caused students to attribute failures to their own ability, again leading to worse performance. Essentially the same result was found for 6/7th graders in Anderman and Young (1994), in which competitive teaching styles was associated with lower understanding of the subject material.

Our work in this paper contributes to the literature on student response to competition by quantifying the extent to which one particular competitively-framed signal of student ability (the student's rank on the University of Bologna medical school entrance exam) affects subsequent academic performance. Our work expands upon the existing literature here by: (1) deriving quantitative measures of effect sizes (2) characterizing the heterogeneity of this effect for students of different genders and different

levels of academic aptitude, (3) investigating this effect for college-age students rather than adolescents.

Our work here also relates to the large literature in economics and psychology on gender differences in response to competition (see e.g. Niederle and Vesterlund (2011) for an overview). In recent years, there has been intense interest in understanding why women are underrepresented in many high-ranking positions in academia, government, and industry. As attainment of high-ranking positions tend to be the outcome of competitive processes, many authors have suggested that gender differences in response to competition could be responsible for a large part of this discrepancy (see e.g. Niederle and Vesterlund (2011), Niederle and Vesterlund (2010), Ellison and Swanson (2010) among many others).

Gender differences in response to competition have been documented in many settings. In an experimental study, Gneezy et al. (2003) found that men and women were equally productive when paid according to a piece rate, but women underperformed relative to men when when payment was structured as a tournament. In another experimental study, Inzlicht and Ben-Zeev (2000) found that this gender difference could be produced by social suggestion, without explicitly altering any incentive structures: merely having women take a math test in a room together with men caused worse performance relative to taking test in a room with other women. This difference has also been found for children, as in Gneezy and Rustichini (2004), which analyzed the problem in the setting of a children's' footrace in Israel. However, while the gender difference in response to competition has been documented in many situations (largely in modern Western societies), some studies have suggested that this may be a consequence of socialization rather than biology. For example, Gneezy et al. (2009) experimentally studies this question in a matriarchal society in India and find that this gender difference is nonexistent, and Andersen et al. (2013) find that the difference appears at around puberty age in a patriarchal society in Tanzania.

Our work in this paper, while performed in a modern Western society, produces results that run counter to most of the work listed above. Specifically, we find that, in the context of the University Bologna's school of medicine, male and female students' academic performance do not exhibit any differential response to entrance exam rank. Thus, it appears that the academic performance of women are no more or less sensitive than the academic performance of men to this particular competitively-framed signal of ability within a Western academic institution.

The paper is organized as follows: Section 3.2 describes the data and provide some

basic summary statistics. Section 3.3 describes our empirical strategy, performs the basic regression analysis, and contains the main results. Section 3.4 performs various robustness checks on our main results. Section 3.5 concludes.

3.2 Setting and data

After graduating from high school, Italian students may choose to enter a variety of universities. It is notable in particular that students do not apply to schools, and then upon arrival, declare their major as is the case in the US. Instead, students apply to particular programs of study with very limited amounts of flexibility. Medical school in Italy is also not a postgraduate program, but instead a 6-year-long program beginning immediately after high school.

The admissions process for each medical school is determined primarily by a single multiple choice exam, administered across Italy all in a single day, at each college of medicine. Students apply to a school by attending the exam at the university which they wish to attend. Thus, students are able to apply only to a single school (this has changed recently, but this fact holds for all years present in our data). This exam is divided into 4 sections: general culture, biology, chemistry, math/physics. Admissions to the university is determined entirely by total score on this test, with tiebreakers being some deterministic and lexicographic function of some subset of these variables: score on each of the 4 subsections, high school graduation grade, and age. The precise tie-breaking method varies from year to year, and will be described in detail below in Section 3.2.2.

The high school graduation grade is determined by an exam taken at the end of a student's high school career. High school graduation grades are not entirely comparable across schools; while one half is prepared by a national board and standardized across all schools, the other half is prepared by the instructors of each high school. Furthermore, the exam questions are generally open-ended, and grading is done by the instructors at each high school.

As soon as all of the medical school entrance exams have been graded, all scores on the exam and the associated rankings are publicly revealed (with test taker names redacted). Thus, everyone gets to see the distribution of scores on each section of the exam, and also get to see their own position within that distribution.

Though students may apply only to a single medical school, they may also apply to

other programs (either at the same university or not). Thus, admissions is performed on a round by round basis, so that each person offered admission who turns it down opens up a slot for the next-best-ranked student.

The analyses in this paper focus on applicants to the University of Bologna's medical school who took the entrance exam between 2005 and 2011. Though the dataset records many attributes of each applicant, we focus on a few variables of interest:

- Outcome variables: measures of academic success. We will focus mainly on FCGPA.
  - Freshman Corrected GPA (FCGPA), defined as the sum of grades received by September 30 of the year after enrollment, weighted by exam units[1] divided by 60 (60 is the theoretical number of units each student should take by the end of freshman year). All grades are on the same scale, of 0 to 30. Thus, FCGPA characterizes a student's performance on the basis of both her grades and the number of units taken during the first year of med school[2].
  - Freshman un-corrected GPA (FGPA), which is just the FCGPA, except without the correction for number of units taken.
  - Freshman total number of units taken (TNU), which is the total number of exam units a student chooses to take during their first year.
- Covariates of interest
  - Position in the admission ranking (rank), with a rank of 1 corresponding to the top performing candidate, 2 the second best candidate, etc. These ranks are determined primarily by entrance test scores; this will be discussed in detail later in Section 3.2.2.
  - The gender of the student.
- Control covariates:
  - The High School Graduation Grade (HSGG).
  - The age of the student.

[1]Ungraded exams are given a grade equal to the average grade received by the cohort in the same time period.

[2]While students have essentially no flexibility in customizing their curriculum in the first year, they can nonetheless choose how many exams to take out of all the ones they are (theoretically) supposed to take by the end of the academic year.

– Score received on each of the four entrance test sections (general culture, biology, chemistry, physics and math)

Due to students attending foreign high schools having high school grades that are often unreliable, we restrict attention to just Italian students. We also exclude students who sit for a medical school exam in a year that is not the same as the year in which they graduate, since it is often the case that these students were previously enrolled in another major in college, and would likely have already completed a part of the first year med school curriculum, especially if they were previously enrolled in a related major (e.g. biology), making their first year GPAs incomparable with those of students who sat for medical school exams immediately after high school. Finally, as our analysis focuses on the effect of ranking on students' academic performance, we restrict attention to students who subsequently matriculate into the University of Bologna's medical school.

3.2.1 Summary statistics

Table 3.1: Summary statistics by gender. Male students rank better, female students get better grades.

	Female		Male		All	
Statistic	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
FCGPA	20.880	7.857	19.376	8.432	20.217	8.147
FGPA	28.083	1.801	27.701	2.052	27.915	1.920
TNU	44.288	15.881	41.755	18.281	43.173	17.021
rank	249.951	200.896	209.031	184.604	231.930	194.893
culture score	23.055	5.917	23.732	6.131	23.354	6.020
bio score	12.091	3.645	12.082	3.795	12.087	3.711
chem score	8.117	2.716	8.374	2.847	8.230	2.777
math/phys score	5.612	2.960	6.543	3.045	6.022	3.032
age	18.957	0.273	18.989	0.437	18.971	0.355
HSGG	94.245	8.486	89.426	10.718	92.123	9.826
observations	906		713		1619	

Summary statistics of our sample, broken down by gender, are presented in Table 3.1. We have a total of around 1600 students in our sample, with about 56% being female. male students appear to perform slightly better on average in the entrance exam relative to female students, but female students tend to have higher FCGPAs. Female

students also tend to have higher high school GPAs relative to male students. When it comes to test subscores, the scores are largely comparable, except male students have slightly higher scores on the math/physics portion of the exam relative to female students. Finally, FGPA exhibits much less variance than FCGPA, indicating that much of the variation in FCGPA is coming for the correction for TNU.

Table 3.2: Summary statistics by year. Culture part of the exam increases in importance over time. Average ranking of enrolled students decrease across time, indicating better students are enrolling year after year.

