8-2018

Essays on Structural Econometric Modeling and Machine Learning

Hajime Shimao
Purdue University

ESSAYS ON STRUCTURAL ECONOMETRIC MODELING AND MACHINE LEARNING

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Hajime Shimao

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2018

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Ralph Siebert, Chair
Krannert School of Management
Dr. Mohitosh Kejriwal
Krannert School of Management
Dr. Justin Tobias
Krannert School of Management
Dr. Stephen Martin
Krannert School of Management
Dr. Joe Mazur
Krannert School of Management

Approved by:

Dr. Brian Roberson, Head of the Graduate Program
Krannert School of Management
Dr. Ralph Siebert, Chair
Krannert School of Management

To Xiaoxiao Li.

ACKNOWLEDGMENTS

I would like to thank my committee members, Ralph Siebert, Mohitosh Kejriwal, Justin Tobias, Stephen Martin, and Joe Mazur for their advice and guidance on my research. I very much appreciate my friend and the co-author of the first two chapters, Junpei Komiyama at University of Tokyo, for his great help and our daily discussion which lead to the research agenda of this dissertation. I would also like to thank the two co-authors of the third chapter, Warut Khern-am-nuai and Kannan Karthik. I am extremely grateful to my parents for all the support and understanding.

Finally, I would like to express my deepest gratitude to my wife, Xiaoxiao Li, for her encouragement, belief in my capability, and everything we had in this four years.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Shimao, Hajime PhD, Purdue University, August 2018. Essays on Structural Econometric Modeling and Machine Learning. Major Professor: Ralph Siebert.

This dissertation is composed of three independent chapters relating the theory and empirical methodology in economics to machine learning and important topics in information age . The first chapter raises an important problem in structural estimation and provide a solution to it by incorporating a culture in machine learning. The second chapter investigates a problem of statistical discrimination in big data era. The third chapter studies the implication of information uncertainty in the security software market.

Structural estimation is a widely used methodology in empirical economics, and a large class of structural econometric models are estimated through the generalized method of moments (GMM). Traditionally, a model to be estimated is chosen by researchers based on their intuition on the model, and the structural estimation itself does not directly test it from the data. In other words, not sufficient amount of attention is paid to devise a principled method to verify such an intuition. In the first chapter, we propose a model selection for GMM by using cross-validation, which is widely used in machine learning and statistics communities. We prove the consistency of the cross-validation. The empirical property of the proposed model selection is compared with existing model selection methods by Monte Carlo simulations of a linear instrumental variable regression and oligopoly pricing model. In addition, we propose the way to apply our method to Mathematical Programming of Equilibrium Constraint (MPEC) approach. Finally, we perform our method to online-retail sales data to compare dynamic model to static model.

In the second chapter, we study a fair machine learning algorithm that avoids a statistical discrimination when making a decision. Algorithmic decision making process now affects many aspects of our lives. Standard tools for machine learning, such as classification and

regression, are subject to the bias in data, and thus direct application of such off-the-shelf tools could lead to a specific group being statistically discriminated. Removing sensitive variables such as race or gender from data does not solve this problem because a *disparate impact* can arise when non-sensitive variables and sensitive variables are correlated. This problem arises severely nowadays as bigger data is utilized, it is of particular importance to invent an algorithmic solution. Inspired by the two-stage least squares method that is widely used in the field of economics, we propose a two-stage algorithm that removes bias in the training data. The proposed algorithm is conceptually simple. Unlike most of existing fair algorithms that are designed for classification tasks, the proposed method is able to (i) deal with regression tasks, (ii) combine explanatory variables to remove reverse discrimination, and (iii) deal with numerical sensitive variables. The performance and fairness of the proposed algorithm are evaluated in simulations with synthetic and real-world datasets.

The third chapter examines the issue of information uncertainty in the context of information security. Many users lack the ability to correctly estimate the true quality of the security software they purchase, as evidenced by some anecdotes and even some academic research. Yet, most of the analytical research assumes otherwise. Hence, we were motivated to incorporate this "false sense of security" behavior into a game-theoretic model and study the implications on welfare parameters. Our model features two segments of consumers, well- and ill-informed, and the monopolistic software vendor. Well-informed consumers observe the true quality of the security software, while the ill-informed ones overestimate. While the proportion of both segments are known to the software vendor, consumers are uncertain about the segment they belong to. We find that, in fact, the level of the uncertainty is not necessarily harmful to society. Furthermore, there exist some extreme circumstances where society and consumers could be better off if the security software did not exist. Interestingly, we also find that the case where consumers know the information structure and weight their expectation accordingly does not always lead to optimal social welfare. These results contrast with the conventional wisdom and are crucially important in developing appropriate policies in this context.