Year	2005	2006	2007	2008	2009	2010	2011
FCGPA	21.779	20.885	21.050	22.156	24.335	21.147	9.059
FGPA	27.657	27.882	27.730	28.142	28.051	27.682	28.319
TNU	46.739	44.390	45.361	46.866	51.819	45.568	19.063
rank	266.185	276.908	266.051	244.736	173.809	192.072	183.621
culture score	17.042	20.764	21.317	19.107	28.382	28.247	31.057
bio score	11.093	12.187	13.916	16.892	8.444	10.284	10.776
chem score	9.929	8.860	8.979	8.049	7.320	6.147	7.939
math/phys score	9.587	4.223	6.168	6.656	5.948	5.461	3.656
age	18.958	18.941	19.003	19.000	18.980	18.954	18.951
HSGG	91.596	91.634	90.174	92.323	92.728	93.468	93.402
observations	243	241	252	247	210	220	206

Summary statistics, broken down by year, are presented in Table 3.2. It is notable that the average rank of students who enroll in the University of Bologna's medical school is decreasing over time, with a major jump from 2008 to 2009, indicating that higher-ranked candidates have become more likely to enroll in recent years. For the entrance exam subscores, it appears that scores on the culture portion of the exam has steadily increased between 2005 and 2011, while the scores on the other sections have decreased to compensate. In particular, the math/physics scores have dropped dramatically over the years in our sample. FCGPA remains fairly stable, with the exception of 2011, for which we only have data on the first half of the school year, resulting in an FCGPA that's half the other years' on average.

We plot the distribution of FCGPA, FGPA, and TNU in Figure 3-1. Female students appear to consistently perform better than male students, with the distribution functions being more concentrated on higher GPA levels. This difference is reflected in both the FGPA and the TNU, indicating that female students both performed better on average in their first year classes, and also took more units.

Figure 3-1: Distribution of FCGPA, FGPA, and TNU, by gender. Female students perform better on all three measures of academic success.

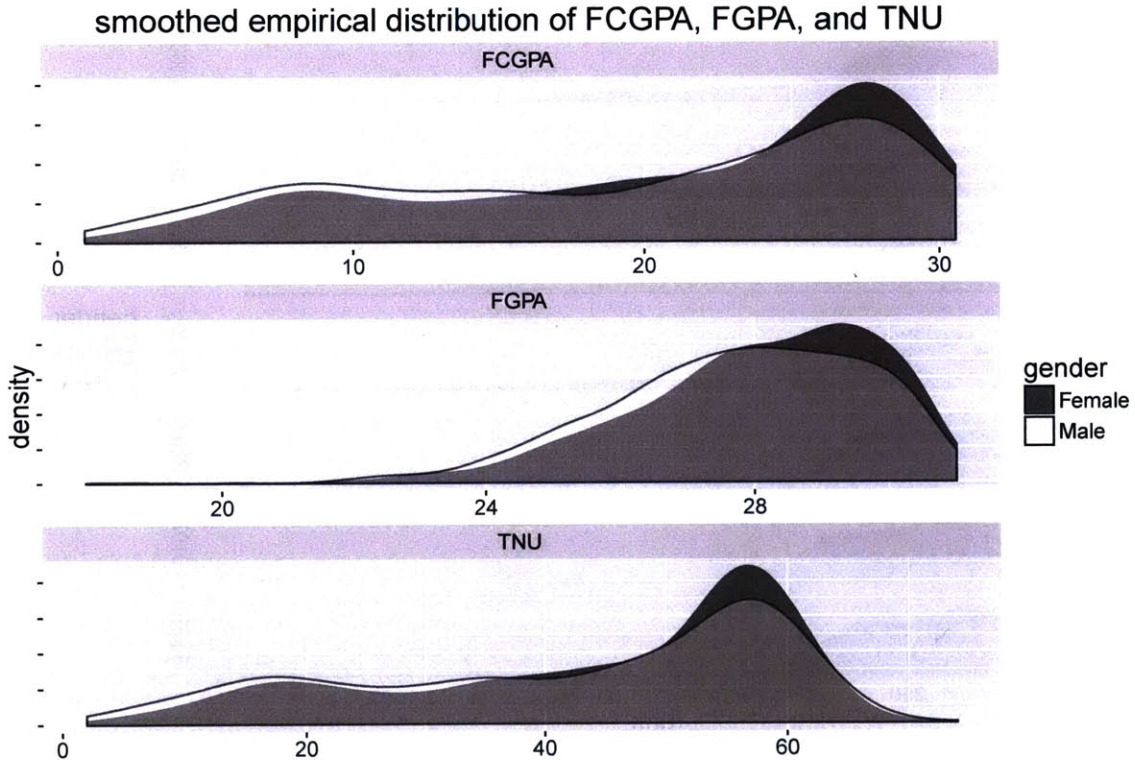
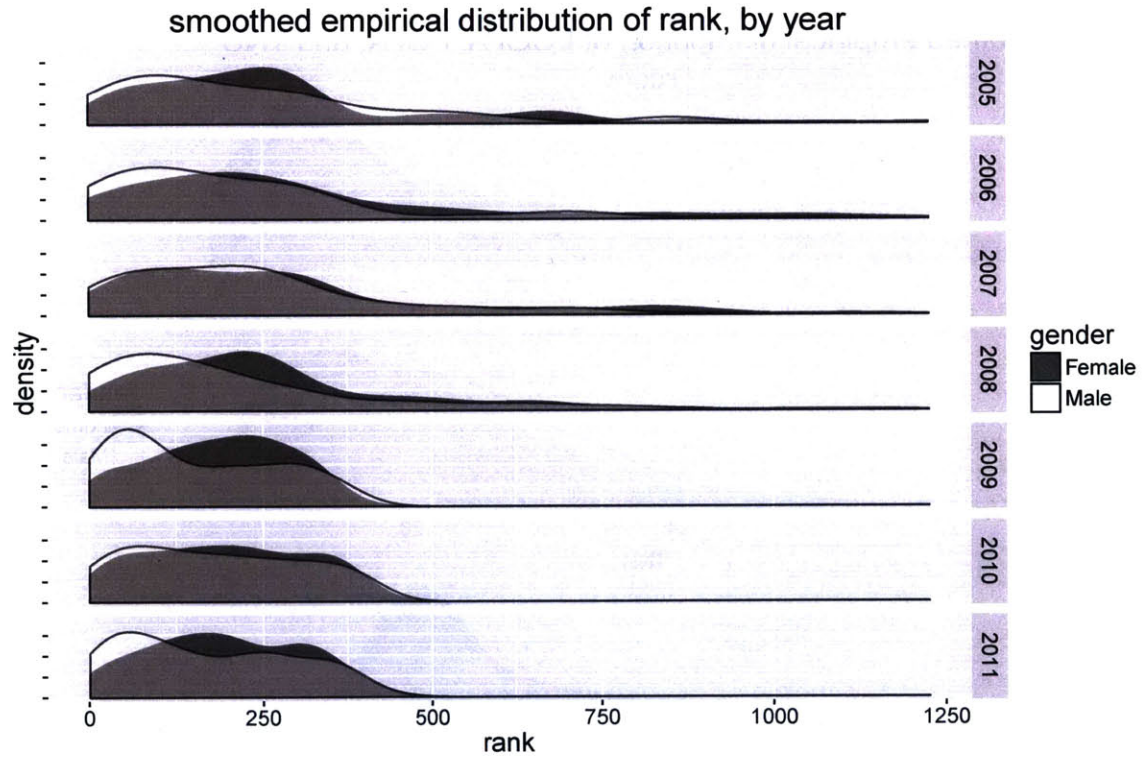


Figure 3-2: Distribution of rank, by gender and year. Male students have better rank.



We also plot the distribution of ranks across genders and years in Figure 3-2. In spite of female students performing better in terms of measures of academic success, male students appear to consistently perform better than female students on the entrance exam. Furthermore, out of the top-100 ranked students each year who enroll in Bologna's medical school, male students outnumbered female students in 5 of the 7 years in our sample, with the gender imbalance tipping more in favor of male students in recent years. This breakdown presented in Table 3.3.

Table 3.3: Fraction of enrolled female students ranked in the top 100 on the entrance exam, by year. This value is computed by taking those students with a rank of 1 through 100, keeping only those who eventually enrolled in the University of Bologna's medical school, and computing the fraction of this remaining sample that is female.

Year	Fraction female in top 100 of rank
2005	0.4821429
2006	0.5087719
2007	0.5666667
2008	0.4307692
2009	0.3623188
2010	0.4590164
2011	0.3770492

3.2.2 Rank

Of central importance to the admission process is the rank of each applicant on the entrance exam. Admissions to the medical school is determined according to a deterministic cut-off rule: if there are 300 available places, the first 300 applicants in the ranking receive admission, whereas individuals further down in the ranking are offered admission only if some people in the pool of initially admitted students do not enroll.

The ranking is a deterministic function of a few attributes: overall test score, test scores on each of the subsection of the test (i.e. general culture, biology, chemistry, math/physics), high school graduation grade, age. This function is lexicographic, so that students are first ranked by the first attribute, and then the second attribute is applied to separate students with identical levels of the first attribute, and so forth. The order in which these attribute enter the lexicographic ranking function varies from year to year; the overall score is always the first attribute, but the subsequent attributes change year to year:

Procedure 6. *Procedure for generating ranks, for each year:*

- *Years 2005-2006: Overall score, general culture score, biology score, chemistry score, math/physics score.*
- *Years 2007-2009: Overall score, general culture score, HS graduation grade, biology score, chemistry score, math/physics score, age.*
- *Years 2010-2011: Overall score, general culture score, biology score, chemistry score, math/physics score, HS graduation grade, age.*

If, in a given year, two or more students tie on all ranking criteria, then they receive the same position in the ranking. This is quite rare, and occurs for only 6 candidates in our sample. When results are published, in addition to her own position in the ranking, each candidate sees the entire anonymized ranking, *including* the overall test score and each subtest score of all other applicants. Regardless of whether such criteria are used as tie-breakers, neither the high school graduation grade nor the age of test takers are visible to candidates.

Given the way ranking is determined, it is not surprising that, in a given year, a candidate's position in the ranking is very strongly (though not perfectly) correlated with some linear combination of the four subtest scores, age, and the high school graduation grade. Table 3.4 reports coefficient estimates and standard errors of a year-by-year regression of rank position on subtest scores, HSGG and age. The R^2 values range between .90 and .94 in every year of observation. Notice that the only significant coefficients are the ones on the four subtest scores: also in the years during which HSGG was used as a tie-breaker, the single most important factor in determining one's position was the overall test score, that is, the sum of the four subtest scores.

3.3 Regression Results

3.3.1 Empirical strategy

The main objective of this paper is to estimate the impact of a student's rank on their subsequent academic performance. In order to do so, it is necessary to disentangle the effect of rank itself from the effect of other things that are correlated with rank and that also affect academic performance (e.g. person-level attributes like baseline

Table 3.4: Student ranking vs subtest scores, HSGG, and age. Each column reports regression estimates of the rank position of a candidate at the entrance test in a given academic year on subtest scores, HSGG and age. For each academic year, the sample consists of Italian students who graduated from High School in the same year of the test and who eventually enrolled at the University of Bologna's school of medicine in that given year.

	Dependent variable: rank position by academic year						
	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
General culture	-29.73	-34.74	-31.92	-25.12	-18.96	-21.16	-22.45
	(1.49)	(1.39)	(1.10)	(1.21)	(0.75)	(0.88)	(1.01)
Biology	-29.62	-32.02	-28.71	-28.66	-19.04	-21.83	-21.78
	(1.56)	(2.23)	(1.27)	(1.66)	(1.11)	(1.30)	(1.12)
Chemistry	-29.86	-34.13	-34.05	-30.45	-21.68	-22.64	-22.06
	(1.33)	(2.27)	(1.59)	(1.79)	(1.60)	(1.43)	(1.59)
Physics and math	-28.39	-27.98	-27.05	-24.13	-19.35	-17.53	-18.52
	(1.31)	(2.22)	(1.60)	(1.92)	(1.24)	(1.69)	(1.31)
HSGG	-0.03	0.20	0.47	0.13	0.12	0.08	0.39
	(0.40)	(0.50)	(0.31)	(0.44)	(0.31)	(0.40)	(0.37)
Age	0.43	17.33	5.42	11.64	10.73	-14.08	-2.97
	(15.22)	(17.41)	(5.90)	(8.90)	(13.24)	(11.11)	(11.00)
Constant	1668.25	1464.71	1674.19	1383.92	932.84	1509.03	1378.13
	(295.79)	(336.75)	(119.56)	(179.52)	(252.83)	(215.38)	(217.77)
Observations	243	241	252	247	210	220	206
R^2	0.91	0.90	0.94	0.91	0.92	0.91	0.92

intelligence, baseline level of motivation, baseline level of knowledge). The model might look something like this:

$$y_i = f(X_i, r_i) + \varepsilon_i \tag{3.1}$$

where y_i is the relevant measure of academic performance (FCGPA, or FGPA, or TNU), X_i is the set of controls covariates consisting of subtest scores, high school graduation grade, and age.

We first assume that ε_i is independent of X_i and r_i. Though less than ideal, this assumption is plausible in light of the fact that X_i contains a large numbers other indicators of ability, so that omitted variable bias is likely to be small. Then, in the

interest of estimating a single marginal effect, we take a first-order approximation to this equation with respect to rank:

$$y_i = \theta_0 + r_i\theta + g(X_i) + \varepsilon_i \tag{3.2}$$

We observe at this point that, in each year, the rank r_i is a deterministic function of the control covariates X_i. That is, without any restrictions on g, it is impossible to estimate the impact of rank. Thus, we go further and stipulate that g is a linear function of X_i, and estimate θ via OLS. Note that this is equivalent to taking the residuals of y_i and r_i after partialling out the controls, and then looking at the correlation between the two. In doing so, we are essentially stipulating that the residual of rank after linearly accounting for the covariates in X represents additional information that's captured by rank, but not by any of the covariates.

This assumption of the linearity of g can more generally be understood as an assumption about how smoothly the various measures of academic performance y_i should vary with the control covariates in X_i. Intuitively, we expect this variation to be fairly smooth; given two students A and B with approximate the same exam scores and high school grade, it's difficult to argue that one of them should have a much higher GPA than the other, even if e.g A ends up with a higher rank than B due to A having performed better on the general culture section of the exam and worse on the biology part (and thus winning the tiebreaker round). The linearity assumption can then be seen as a restrictive first attempt at capturing this assumption of smoothness. The main analyses of the paper in Section 3.3 is carried through with this linearity assumption. Relaxations of this assumption and other robustness checks to ensure that the results are not purely driven by this linearity assumption are detailed in Section 3.4.

3.3.2 Academic performance and rank

We now analyze whether students' ranks on the entrance exam has an effect on their subsequent academic performance. To do so, we assume as in the previous section that the component of rank which is left over after accounting for subtest scores, HSGG, age in a linear fashion can be understood as not containing any correlation with the potential counterfactual values of FCGPA under different levels of the rank treatment. Thus, we use ordinary least square to assess the impact on students' freshman CGPA of rank and analyze whether this impact is positive (higher values of rank causes, holding all else

constant, a better FCGPA) or negative. This produces the the regression specification in Equation 3.3. As the best linear predictor of rank as a function of subtest scores, HSGG, and age varies from year to year, we interact these attributes with year fixed effects (so that the D_i in Equation 3.3 are indicators for each of the years 2006-2011, with 2005 omitted for collinearity).

$$FCGPA_i = [1 \;\; D_i'] \times [\alpha + H_i\beta + T_i'\gamma + A_i\eta] + rank_i\theta + \varepsilon_i. \tag{3.3}$$

The results of this regression are described in Table 3.5. It appears that after linearly partialling out subtest scores, HSGG, and age, the estimated coefficient of rank on FCGPA is significantly negative. This suggests that on average, receiving a higher (i.e worse) rank leads to worse FCGPA. Under the assumptions made above, this can be taken to be a causal effect, which then suggests something of a demotivation story, in which students who get ranked higher get demotivated about their academic ability, and thus perform worse in their first year of medical school.

Table 3.5: Regression of FCGPA on subtest scores, HSGG, age, year of entrance test fixed effects and interactions, rank. The table reports the estimate and standard error of θ from the regression of Equation 3.3.