1. CROSS VALIDATION BASED MODEL SELECTION VIA GENERALIZED METHOD OF MOMENTS

1.1 Introduction

Structural estimation of economic models is one of the most widely used methodologies in empirical economics nowadays in variety of fields.Structural estimation enables researchers to interpret latent variable, as well as it allows researchers to perform counterfactual simulations. Arguably, however, one of the largest shortcoming in the structural estimation procedure lies in the selection of a proper model. That is, the specification of estimation models is usually chosen by researchers and implementation of structural estimation itself does not directly address on it from the data, because the estimation is performed by assuming the model reflects the true data generating process ((Angrist and Pischke, 2010)). On a paper it is a common practice for economists to verbally argue and defend their model specification in a descriptive way. However, since the validity of the counterfactual simulation crucially depends on the goodness of the model, verifying and choosing a proper model empirically is of particular importance. Especially, we often simplify a model for the ease of tractability: Such simplifications is preferred to be subject to some assessment.

When a structural model is estimated in economics, researchers often use generalized method of moments (GMM) as well as maximum likelihood. As to selecting a true model, (Smith, 1992) and (Rivers and Vuong, 2002) offer a model selection procedure for GMM based on the difference of empirical moments. Their core idea is a simple use of the GMM minimand as a fitness of the model with the observed data: That is, to select the model of the smallest GMM minimand when it is estimated[1]. Although such a procedure is

[1]The theory provided in (Rivers and Vuong, 2002) applies to broader range of model selection criteria. However, it is often implemented as GMM minimand comparison. See (Bonnet and Dubois, 2010) or (Berto Villas-Boas, 2007) for example.

asymptotically consistent in choosing a true or ”better” model, the performance of model selection with limited sample size is still uncertain. In some applications, economists have to make an inference from a relatively small number of observations. Given a limited size of the sample, their procedures may be subject to ”over-fitting”: excessively complicated models can fit tighter to the observations in hand with better ”goodness-of-fit” criterion, and thus is selected as a better model even if the model is not very true.

To avoid over-fitting problem, some model selection criteria such as AIC-GMM or BIC-GMM ”penalize” the number of parameters in a model ((Andrews, 1999)). However, the complexity of economic models is not simply measured by the number of parameters. Structural model may include non-parametric components in specification (e.g., (Gautier and Kitamura, 2013)), where we cannot apply a penalization based on number of parameters. Additionally, estimation procedure sometimes involves nonparametric approximation only for certain models. For example, estimation of dynamic demand model in (Gowrisankaran and Rysman, 2012) includes a nonparametric approximation of a value function, which may make their model more flexible than static demand model. To date, it is not well understood how these factors contribute to the over-fitting issue nor how to penalize its flexibility.

In this paper, we offer a novel approach to this problem that helps researchers to identify the best model specification from the data. Our idea is to apply the cross-validation (CV) method, which is commonly used in other areas such as machine learning, in evaluating the predictive power of the model. The main idea behind cross-validation is to split the data into several portions so that test of a model fit is implemented on a different data from the one used for estimating parameters. As a result, the estimated moment suffers a smaller over-fitting than in-sample model selection.

The largest advantage of sample splitting lies in its wide range of potential applications. On applying CV, one does not need to take the number of model parameters explicitly. As a result, it can select the true model among parametric, non-parametric and even semi-parametric models. Moreover, CV can be applied not only in selecting models, but also selecting hyper-parameters of estimation and even estimation method itself. For example,

estimation of dynamic model often includes approximation of value function on a discrete grid space, where the coarseness of the grid space has not been paid adequate attention though it heavily influences the performance of estimation. As to the example of estimation method, random coefficient demand system can be estimated in various specifications, such as parametric or non-parametric, through various methodologies such as nested fixed point algorithm or constrained optimization approach (MPEC, (Su and Judd, 2012)) and they may yield different results especially in limited sample size.

Economists typically evaluate estimation techniques and model specification by checking how the true parameters are recovered in a Monte-Carlo simulation. However, the best specification or methodology may vary across different data or the ”true” data generating process that researchers do not observe. Thus, it is preferable to make an assessment in real-world data as well, and CV offers a practical approach to that end. Taking a wide range of applications into consideration, conducting CV in selecting models deserves a significant portion of attention.

Although CV is commonly used in data science fields such as machine learning and data mining, its applicability to economic models is not obvious. In machine learning and data mining, the primal concern lies in how accurate the prediction of a regressor or classifier is. Meanwhile, in empirical economics, identifying the model reflecting the reality closer and estimating its model parameters are of primal concern, and machine learning literature does not provide a sufficient guarantee in identification of a model. This gap remains to be closed in applying data science methods in econometrics. Taking this into consideration, we propose an identifiable CV method for GMM.

We first prove the consistency of cross-validation algorithm: That is, the algorithm identifies a correctly specified model from misspecified models with the probability approaching to 1 as the number of data increases. When a model is estimated through likelihood maximization, (Yang, 2007) proved the consistency of the cross-validation in non-parametric regression model selection. We prove an analogous result for GMM version of CV algorithm.

After giving the consistency, we test the performance of our cross-validation algorithm with a limited number of samples by Monte-Carlo simulation. Firstly, we examine a simple instrumental variable regression. We observe our algorithm selects a correctly specified mode over a misspecified model with high probability even when data size is limited. Importantly, our algorithm finds the correctly specified model even when the alternative model has higher flexibility (i.e., more parameters) than the true model, suggesting that it is robust to over-fitting. Furthermore, we compare the performance of our algorithm with Rivers-Vuong type GMM minimand comparison approach and also approaches based on GMM-AIC and GMM-BIC criteria that (Andrews, 1999) suggested. The result implies that the comparison of GMM minimand suffers over-fitting, and as a result it often selects a misspecified model of higher complexities. Though GMM-AIC and GMM-BIC based approaches attempt to solve the over-fitting problem by penalizing the flexibility of model, their performance turns out to be extremely sensitive to the model specification, and as a result, they often fail to find the correctly specified model.