	Dependent variable:
	FCGPA
rank	−0.0112***
	(0.00243)
Observations	1,619
Note:	*p<0.1; **p<0.05; ***p<0.01

We may also investigate whether the effect of rank on FCGPA by year. To do so, we take the specification in Equation 3.3 and interact rank with year-fixed effects, producing the specification in Equation 3.4. Given the coefficient estimates from this, it is possible to analyze the effect of rank on FCGPA for each year pair. To do so, we simply sum the relevant covariates; for example to get the effect of rank on FCGPA for students in the year 2006, one would simply take the sum of two coefficients: [rank + (Test year 2006) * rank]. The results from this exercise are listed in Table 3.6. We find that the effect of rank on FCGPA is negative in every year in our sample, though this

effect achieves significance at the 5% level only in 2005, while in 2006 and 2009 the effect is significant at the 10% level. The effect is not significant in other years.

$$FCGPA_i = [1 \;\; D_i'] \times [\alpha + H_i\beta + T_i'\gamma + A_i\eta + rank_i\theta] + \varepsilon_i. \tag{3.4}$$

Table 3.6: Total effect of rank on FCGPA, for each year, for the regression in Equation 3.4. The residual impact of rank on GPA appears to be negative in all years, but only significant in 2005.

	FCGPA	se
exam year 2005	-0.0255***	(0.00586)
exam year 2006	-0.00789*	(0.00456)
exam year 2007	-0.0101	(0.00647)
exam year 2008	-0.0055	(0.00553)
exam year 2009	-0.0172*	(0.0102)
exam year 2010	-0.00753	(0.00828)
exam year 2011	-0.00671	(0.00931)
Observations	1619	
*** p<0.01, ** p<0.05, * p<0.1		

3.3.3 Quartile differences

We are also interested in understanding whether the effect of rank differs at different levels of rank, as we suspect it may be the case that students ranked near to top might have different reactions to ranking when compared to those who with a position close to (or even below) the admission cut-off point. In order to do so, we put in linear spline for rank. Instead of having GPA be a linear function of rank, we set cut-points at the 1st, 2nd, and 3rd quartiles (i.e. the 25, 50, and 75th percentiles), together with the appropriate linear spline functions. These cutoffs are denoted $RQ1, RQ2, RQ3$. These quartile are separately computed for each year.

We first run this specification without interaction rank with year, so as to get mean effect of rank on FCGPA across all years. This specification is described in Equation 3.5.

$$FCGPA_i = [1 \;\; D_i'] \times [\alpha + H_i\beta + T_i'\gamma + A_i\eta] + rank_i\theta$$
$$+ [(rank_i - RQ1_i)^+, (rank_i - RQ2_i)^+, (rank_i - RQ3_i)^+]'\mu + \varepsilon_i. \quad (3.5)$$

The output of this regressions allows us to find the total effect of rank on FCGPA for each quartile (e.g. to get the effect of rank on FCGPA for students in the 3rd quartile, one would simply take the sum of four coefficients: [rank + (rank first quartile spline) + (rank second quartile spline)). These total effects are listed in Table 3.7. It can be seen from this table that the only quartile for which the total effect of rank on FCGPA is significantly nonzero is the third quartile, where this total effect is negative. This suggests that this demotivation effect is mostly driven by students in the third quartile.

Table 3.7: Total effect of rank on FCGPA, for each quartile, for the regression in Equation 3.5. The only quartile for which this effect is significant is the third quartile, and it is significantly negative.

	FCGPA	se
1st quartile	0.0199	(0.0179)
2nd quartile	0.0129	(0.00795)
3rd quartile	-0.0208***	(0.00695)
4th quartile	-0.000765	(0.00429)
Observations	1619	
*** p<0.01, ** p<0.05, * p<0.1		

We now interact rank with year, and also include quartile splines as before. This specification is described in Equation 3.6.

$$FCGPA_i = [1 \;\; D_i'] \times [\alpha + H_i\beta + T_i'\gamma + A_i\eta + rank_i\theta]$$
$$+ [(rank_i - RQ1_i)^+, (rank_i - RQ2_i)^+, (rank_i - RQ3_i)^+]'\mu + \varepsilon_i. \quad (3.6)$$

We may also analyze the total effect of rank on FCGPA for each (year, quartile) pair, which we do in the same manner as above (e.g. to get the effect of rank on FCGPA for students in the 3rd quartile in the year 2006, one would simply take the sum of four coefficients: [rank + (rank first quartile spline) + (rank second quartile spline) + (test year 2006) * rank]). These results are listed in Table 3.8.

Table 3.8: Total effect of rank on FCGPA, for each (year, quartile) pair, for the regression in Equation 3.6. Effects of rank on FCGPA are more negative for higher quartiles of rank. Furthermore, the significant effects are all negative, and mostly occur in the third quartile of rank.

	FCGPA	se
exam year 2005, 1st quartile	-0.00353	(0.0206)
exam year 2005, 2nd quartile	-0.00582	(0.0107)
exam year 2005, 3rd quartile	-0.0411***	(0.0104)
exam year 2005, 4th quartile	-0.0176**	(0.00734)
exam year 2006, 1st quartile	0.0135	(0.0194)
exam year 2006, 2nd quartile	0.0112	(0.00916)
exam year 2006, 3rd quartile	-0.024***	(0.00891)
exam year 2006, 4th quartile	-0.000526	(0.00573)
exam year 2007, 1st quartile	0.0139	(0.0217)
exam year 2007, 2nd quartile	0.0116	(0.0117)
exam year 2007, 3rd quartile	-0.0236**	(0.0112)
exam year 2007, 4th quartile	-0.000112	(0.00845)
exam year 2008, 1st quartile	0.0173	(0.0197)
exam year 2008, 2nd quartile	0.015	(0.00995)
exam year 2008, 3rd quartile	-0.0202**	(0.0093)
exam year 2008, 4th quartile	0.00327	(0.00659)
exam year 2009, 1st quartile	0.00943	(0.0205)
exam year 2009, 2nd quartile	0.00715	(0.0122)
exam year 2009, 3rd quartile	-0.0281**	(0.0115)
exam year 2009, 4th quartile	-0.0046	(0.0108)
exam year 2010, 1st quartile	0.0193	(0.0193)
exam year 2010, 2nd quartile	0.017	(0.0107)
exam year 2010, 3rd quartile	-0.0183*	(0.00968)
exam year 2010, 4th quartile	0.00523	(0.00905)
exam year 2011, 1st quartile	0.0215	(0.02)
exam year 2011, 2nd quartile	0.0192	(0.0117)
exam year 2011, 3rd quartile	-0.016	(0.0106)
exam year 2011, 4th quartile	0.00747	(0.0101)
Observations	1619	
*** p<0.01, ** p<0.05, * p<0.1		

One clear conclusion of this table is that the effect of rank on FCGPA is more positive at lower ranks relative to higher ranks. This appears to indicate that students

in higher rank quartiles (e.g. students who placed worse on the entrance exam) get more demotivated by a marginal increase in rank (i.e. being being ranked one place worse), so that the associated decrease in GPA is larger for these worse-ranked students relative to the better-ranked ones. Furthermore, the statistically significant rank coefficients are all in the third quartile, and are all negative, similar to what we found above in Table 3.7.

3.3.4 Quantity or Performance?

In the above analyses, we've focused on FCGPA as the relevant left hand variable, as it incorporates two different measures of academic performance: the number of units taken and the average grade received by enrolled students during their freshman year. In this section, we break down FCGPA into these two components of grade vs units, and analyze whether the results of Section 3.3.2 the previous analyses of rank on FCGPA operate via gpa or total units taken.

For this purpose, we now analyze two additional measures of academic performance: FGPA (the student's first year GPA, without adjusting for units taken), and the TNU, which is the total number of units accumulated by a student during their freshman year. We then regress these two outcomes measures using the same specification as in Equation 3.6. That is, we run:

$$\begin{aligned} y_i = &[1 \;\; D_i'] \times [\alpha + H_i\beta + T_i'\gamma + A_i\eta + rank_i\theta] \\ &+ [(rank_i - RQ1_i)^+, (rank_i - RQ2_i)^+, (rank_i - RQ3_i)^+]'\mu + \varepsilon_i. \end{aligned} \tag{3.7}$$

where y_i is either student i's FGPA or TNU. The results are described in Table 3.9.

From Table 3.9, it is clear that the effect of rank on FGPA is insignificant for most years and quartiles, with only the effect in 2008 for the 4th quartile being significant at the 5% level. On the other hand, the effect of rank on TNU is significantly negative in many years for the 3rd and 4th quartiles. This implies that a higher (i.e. worse) ranking doesn't cause students to do worse in their classes, but rather to take fewer classes.

As rank appears to decrease the number of units students take, but not their performance in each class, it follows that the effect of rank on FCGPA operates largely through TNU, and not through FGPA. We noted above in Table 3.8 that the effects of rank on FCGPA for most years and quartiles are negative, with the higher rank quart-

Table 3.9: Total effect of rank on FGPA (un-corrected first year GPA) and TNU (total number of units taken in first year), as in Equation 3.7. FGPA is largely unaffected by rank, whereas in many years and quartiles TNU is negatively impacted by rank.

	FGPA		TNU	
	estimate	se	estimate	se
exam year 2005, 1st quartile	0.00153	(0.00581)	-0.0107	(0.0432)
exam year 2005, 2nd quartile	-0.00104	(0.00301)	-0.0159	(0.0224)
exam year 2005, 3rd quartile	-0.00294	(0.00292)	-0.0866***	(0.0217)
exam year 2005, 4th quartile	-0.00195	(0.00207)	-0.0392**	(0.0154)
exam year 2006, 1st quartile	0.00441	(0.00547)	0.0253	(0.0407)
exam year 2006, 2nd quartile	0.00184	(0.00258)	0.0201	(0.0192)
exam year 2006, 3rd quartile	-6.32e-05	(0.00251)	-0.0506***	(0.0187)
exam year 2006, 4th quartile	0.000926	(0.00161)	-0.00324	(0.012)
exam year 2007, 1st quartile	0.00323	(0.00613)	0.0334	(0.0456)
exam year 2007, 2nd quartile	0.000665	(0.00329)	0.0282	(0.0245)
exam year 2007, 3rd quartile	-0.00124	(0.00317)	-0.0424*	(0.0236)
exam year 2007, 4th quartile	-0.000251	(0.00238)	0.00492	(0.0177)
exam year 2008, 1st quartile	0.00772	(0.00554)	0.0291	(0.0412)
exam year 2008, 2nd quartile	0.00515*	(0.00281)	0.0238	(0.0209)
exam year 2008, 3rd quartile	0.00324	(0.00262)	-0.0468**	(0.0195)
exam year 2008, 4th quartile	0.00423**	(0.00186)	0.000538	(0.0138)
exam year 2009, 1st quartile	0.00636	(0.00578)	0.0121	(0.043)
exam year 2009, 2nd quartile	0.00379	(0.00343)	0.00689	(0.0255)
exam year 2009, 3rd quartile	0.00189	(0.00325)	-0.0638***	(0.0241)
exam year 2009, 4th quartile	0.00288	(0.00306)	-0.0164	(0.0227)
exam year 2010, 1st quartile	0.00582	(0.00543)	0.0388	(0.0404)
exam year 2010, 2nd quartile	0.00325	(0.00301)	0.0336	(0.0223)
exam year 2010, 3rd quartile	0.00135	(0.00273)	-0.0371*	(0.0203)
exam year 2010, 4th quartile	0.00234	(0.00255)	0.0103	(0.019)
exam year 2011, 1st quartile	0.00436	(0.00564)	0.0435	(0.042)
exam year 2011, 2nd quartile	0.00179	(0.00331)	0.0383	(0.0246)
exam year 2011, 3rd quartile	-0.000113	(0.00299)	-0.0324	(0.0222)
exam year 2011, 4th quartile	0.000877	(0.00284)	0.015	(0.0211)
Observations	1,619		1,619	

*** p<0.01, ** p<0.05, * p<0.1

iles, in particular the 3rd quartile being most significantly negative. This pattern is mirrored in the effect of rank on TNU in Table 3.9. This suggests a possible mechanism through which rank is affecting student performance: being assigned to a lower rank

causes students to be somewhat more conservative about their academic ability, and as a result causes these students choose to focus attention on a few classes, thus focusing on a few classes instead of spreading themselves to thin.

3.3.5 Gender differences

We now turn to the question of gender differences: do male and female students react differently to rank when it comes to academic performance? To that end, we run the baseline regression above in Equation 3.3, but now with gender dummies. Precisely, we construct G_i, a length-2 vector, where the first entry is 1, and the second entry is an indicator for whether a student is male. Then, we interact this indicator with rank, and look at whether effect of rank on FCGPA differs across gender. This specification corresponds to Equation 3.8, and the results are listed in Table 3.10.

$$FCGPA_i = [1 \;\; D_i'] \times [\alpha + H_i\beta + T_i'\gamma + A_i\eta] + rank_i \times G_i'\nu + \varepsilon_i. \tag{3.8}$$

Table 3.10: Coefficient and standard errors of ν from the regression in Equation 3.8. There does not appear to be any differences in response of FCGPA to rank between genders.

	Dependent variable:
	FCGPA
rank	-0.011^{***}
	(0.002)
rank * male	0.0004
	(0.002)
Observations	1,619
Note:	*p<0.1; **p<0.05; ***p<0.01

As is apparent, the coefficient on male is insignificant, suggesting that there are not any significant gender differences in students' FCGPA response to ranking.

Now that FCGPA doesn't appear to be differentially affected by rank across gender, we may further break down FCGPA into the two components of FGPA and TNU, and ask whether these two measures of academic performance exhibit different impacts from

rank. To do so, we run the same regression as Equation 3.8, except with the left hand variable replaced with FGPA or TNU:

$$y_i = [1 \;\; D_i'] \times [\alpha + H_i\beta + T_i'\gamma + A_i\eta] + rank_i \times G_i'\nu + \varepsilon_i. \tag{3.9}$$

where y_i is FGPA or TNU. The results of this regression are in Table 3.11. There appears no significant differential between male and female students on both total number of units vs GPA. Students of both genders appear to adjust primarily on TNU, and not so much on GPA.

Table 3.11: Coefficient estimates and standard errors of ν from the regression in Equation 3.9, for left hand variables of FGPA and TNU. Neither FGPA nor TNU seem to exhibit any differential response to rank between genders.

	Dependent variable:	
	FGPA	TNU
	(1)	(2)
rank	0.00003	−0.024***
	(0.001)	(0.005)
rank * male	0.0001	−0.00001
	(0.0004)	(0.003)
Observations	1,619	1,619
Note:	*p<0.1; **p<0.05; ***p<0.01	

3.4 Robustness checks

Throughout the main analyses performed in Section 3.3, we've made the assumption that linearly accounting for the control covariates (subtest scores, high school grade, and age) leaves us with the a residual on rank that is orthogonal to the best predictive model of academic performance give the non-rank covariates. This assumption is made as a first attempt at capturing the intuition that academic performance should vary smoothly with the listed covariates, but is likely too strong of an assumption to be true. That is, while it's reasonable to expect that the best predictor should vary smoothly

as a function of the control covariates, it's less reasonable to assume that this function is exactly linear. Thus, it is necessary to show that the main results of Section 3.3 are not merely artifacts generated by this linearity assumption. We do so in this section.

3.4.1 Quadratic controls

One very straightforward way to see if the various estimates in Section 3.3 are actually real or illusions due to using only linear controls is to perform the exact same analysis, but with higher order controls. This relaxes the assumption made previously about the smoothness of the best predictor of FCGPA as a function of the control variables: instead of assuming that the best predictor must be linear, we instead assume that it may have higher order terms.

In this section, we do this by using quadratic controls. To be precise: in the work of Section 3.3, we used the following controls:

$$[1 \;\; D_i'] \times [\alpha + H_i\beta + T_i'\gamma + A_i\eta]$$

where H_i high school graduation grade, T_i the scores on each of the 4 sub-tests making up the entrance exam, A_i age, and D_i indicators for the various years in our data sample. In this section, we relax the linearity assumption by allowing for all squares and interaction terms, which produces the following quadratic controls:

$$QC_i := [1 \;\; D_i'] \times \left[1, H_i, T_i, A_i, H_i^2, T_i^2, A_i^2, H_iT_i, H_iA_i, T_iA_i\right]'\gamma$$

Where the T_i^2 consists of all square and interactions between the 4 subtest scores, H_iT_i consists of all the interactions of H_i with the various subtest scores, and likewise for T_iA_i. With this, we the proceed to re-do the main analyses of Section 3.3.

We first run the analogue of the base rank regression of Equation 3.3, except now with quadratic controls:

$$FCGPA_i = QC_i + rank_i\theta + \varepsilon_i. \tag{3.10}$$

The results are listed in Table 3.12. As before, the mean effect of rank on FCGPA is negative and significant at 5%.