Secondly, we conduct another experiment in more complex nonlinear models. We use a collusive pricing model similar to the ones of (Bresnahan, 1987) and (Hu et al., 2014), where their objective of model selection is to detect a potential tacit collusion from the sales and price data. We simulate the price and quantity data from perfectly competitive setting and partially collusive setting, and test if our algorithm discovers the true conduct or not. We show that our cross-validation procedure generally perform well to identify the true pricing structure from a limited amount of data. We show how CV outperforms the simple GMM fitting comparison without data split.

In addition, we propose a method to apply cross-validation algorithm when estimation is based on Mathematical Programming of Equilibrium Constraint (MPEC) approach. MPEC is proposed by (Su and Judd, 2012) and is one of the state-of-the-art estimation methodologies. MPEC achieves high computational efficiency by avoiding the nested fixed point algorithm, and its convenience is earning significant attention especially in the industrial organization research community. Though application of CV to MPEC is not straightforward, we provide a modified algorithm of CV applicable to MPEC estimation.

Finally, we perform our algorithm on a cutting-edge structural model with real-world data. The model we adopt is dynamic demand and dynamic pricing model of (Conlon, 2012). The dynamic models are considered to be the recent frontier of the industrial organization community and used in many applications (such as (Lee, 2013)). However, the superiority of the dynamic models compared with static models on its explainability of the consumer behavior is not sufficiently supported. Likewise, the dynamic pricing model is a frontier research topic in the industrial organization ((Nair, 2007),(Luo, 2015)), but its empirical support against static model is only descriptive. We apply our CV algorithm to the market data of an online retailer based in the UK to test dynamic models against static models. We show that the results are mixed across different products, even though they are sold by the same retailer.

The paper proceeds as follows. In Section 2, we formally introduce cross-validation in GMM and discuss its econometric property. In particular, we prove its asymptotic consistency. In Section 3, we demonstrate a Monte-Carlo experiment of model selection in IV regression. In Section 4, we perform a further experiment in an oligopolistic pricing model as a nonlinear example. Section 5 explains how we can modify the algorithm when it is applied to MPEC approach. Section 6 presents the setup and results of the real-world application of the dynamic pricing model using online-retailer data. Section 7 concludes the paper.

1.2 Cross-validation Approach to GMM Model Selection

1.2.1 Setup

Let $\mathbf{v} = \{v_t\}$ be a random vector of observed data in $\mathbf{V} \subset R^d$. Let $\mathcal{M}_i$ for $i = 1, 2$ be the two candidate models to explain the observed data. Each model, if correctly specified, is characterized by a set of moment conditions $f^{(i)} : \mathbf{V} \times \Theta^{(i)} \to \mathcal{R}^{q_i}$ such that

$$\mathcal{M}_i \Rightarrow E[f^{(i)}(v_t, \theta_0^{(i)})] = 0 \text{ for a unique } \theta_0^{(i)} \in \Theta^{(i)}$$

where $\theta^{(i)} \in \Theta^{(i)}$ denotes the parameters of a model i to be estimated. Let p_i be the dimension of $\theta^{(i)}$. Given the observation $\{v_t\}_{t=1,,,T}$, the parameters of each model are estimated via GMM;

$$\hat{\theta}_T^{(i)} = \underset{\theta^{(i)} \in \Theta^{(i)}}{\arg\min}\, Q_T^{(i)}(\theta^{(i)}) \tag{1.1}$$

where

$$Q_T^{(i)}(\theta^{(i)}) = \quad \frac{1}{T} \sum_{t=1}^{T} f^{(i)}(v_t, \theta^{(i)}) \quad W_T^{(i)} \quad \frac{1}{T} \sum_{t=1}^{T} f^{(i)}(v_t, \theta^{(i)}) \quad .$$

Let $\underset{T\to\infty}{\text{plim}} W_T^{(i)} = W^{(i)}$, and the population analogue of the moment conditions be

$$Q_0^{(i)}(\theta^{(i)}) = E[f^{(i)}(v_t, \theta^{(i)})]\, W^{(i)} E[f^{(i)}(v_t, \theta^{(i)})].$$

Assume that $\underset{T\to\infty}{\text{plim}} \theta_T^{(i)} = \theta_0^{(i)}$ exists. The null hypothesis is that $\mathcal{M}_1$ and $\mathcal{M}_2$ are asymptotically equivalent;

$$\mathrm{H}_0 : Q_0^{(1)}(\theta_0^{(1)}) = Q_0^{(2)}(\theta_0^{(2)}).$$

Two alternative hypotheses are that $\mathcal{M}_1$ is asymptotically better than $\mathcal{M}_2$ or the other way around;

$$\mathrm{H}_1^{(a)} = Q_0^{(1)}(\theta_0^{(1)}) < Q_0^{(2)}(\theta_0^{(2)}),$$
$$\mathrm{H}_1^{(b)} = Q_0^{(1)}(\theta_0^{(1)}) > Q_0^{(2)}(\theta_0^{(2)}).$$

1.2.2 Cross-validation

Cross-validation is a model selection procedure in which the data is split into two subsets called training set and validation set. The set of parameters of each model is estimated in the training set, and its goodness is evaluated with the validation set. Let $r \geq 2$, $k < r$ be integers. In leave-k-out r-fold cross-validation ((k, r)-CV), we first split T datapoints into r disjoint subsets. At each round of CV, We use $r - k$ of them as the training data, and the other k as the validation data. Multiple number of rounds among possible splits are performed to reduce variability.