We now turn to the effect of rank on FCGPA for each quartile, replicating the regression of Equation 3.5, except with quadratic controls:

Table 3.12: Basic regression of FCGPA on rank, now with quadratic controls as in Equation 3.10. The effect of rank is still negative and significant.

	Dependent variable:
	FCGPA
rank	−0.0263***
	(0.00564)
Observations	1,619
Note:	*p<0.1; **p<0.05; ***p<0.01

$$FCGPA_i = QC_i + rank_i\theta + [(rank_i - RQ1_i)^+, (rank_i - RQ2_i)^+, (rank_i - RQ3_i)^+]'\mu + \varepsilon_i. \quad (3.11)$$

The results are detailed in Table 3.13. We find that just as with linear controls, effect of rank on FCGPA is very significantly negative in the third quartile. In addition, with this specification, we further find that the effect is significantly negative in the fourth quartile, whereas under linear controls this effect was insignificant.

Table 3.13: Total effect of rank on FCGPA, for each quartile, now with quadratic controls as in in Equation 3.11. The effect is significantly negative in both the third and fourth quartiles at 5%.

	FCGPA	se
1st quartile	-0.00337	(0.0201)
2nd quartile	-0.00394	(0.00982)
3rd quartile	-0.0301***	(0.00855)
4th quartile	-0.0172**	(0.00708)
Observations	1619	
*** p<0.01, ** p<0.05, * p<0.1		

We may also re-visit how the effect of rank on FCGPA varies with gender by running the following regression, which is analogous to Equation 3.8 except with quadratic

controls:

$$FCGPA_i = QC_i + rank_i \times G_i'\nu + \varepsilon_i. \tag{3.12}$$

The results of this regression can be found in Table 3.14, where we, as before, find that the two genders appear to respond in similar ways to rank.

Table 3.14: Effect of rank on gender, difference between male and female students, from the regression in Equation 3.12. There does not appear to be any differences in response of FCGPA to rank between genders.

	Dependent variable:
	FCGPA
rank	-0.0254^{***}
	(0.00563)
rank * male	-0.00125
	(0.0017)
Observations	1,619
Note:	*p<0.1; **p<0.05; ***p<0.01

Finally, we investigate whether the previous results about FGPA vs TNU apply as well with quadratic controls. Thus, we run the specification of Equation 3.7 with quadratic controls, as follow:

$$\begin{aligned} y_i =& QC_i + [1 \;\; D_i'] \times rank_i\theta \\ &+ [(rank_i - RQ1_i)^+, (rank_i - RQ2_i)^+, (rank_i - RQ3_i)^+]'\mu + \varepsilon_i. \end{aligned} \tag{3.13}$$

where y_i is either student i's FGPA or TNU. The results are described in Table 3.15. We find that the results agree broadly with the what we found previously: rank has little to no effect on freshman GPA in all years and quartiles, while it has a significant negative effect on total number of units taken in the first year, especially in earlier years and in the latter quartiles. However, in addition to this, we also find that, with quadratic controls, the negative effects on TNU is significant in 2005 even for the first two quartiles.

Thus, using quadratic controls in place of linear ones don't seem to particularly affect the main results of Section 3.3, suggesting that these effects are not purely a

Table 3.15: Total effect of rank on FGPA (un-corrected first year GPA) and TNU (total number of units taken in first year), with quadratic controls as in Equation 3.13. FGPA is largely unaffected by rank, whereas in many years and quartiles TNU is negatively impacted by rank.

	FGPA		TNU	
	estimate	se	estimate	se
exam year 2005, 1st quartile	-0.00509	(0.00689)	-0.113**	(0.0509)
exam year 2005, 2nd quartile	-0.00594	(0.00483)	-0.109***	(0.0357)
exam year 2005, 3rd quartile	-0.00788	(0.00479)	-0.171***	(0.0354)
exam year 2005, 4th quartile	-0.00653	(0.00434)	-0.134***	(0.0321)
exam year 2006, 1st quartile	0.00304	(0.00603)	-0.0219	(0.0446)
exam year 2006, 2nd quartile	0.00219	(0.00356)	-0.0173	(0.0264)
exam year 2006, 3rd quartile	0.000247	(0.00336)	-0.0796***	(0.0248)
exam year 2006, 4th quartile	0.0016	(0.00305)	-0.0426*	(0.0225)
exam year 2007, 1st quartile	0.00332	(0.00744)	0.0077	(0.055)
exam year 2007, 2nd quartile	0.00247	(0.00527)	0.0123	(0.039)
exam year 2007, 3rd quartile	0.000528	(0.00502)	-0.05	(0.0371)
exam year 2007, 4th quartile	0.00188	(0.00485)	-0.013	(0.0358)
exam year 2008, 1st quartile	0.0106	(0.00724)	-0.0182	(0.0535)
exam year 2008, 2nd quartile	0.0097*	(0.00542)	-0.0136	(0.0401)
exam year 2008, 3rd quartile	0.00776	(0.00508)	-0.0758**	(0.0376)
exam year 2008, 4th quartile	0.00911*	(0.00499)	-0.0389	(0.0369)
exam year 2009, 1st quartile	0.000176	(0.00724)	-0.0068	(0.0535)
exam year 2009, 2nd quartile	-0.000674	(0.00527)	-0.00222	(0.039)
exam year 2009, 3rd quartile	-0.00261	(0.00516)	-0.0645*	(0.0382)
exam year 2009, 4th quartile	-0.00126	(0.00502)	-0.0275	(0.0371)
exam year 2010, 1st quartile	0.00459	(0.00661)	0.0387	(0.0489)
exam year 2010, 2nd quartile	0.00374	(0.00453)	0.0432	(0.0335)
exam year 2010, 3rd quartile	0.0018	(0.00433)	-0.019	(0.032)
exam year 2010, 4th quartile	0.00315	(0.00422)	0.018	(0.0312)
exam year 2011, 1st quartile	0.0119	(0.0075)	0.0498	(0.0555)
exam year 2011, 2nd quartile	0.011*	(0.00581)	0.0543	(0.043)
exam year 2011, 3rd quartile	0.00911*	(0.00551)	-0.00792	(0.0408)
exam year 2011, 4th quartile	0.0105*	(0.00545)	0.0291	(0.0403)
Observations	1,619		1,619	

*** p<0.01, ** p<0.05, * p<0.1

consequence of misspecification.

3.4.2 Placebo regressions

In Section 3.3, we showed two statistically significant effects, the first being that the mean effect of rank on FCGPA is negative, and the second being that this effect is concentrated on students in the third quartile of rank. Here, we perform some placebo tests to demonstrate that these two results are not due purely to higher-order correlations between FCGPA and residual of rank after linearly controlling for the various covariates.

We perform the placebo tests as follows:

1. Consider the 7 covariates (total score, culture score, biology score, chemistry score, physics score, high school graduation grade, age)

2. Consider all 5040 permutations of the names of these 7 covariates (e.g. so maybe 'age' now refers to what is actually 'total score', and 'high school graduation grade' now refers to what is actually 'physics score', and so forth)

3. Exclude permutations that map 'total score' onto itself, leaving us with 4320 permutations.

4. For each of these 4320 remaining permutations:

   (a) Construct a placebo-rank, using these new permuted attribute names, as described in Procedure 6.

   (b) Run the regressions described in Equations 3.3, and record the p-value of the coefficient on rank.

   (c) Run the regressions described in Equations 3.5, construct the total effect of rank on FCGPA for each quarter, and record the p-value of this effect for each quarter.