Namely, let

$$\mathcal{N}_{T_{j,r}} = \{\lfloor T(j-1)/r \ \ +1,\ T(j-1)/r \ \ +2, \ldots,\ Tj/r\rfloor\}$$

be the indices of the j-th split. Let $\{S \subset \{1,2,\ldots,r\} : |S| = r-k\}$ and

$$\mathcal{N}_S = \bigcup_{j \in S} \mathcal{N}_{T_{j,r}}$$

be subset of datapoints consisted of folds in S. The moment on this datapoints is denoted as

$$Q_S^{(i)}(\theta^{(i)}) = \quad \frac{1}{|\mathcal{N}_S|} \sum_{t \in \mathcal{N}_S} f^{(i)}(v_t, \theta^{(i)}) \quad W_S^{(i)} \quad \frac{1}{|\mathcal{N}_S|} \sum_{t \in \mathcal{N}_S} f^{(i)}(v_t, \theta^{(i)}) \quad ,$$

and the model trained to minimize the moment is denoted as

$$\hat{\theta}_S^{(i)} = \underset{\theta^{(i)} \in \Theta^{(i)}}{\arg\min}\, Q_S^{(i)}(\theta^{(i)}).$$

Once the model is trained, it is validated by the rest of datapoints as:

$$Q_{S,\text{valid}}^{(i)}(\hat{\theta}_S^{(i)}) = \left\{ \frac{1}{|\mathcal{N}_{\setminus S}|} \sum_{t \in \mathcal{N}_{\setminus S}} f^{(i)}(v_t, \hat{\theta}_S^{(i)}) \right\} W_S^{(i)} \left\{ \frac{1}{|\mathcal{N}_{\setminus S}|} \sum_{t \in \mathcal{N}_{\setminus S}} f^{(i)}(v_t, \hat{\theta}_S^{(i)}) \right\},$$

where $\mathcal{N}_{\setminus S} = \{1,\ldots,T\} \setminus \mathcal{N}_S$. In (k,r)-CV, the averaged validation score of each model

$$Q_{\text{valid}}^{(i)} = \frac{1}{{}_rC_k} \sum_{S \subset \{1,2,\ldots,r\} : |S| = r-k} Q_{S,\text{valid}}^{(i)}(\hat{\theta}_S^{(i)})$$

is compared, and the model of smaller averaged validation score is selected. The procedure is summarized in Algorithm 1.

1.2.3 Consistency of CV in Model Selection

In this section, we derive the consistency of CV in GMM model selection. Let one of the models is misspecified. Without loss of generality, we assume the first model is the true model[2]. The true model satisfies the following moment condition:

$$\mathbf{E}[f^{(1)}(v_t, \theta_0^{(1)})] = 0.$$

[2]Of course, the model selection method should not exploit this fact.

The latter model is assumed to be misspecified: that is, for any $\theta^{(2)}$ the following holds:

$$\mathbf{E}[f^{(2)}(v_t, \theta^{(2)})] > 0.$$

The misspecification is divided into two local and non-local ones (Hall, 2005).

Assumption 1.2.1 *The false model is globally misspecified if there exists* $\mu(\theta)$ *such that* $||\mu(\theta)|| > 0$ *and*

$$\inf_{\theta^{(2)} \in \Theta^{(2)}} \mathbf{E} \ f^{(2)}(v_t, \theta^{(2)}) \ = \mu(\theta).$$

Alternatively, we can make a weaker assumption that the sample moment of the misspecified model converges to zero slower than that of the true model. This assumption covers cases where the misspecified model is more general (or too general) than the true model. This is the case, for example, the utility function in the true model is a linear function of price but the misspecified model incorporates higher order polynomials.

Assumption 1.2.2 *The false model is said to be locally misspecified if, for every* $\in (0, 1)$*, there exists* $c > 0$ *such that, when* T *is sufficiently large,* $P[Q^{(1)}_{\text{valid}} < Q^{(2)}_{\text{valid}}] \geq 1 - \ .$

Note that, in either definition of misspecification, the researcher does not know which model is true, and our interest lies in consistently choosing the true model over a misspecified model based on the dataset.

In the previous literature, Smith (1992) offers a pairwise comparison process for consistent model selection. However, it has some practical disadvantages when applied to empirical research: (i) A pairwise comparison could be extremely demanding if the space of candidate models is large, and (ii) it may be subject to over-fitting problem. To avoid those issues, the most common practice in the field of machine learning is to apply cross-validation (CV) algorithm. In the literature in statistics, Yang (2006,2007) have shown that even the simplest CV procedure can find a true model consistently when the data structure is regression form, i.e. $y_i = f(x_i) + \ _i$. Likewise to the literature, we define a consistent model selection as below:

Definition 1.2.1 *Assume that model* 1 *is correct while model* 2 *is wrong in a sense that it is globally misspecified. A selection rule is said to be consistent if the probability of selecting model* 1 *approaches* 1 *as* $T \longrightarrow \infty$.

To derive the consistency of CV, we define the following assumptions.

Assumption 1.2.3 (strict stationarity) $\mathbf{v} = \{v_t\}$ *is a strictly stationary process.*

Assumption 1.2.4 (regularity condition) *Let* $f^{(i)}(v_t, \theta)$ *and its population analogue* $\mathbf{E}[f^{(i)}(v_t, \theta)]$ *be continuous on* $\theta^{(i)}$ *for each* v_t. *Let* $\Theta^{(i)}$ *be compact and* $\mathbf{E}[\sup_{\theta^{(i)} \in \Theta^{(i)}} f^{(i)}(v_t, \theta)]$ *be bounded.*

Assumption 1.2.5 (ergodicity) $\mathbf{v} = \{v_t\}$ *is an ergodic process.*

Assumption 1.2.6 (identification condition) *Let*

$$\mathbf{E} \quad \frac{\partial f^{(i)}(v_t, \theta_0^{(i)})}{\partial \theta^{(i)}}$$

have rank d.

In the following we prove the following theorem.

Theorem 1.2.1 *Let Assumptions 1.2.3–1.2.6 hold. Then,* (r, k)*-CV is consistent.*

Proof of Theorem 1.2.1

We first states lemmas that are proven in (Hall, 2005), and by using them we prove the theorem.

Lemma 1.2.1 (Consistency of the estimator in the correct model, Theorem 3.1 in (Hall, 2005)) *Let* $S \subset \{1, \ldots, r\}, |S| = r - k$ *be any split in* (k, r)*-CV, and model* 1 *be correctly specified. Let Assumptions 1.2.3–1.2.6 hold. Then,*

$$\hat{\theta}_S^{(1)} \to^p \theta_0^{(1)} \tag{1.2}$$

as $T/r \to \infty$.

Lemma 1.2.2 (Property of a globally misspecified estimator, Theorem 5.2 in (Hall, 2005)) *Let $S \subset \{1, \ldots, r\}, |S| = r - k$ be any split in (k, r)-CV. Let Assumptions 1.2.3–1.2.6 hold. Then, here exists $c > 0$ such that*

$$Q_0^{(i)}(\hat{\theta}_S^{(i)}) \xrightarrow{p} c \tag{1.3}$$

as $T/r \to \infty$.

Lemma 1.2.3 (Uniform convergence of the moment, Lemma 3.1 in (Hall, 2005)) *Let Assumptions 1.2.3–1.2.6 hold. Then,*

$$\sup_{\theta^{(1)} \in \Theta^{(1)}} |Q_{S,\text{valid}}^{(i)}(\theta^{(1)}) - Q_0^{(1)}(\theta^{(1)})| \xrightarrow{p} 0 \tag{1.4}$$

$$\sup_{\theta^{(2)} \in \Theta^{(2)}} |Q_{S,\text{valid}}^{(2)}(\theta^{(2)}) - Q_0^{(2)}(\theta^{(2)})| \xrightarrow{p} 0 \tag{1.5}$$

Proof [Proof of Theorem 1.2.1] We show that,

$$\sup_{\theta^{(1)} \in \Theta^{(1)}} |Q_{\text{valid}}^{(1)}| \xrightarrow{p} 0 \tag{1.6}$$

and there exists $c > 0$ such that

$$|Q_{\text{valid}}^{(2)}| \xrightarrow{p} c \tag{1.7}$$

which imply Theorem 1.2.1. First,

$$\begin{aligned} |Q_{\text{valid}}^{(1)} - Q_0^{(1)}(\theta_0^{(1)})| &\leq \sup_{S \in \{1,\ldots,r\}: |S| = r-k} |Q_{S,\text{valid}}^{(1)}(\hat{\theta}_S^{(1)}) - Q_0^{(1)}(\theta_0^{(1)})| \\ &\leq \sup_{S \in \{1,\ldots,r\}: |S| = r-k} |Q_{S,\text{valid}}^{(1)}(\hat{\theta}_S^{(1)}) - Q_0^{(1)}(\hat{\theta}_S^{(1)})| + |Q_0^{(1)}(\hat{\theta}_S^{(1)}) - Q_0^{(1)}(\theta_0^{(1)})| \end{aligned}$$