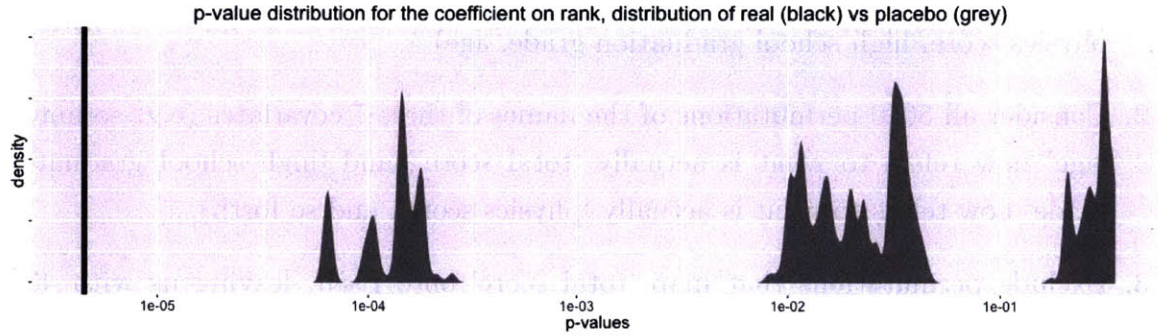
5. Plot the true p-value of the coefficient on rank (from Equations 3.3), along with the distribution of placebo p-values generated over the 4320 permutation runs.

6. Plot the true p-values of the total effect of rank on FCGPA for each quartile (from Equations 3.5), along with the distribution of placebo p-values generated over the 4320 permutation runs.

The intuition for performing the placebo tests is this: each placebo rank is still a function of the control covariates, so that for any placebo rank, we expect there

to be some left-over correlation between this placebo rank and FCGPA after linearly controlling for the control covariates. We have no ex-ante reason to suspect that this left-over correlation is particularly strong for the real rank vs the placebo ranks. Thus, if it were the case that our significant effects in Section 3.3 were consequences purely of this left-over higher order correlation, we would expect these effects to be equally significant for each of the placebo ranks as for the real rank.

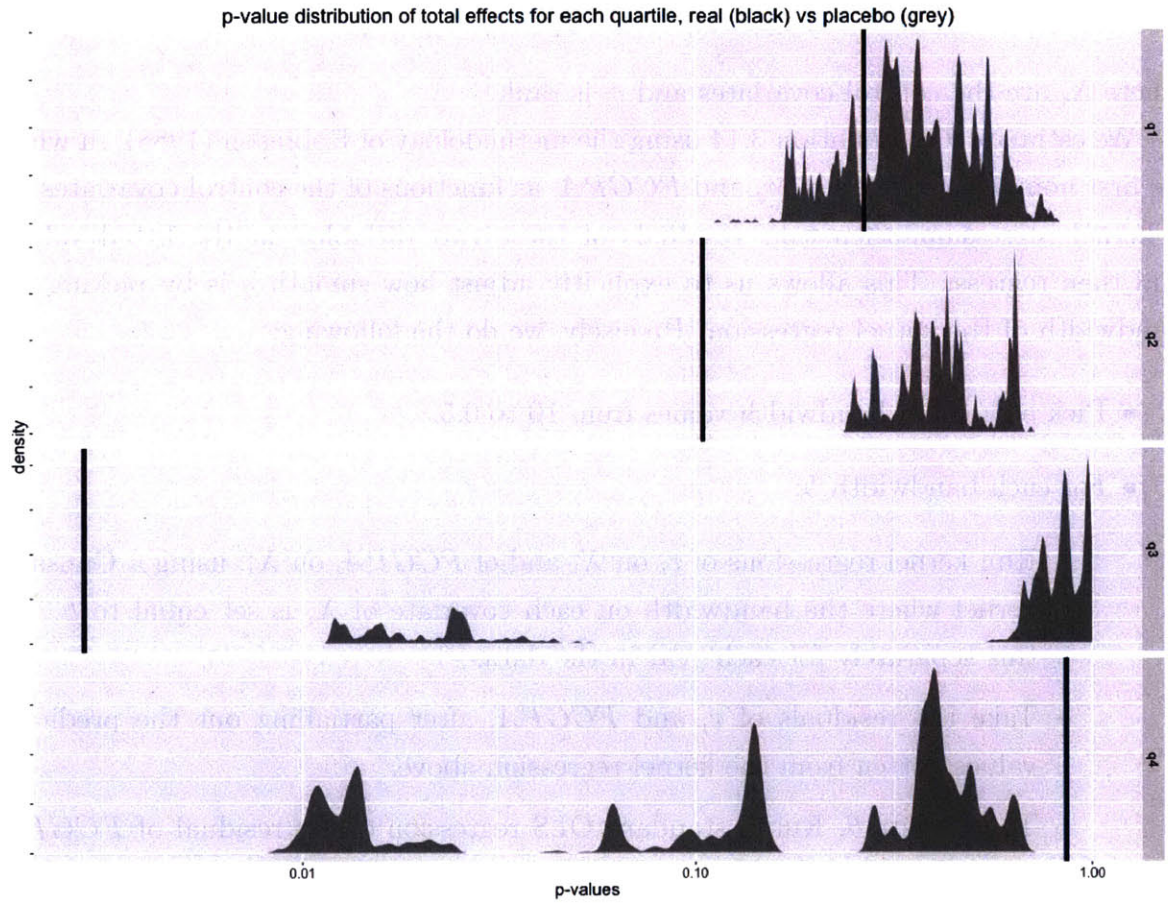
Figure 3-3: Distribution of p-values of the coefficient on rank from the regression in Equation 3.3, real vs placebo. The real p-value is orders of magnitude smaller than the placebo p-values.



The placebo vs real p-value graph for the baseline regression of Equation 3.3 is presented in Figure 3-3. As is apparent, the p-value on rank from using the true rank is much lower by an order of magnitude than the lowest p-value on rank from using any placebo rank. Thus, this suggests that the significance of the coefficient on rank in Equation 3.3 is not purely an artefact of using only linear controls.

Similarly, the placebo vs real p-value graph for the rank-quartile regression of Equation 3.5 is presented in Figure 3-4. Previously, using the true rank, we found that the effect of rank on FCGPA was significant and negative for the third quartile, whereas the effects where not significant for the other quartiles. As the figure indicates, for the third quartile, the p-value of the effect of rank on FCGPA is much smaller than any of the p-values gotten from using placebo rank, again suggesting that this effect is real rather than a statistical error. Curiously, the real p-value of the effect on rank on FCGPA for the second quartile is also much lower than any of the placebo ones, indicating that, while this effect is not significant, the effect is still much more significant than any of the placebo effects.

Figure 3-4: Distribution of p-values of the total effect of rank on FCGPA, for each rank quartile, from the regression in Equation 3.5, real vs placebo. The real p-value for the third quartile is orders of magnitude smaller than the placebo p-values.



3.4.3 Semiparametric regression

In Sections 3.3 and 3.4.1, we restricted the smoothness of the best predictor of FCGPA as a function of the control covariates through parametric assumptions. In both of those cases, the effect of rank on FCGPA was found to be significant and negative. In this section, we try to further understand just how non-smooth this best predictor function has to be in order for the correlation between rank and FCGPA to become insignificant. To that end, we run a semiparametric regression with a wide range of smoothness assumptions about the best predictor g of FCGPA as a function of covariates:

$$FCGPA_i = \theta_0 + r_i\theta + g(X_i) + \varepsilon_i \tag{3.14}$$

where X_i are the control covariates and r_i is rank.

We estimate θ in Equation 3.14 using the methodology of Robinson (1988), in which we first nonparametrically fit r_i and $FCGPA_i$ as functions of the control covariates via a kernel regression, then take residuals of these two variables on the fitted values, and then regress. This allows us to explicitly adjust how smooth g is by picking the bandwidth of the kernel regression. Precisely, we do the following:

- Pick a range of bandwidth values from 10 to 0.5.
- For each bandwidth h:
  - Run kernel regressions of r_i on X_i and of $FCGPA_i$ on X_i, using a Gaussian kernel where the bandwidth on each covariate of X_i is set equal to h. Do this separately for each year in the data.
  - Take the residuals of r_i and $FCGPA_i$ after partialling out the predicted values gotten from the kernel regression above.
  - To estimate θ, Run a standard OLS regression of the residual of $FCGPA_i$ on the residual of r_i.
- Produce confidence intervals by generating bootstrap draws of the data and re-running the entire above process.