Inequality (1.4) implies the first term converges to zero in probability, and the second term converges to zero in probability by (1.2). In other words,

$$|Q_{\text{valid}}^{(1)}(\theta^{(1)}) - Q_0^{(1)}(\theta_0^{(1)})| \xrightarrow{p} 0 \tag{1.8}$$

and by Assumption 3.3 in (Hall, 2005),

$$Q_0^{(1)}(\theta_0^{(1)}) = 0 \tag{1.9}$$

and thus inequality (1.6) is derived. We next show (1.7). We have,

$$
\begin{aligned}
&Q^{(2)}_{\mathrm{valid}} \\
&\geq Q^{(2)}_0(\theta^{(2)}_0) - \frac{1}{{}_rC_k} \sum_{S\in\{1,\ldots,r\}:|S|=r-k} |Q^{(2)}_{S,\mathrm{valid}}(\hat{\theta}^{(2)}_S) - Q^{(2)}_0(\hat{\theta}^{(2)}_S)| - |Q^{(2)}_0(\hat{\theta}^{(2)}_S) - Q^{(2)}_0(\theta^{(2)}_0)| \quad ,
\end{aligned}
$$

where the first term of the RHS converges to $c > 0$ in probability by (1.2). The second term converge to zero in probability by (1.5). The third term goes to zero in probability by our assumption. Therefore (1.7) holds. ■

1.2.4 Statistical testing

This section proposes a statistical hypothesis testing on our CV-based model selection. Let

$$R_{\mathrm{CV}} = \frac{|\mathcal{N}_S|^{1/2}(Q^{(1)}_{\mathrm{valid}} - Q^{(2)}_{\mathrm{valid}})}{\hat{\sigma}^2}$$

be the test statistics that indicates either the first or the second hypothesis is better than the other. Here, $\hat{\sigma}^2$ is the estimator of the limiting variance σ^2_0 of R_{CV}. The null hypothesis of the test is

$$\mathrm{H}_0 : Q^{(1)}_0(\theta^{(1)}_0) = Q^{(2)}_0(\theta^{(2)}_0).$$

These are two alternative hypotheses of interest: The first one indicates $\mathcal{M}_1$ is better than $\mathcal{M}_2$. That is,

$$\mathrm{H}^{(a)}_1 : Q^{(1)}_0(\theta^{(1)}_0) < Q^{(2)}_0(\theta^{(2)}_0)$$

and the second one indicates $\mathcal{M}_2$ is better than $\mathcal{M}_2$:

$$\mathrm{H}^{(b)}_1 : Q^{(1)}_0(\theta^{(1)}_0) > Q^{(2)}_0(\theta^{(2)}_0).$$

Following (Rivers and Vuong, 2002; Hall and Pelletier, 2011), we discuss conditions where the statistics R_{CV} is asymptotically normal. We first consider the testing statistics in the general case in Section 1.2.4. Moreover, we show in the case the dependency among splits are sufficiently small in Section 1.2.4, where the statistics is represented in a much computationally efficient way.

We pose the following assumption on the structure of the weight matrix that is essentially the same as (Hall and Pelletier, 2011):

Assumption 1.2.7 (parameterization of the weight matrix) *Let $W^{(i)}$ depends on a vector nuisance parameter $\tau_0^{(i)}$ and $\hat{\tau}_S^{(i)}$ is the estimator of $\tau_0^{(i)}$ as $W^{(i)} = W^{(i)}(\tau_0^{(i)})$ and $W_S^{(i)} = W_S^{(i)}(\hat{\tau}_S^{(i)})$. It is assumed that the nuisance parameter satisfies*

$$|\mathcal{N}_S|^{1/2}(\hat{\tau}_S^{(i)} - \tau_0^{(i)}) = -A_*^{(i)}|\mathcal{N}_S|^{-1/2} \sum_{t \in \mathcal{N}_S} Y_t^{(i)} + o_p(1)$$

for some symmetric matrix of constants $A_*^{(i)}$ and data-dependent vector $Y_t^{(i)}$, and the weight matrix satisfies

$$|\mathcal{N}_S|^{1/2} \quad \text{vech}[W_S^{(i)}] - \text{vech}[W^{(i)}] \quad = \Delta^{(i)}|\mathcal{N}_S|^{1/2}(\hat{\tau}_S^{(i)} - \tau_0^{(i)}) + o_p(1)$$

for some matrix of constants $\Delta^{(i)}$.

To discuss statistical testing, we need to have asymptotic normality property. The following assumption guarantees that the moment is "well-behaved" around the optimal value $\theta_0^{(i)}$.

Assumption 1.2.8 (regularity condition on derivative)

- *The derivatie matrix $\partial f^{(i)}(v, \theta^{(i)})/\partial\theta^{(i)}$ exists and is continuous on $\Theta^{(i)}$ for each v.*
- *$\theta_0^{(i)}$ lies in the interior of $\Theta^{(i)}$.*
- *$\mathbb{E}[\partial f^{(i)}(v, \theta_0^{(i)})/\partial\theta^{(i)}\,]$ exists and is finite.*
- *$\mathbb{E}[\partial f^{(i)}(v, \theta^{(i)})/\partial\theta^{(i)}\,]$ continuous on some neighborhood N of $\theta_0^{(i)}$.*
- $\sup_{\theta^{(i)} \in N} ||(1/T) \sum_{t=1}^{T} \partial f^{(i)}(v, \theta^{(i)})/\partial\theta^{(i)} - \mathbb{E}[\partial f^{(i)}(v, \theta^{(i)})/\partial\theta^{(i)}\,]|| \xrightarrow{p} 0$.

For the ease of discussion, we further add the following notation. Let $F_S^{(i)} = |\mathcal{N}_{\setminus S}|^{-1/2} \sum_{t \in \mathcal{N}_{\setminus S}} \{f^{(i)}(v_t,$ Let $G_0^{(i)} = \mathbb{E}[\partial f^{(i)}(v_t, \theta^{(i)})/\partial\theta^{(i)}]$, and its empirical counterpart be $G_S^{(i)} = |\mathcal{N}_S|^{-1} \sum_{t \in \mathcal{N}_{\setminus S}} (\partial f^{(i)}(v_t, \theta^{(i)})/\partial$ Let $\{S_1, \ldots, S_{rC_k}\} = \{S \subset \{1, 2, \ldots, r\} : |S| = r - k\}$ be the set of all splits. We also denote $\theta = (\theta^{(1)}, \theta^{(2)})$, and θ_0 and $\hat{\theta}_S$ are defined in the same way.

General splitting

Then, V_* is the

$$V_* = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,rC_k} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,rC_k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{rC_k,1} & c_{rC_k,2} & \cdots & c_{rC_k,rC_k} \end{pmatrix}$$

where $c_{j,j}$ is a submatrix such that

$$c_{j,j} = \lim_{T\to\infty} \text{Cov} \begin{pmatrix} \xi_t(\theta_0), & \xi_t(\theta_0) \\ t\in\mathcal{N}_{\setminus S_j} & t\in\mathcal{N}_{\setminus S_j} \end{pmatrix}$$

$$\xi_t(\theta) = f^{(1)}(v_t,\theta^{(1)}) - \mathbb{E}[f^{(1)}(v_t,\theta^{(1)}], Y_t^{(1)}, f^{(2)}(v_t,\theta^{(2)}) - \mathbb{E}[f^{(2)}(v_t,\theta^{(2)})], Y_t^{(2)} \quad .$$

Moreover,

$$R_* = R_*^{(1)}, R_*^{(2)}, R_*^{(1)}, R_*^{(2)}, \ldots, R_*^{(1)}, R_*^{(2)}$$

$$R_*^{(i)} = \begin{bmatrix} 2W^{(i)}\mathbb{E}[f^{(1)}(v_t,\theta^{(1)})] \\ -A_*^{(i)}\Delta^{(i)} B_i\mathbb{E}[f^{(1)}(v_t,\theta^{(1)})] \otimes \mathbb{E}[f^{(1)}(v_t,\theta^{(1)})] \end{bmatrix}$$

where B_i is the $q_i^2 \times q_i(q_i+1)/2$ matrix such that $\text{vec}(W^{(i)}) = B_i\text{vech}(W^{(i)})$.

Assumption 1.2.9 *1. Assume that* $[F_{S_1}^{(1)}, F_{S_1}^{(2)}, F_{S_2}^{(1)}, F_{S_2}^{(2)}, \ldots, F_{S_{rC_k}}^{(1)}, F_{S_{rC_k}}^{(2)}] \to N(0,\Sigma(\theta))$.

Where $\Sigma(\theta)$ *is a positive semi-definite matrix of constants.*

2. *rank*$\{G_0^{(i)}\} = d$.

3. $S^{1/2}(\hat{\theta}_S^{(i)} - \theta^{(i)}) = O_p(1)$.

4. *The empirical estimator of each* $\Sigma(\theta)$ *converges as* $\hat{\Sigma}(\hat{\theta}_S) \to \Sigma(\theta_0)$.

Theorem 1.2.2 (asymptotic normality of R_{CV}) *Assume that both models* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *are misspecified. Assume that Assumption 1.2.9 holds. Assume Assumptions 1.2.3, 1.2.4, 1.2.5, and 1.2.8 hold. Assume that the null hypothesis* H_0 *holds. Let* $W^{(i)} = I_{q_i}$. *Then,*

$$R_{\text{CV}} \to N(0,1).$$

Remark 1.2.1 *Theorem 1.2.2 poses the assumption that both models are misspecified. As discussed in (Hall and Pelletier, 2011), this assumption is essential: One can check that, under correctly specified models, the distribution of R_{CV} does not have asymptotic normality.*

Remark 1.2.2 *As discussed in (Hall and Inoue, 2003) a constant weight matrix has the best rate of convergence in misspecified models and thus the assumption of identity $W^{(i)}$ in Theorem 1.2.2 is reasonable.*

Proof [Proof of Theorem 1.2.2] The theorem is an extension of Theorem 1 in (Hall and Pelletier, 2011) to multiple splitting. The mean value theorem applied to $Q^{(i)}_{S,\mathrm{valid}}(\hat{\theta}^{(i)}_S)$ around $\theta^{(i)}_0$, we obtain

$$Q^{(i)}_{S,\mathrm{valid}}(\hat{\theta}^{(i)}_S) = Q^{(i)}_{S,\mathrm{valid}}(\theta^{(i)}_0) + \left\{ \frac{\partial Q^{(i)}_{S,\mathrm{valid}}(\theta^{(i)})}{\partial \theta^{(i)}} \right\}_{\theta^{(i)}=\bar{\theta}^{(i)}_S} (\hat{\theta}^{(i)}_S - \theta^{(i)}_0)$$

where $\bar{\theta}^{(i)}_S = \lambda_S \theta^{(i)}_0 + (1-\lambda_S)\hat{\theta}^{(i)}_S$ for some $\lambda_S \in [0,1]$. Let

$$\Phi^{(i)}(\theta^{(i)}_0) = 2G^{(i)}_0(\theta^{(i)}_0)\, W^{(i)} \mathbb{E}[f^{(i)}(v_t, \theta^{(i)}_0)].$$