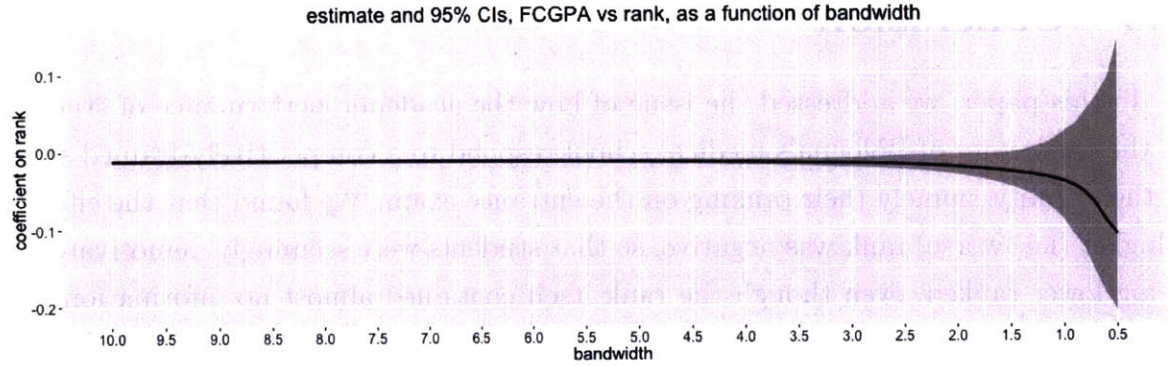
By varying the bandwidth parameter h, we may pick more or less smooth mappings of the control covariates to $FCGPA$, thus allowing us to see how the estimated coefficients vary with smoothness of the g function. Naturally, large values of the bandwidth

correspond to much smoother functions of rank as a function of control covariates, and in such cases we would expect to have enough residual variation in rank left over after partialling out the fitted values to precisely estimate θ. On the other hand, if the bandwidth is low, then rank will be fitted as a highly non-smooth function of the control covariates, thus leaving little residual variation in rank and FCGPA.

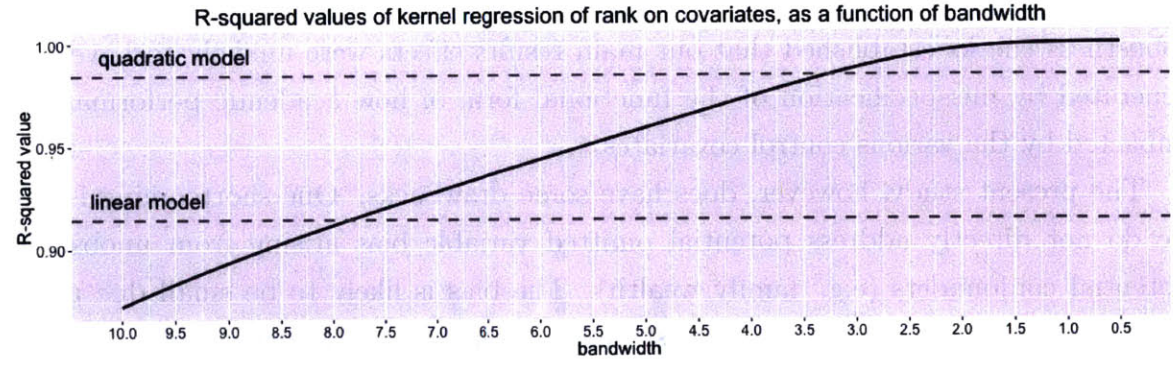
We plot the estimates of θ along with the associated confidence intervals in Figure 3-5a. As the figure shows, θ is precisely estimated for all values of the bandwidth that are

Figure 3-5

(a) Semiparametric estimates of the effect on rank on FCGPA from Equation 3.14, for a variety of bandwidth values. Confidence bands computed over 500 bootstrap runs. The effect is significantly negative for values of the bandwidth that are not too small.



(b) R^2 values from the kernel regression of rank on the control covariates, for a variety of bandwidth values. For bandwidth values below 2, the R^2 is extremely close to 1. The R^2 values from the linear and quadratic regressions of rank on the control covariates are plotted for comparison purposes.



not too small. The effect is significantly negative for all values of bandwidth larger than 1.4. This bandwidth value of 1.4 amounts to an assumption of extreme non-smoothness of the g function: performing the kernel regression of the control covariates on rank using a bandwidth value of 1.4 yields an R^2 value of 0.9996, which is much larger than

the R^2 value of the linear regression of rank on the covariates (about 0.9144), and the R^2 value of the quadratic regression of rank on the covariates (about 0.9845). The entire path of R^2 values, along with comparisons to the linear and quadratic R^2 values is presented in Figure 3-5b.

Thus, this analysis gives us more assurance that the average effect of rank on FCGPA is indeed significantly negative. For no values of the bandwidth is the effect positive, and it is significantly negative for all non-extreme values of bandwidth. To get an insignificant effect, one would have to argue that the best predictor of $FCGPA$ as a function of the controls covariates is so non-smooth that only 0.0004 fraction of the residual variation of rank is leftover after accounting for this function of the controls.

3.5 Conclusion

In this paper, we addressed the issue of how the academic performance of students at the University of Bologna's medical school respond to a competitively-framed signal of their ability, namely their ranking on the entrance exam. We found that the effect of a higher (i.e. worse) rank was negative, so that students were seemingly demotivated by being lower ranked, even though the rank itself contained almost no information that wasn't already available through other means. Furthermore, we found that this demotivational effect primarily affects students ranked in the third quartile on the entrance exam, equally affects male and female students, and operates by affecting courseload rather than GPA. The main results of the paper, which were obtained through OLS, were found to be robust to a number of alternative specifications. In particular, these robustness checks established that our main results effects were unlikely to have been generated by misspecification of the functional form of how academic performance is impacted by the various control covariates.

The present study, however, does have some drawbacks. One shortcoming is that we do not directly address potential omitted variable bias arising from unobserved potential confounders (e.g. family wealth). The bias is likely to be small due purely to the fact that we already observe so many indicators of student's prior academic performance, though of course we can't rule it out entirely. Ideally, we would like to find an instrument for rank, and use that to estimate causal impact, but we are currently unaware of any such instruments. Another limitation of the current study is that we've focused on estimating the mean impact of rank on academic performance, with some

limited heterogeneity for gender and different rank quartiles. It may be worth estimating the treatment effect at each level of rank, e.g. by doing some sort of propensity score weighting where we model rank as a continuous treatment variable. Finally, we've interpreted the effect of rank on academic performance as a psychological effect, where students who get ranked worse feel demotivated and thus perform worse academically. While some sort of psychological channel seems like the only plausible mechanism by which rank could impact academic performance, a more thorough investigation may be worthwhile, and could possibly be carried out e.g. through surveys designed to elicit the students' level of motivation and their amount of effort put into studying for exams.

Bibliography

Carole Ames and Jennifer Archer. Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of educational psychology*, 80 (3):260, 1988.

Eric M Anderman and Allison J Young. Motivation and strategy use in science: Individual differences and classroom effects. *Journal of research in science teaching*, 31 (8):811–831, 1994.

Steffen Andersen, Seda Ertac, Uri Gneezy, John A List, and Sandra Maximiano. Gender, competitiveness, and socialization at a young age: Evidence from a matrilineal and a patriarchal society. *Review of Economics and Statistics*, 95(4):1438–1443, 2013.

Glenn Ellison and Ashley Swanson. The gender gap in secondary school mathematics at high achievement levels: Evidence from the american mathematics competitions. *The Journal of Economic Perspectives*, 24(2):109–128, 2010.

Uri Gneezy and Aldo Rustichini. Gender and competition at a young age. *The American Economic Review*, 94(2):377–381, 2004.

Uri Gneezy, Muriel Niederle, Aldo Rustichini, et al. Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, 118(3):1049–1074, 2003.

Uri Gneezy, Kenneth L Leonard, and John A List. Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, 77(5):1637–1664, 2009.

Michael Inzlicht and Talia Ben-Zeev. A threatening intellectual environment: Why

females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science*, 11(5):365–371, 2000.

Judith L Meece, Eric M Anderman, and Lynley H Anderman. Classroom goal structure, student motivation, and academic achievement. *Annu. Rev. Psychol.*, 57:487–503, 2006.

Muriel Niederle and Lise Vesterlund. Explaining the gender gap in math test scores: The role of competition. *The Journal of Economic Perspectives*, 24(2):129–144, 2010.

Muriel Niederle and Lise Vesterlund. Gender and competition. *Annu. Rev. Econ.*, 3 (1):601–630, 2011.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

Robert W Roeser and Jacquelynne S Eccles. Adolescents' perceptions of middle school: Relation to longitudinal changes in academic and psychological adjustment. *Journal of Research on Adolescence*, 8(1):123–158, 1998.