From our assumptions, we obtain

$$Q^{(i)}_{S,\mathrm{valid}}(\hat{\theta}^{(i)}_S) = Q^{(i)}_{S,\mathrm{valid}}(\theta^{(i)}_0) + \quad \frac{\partial \hat{\theta}^{(i)}_S}{\partial \theta^{(i)}} \quad (\hat{\theta}^{(i)}_S - \theta^{(i)}_0) + o_p(|\mathcal{N}_{\setminus S}|^{-1/2}),$$

and thus

$$\begin{aligned} |\mathcal{N}_{\setminus S}|^{1/2}[Q^{(1)}_{S,\mathrm{valid}}(\hat{\theta}^{(1)}_S) - Q^{(2)}_{S,\mathrm{valid}}(\hat{\theta}^{(2)}_S)] = |\mathcal{N}_{\setminus S}|^{1/2}[Q^{(1)}_{S,\mathrm{valid}}(\theta^{(1)}_0) - Q^{(2)}_{S,\mathrm{valid}}(\theta^{(2)}_0)] \\ + \quad \Phi^{(1)}(\theta^{(1)}_0) \quad S^{1/2}(\hat{\theta}^{(1)}_S - \theta^{(1)}_0) \\ - \quad \Phi^{(2)}(\theta^{(2)}_0) \quad S^{1/2}(\hat{\theta}^{(2)}_S - \theta^{(2)}_0) \\ + o_p(1). \end{aligned} \tag{1.10}$$

Note that the GMM estimator minimizes the moment condition, which implies $G^{(i)}_S(\hat{\theta}^{(i)}_S)\, W^{(i)}_S(1/|\mathcal{N}_{\setminus S}|^{-1})$ 0. This fact implies the third and fourth terms of (1.11) vanishes. Namely,

$$\begin{aligned} |\mathcal{N}_{\setminus S}|^{1/2}[Q^{(1)}_{S,\mathrm{valid}}(\hat{\theta}^{(1)}_S) - Q^{(2)}_{S,\mathrm{valid}}(\hat{\theta}^{(2)}_S)] = |\mathcal{N}_{\setminus S}|^{1/2}[Q^{(1)}_{S,\mathrm{valid}}(\theta^{(1)}) - Q^{(2)}_{S,\mathrm{valid}}(\theta^{(2)})] \\ + o_p(1). \end{aligned} \tag{1.11}$$

With the choice $W^{(i)} = I_{q_i}$ for the weighting matrix, and by using the symmetry of the moment we obtain

$$|\mathcal{N}_{\setminus S}|^{1/2}[Q^{(1)}_{S,\mathrm{valid}}(\hat{\theta}^{(1)}_S) - Q^{(2)}_{S,\mathrm{valid}}(\hat{\theta}^{(2)}_S)] =$$

$$2 \quad \mu^{(1)}(\theta^{(1)}_0)|\mathcal{N}_{\setminus S}|^{-1/2} \sum_{t \in \mathcal{N}_{\setminus S}} [f^{(1)}(v_t, \theta^{(1)}_0) - \mu^{(1)}(\theta^{(1)}_0)]$$

$$- \mu^{(2)}(\theta^{(2)}_0)|\mathcal{N}_{\setminus S}|^{-1/2} \sum_{t \in \mathcal{N}_{\setminus S}} [f^{(2)}(v_t, \theta^{(2)}_0) - \mu^{(2)}(\theta^{(2)}_0)] \quad + o_p(1),$$

which, combined with our assumptions, completes the proof. ■

When dependency among validation splits is small

Calculating the asymptotic variance of Theorem 1.2.2 requires a calculation of a matrix with its size proportional to the number of splits, which in some cases is computationally prohibitive. This section consider the case where the dependency between the validation data is sufficiently small. In such a case, we can circumvent the computation of a large matrix.

In particular, the leave-one-out CV (special case of our CV with $k = 1$) when each datapoint is identically and independently distributed (i.i.d), the following assumption holds:

Assumption 1.2.10 *Assume that each validation split $\{\mathcal{N}_{\setminus S_j}\}$ is independent and identically distributed.*

Theorem 1.2.3 (asymptotic normality of R_{CV}, Leave-one-out) *Let assumptions in Theorem 1.2.2 hold. Let Assumption 1.2.10 holds. Then, the limit variance is written as*

$$\sigma^2 = \sum_{S \in \{S \subset \{1,2,\ldots,r\} : |S| = r-k\}} R^{\mathrm{single}}_* \, V^{\mathrm{single}}_*(S) R^{\mathrm{single}}_*$$

where

$$
\begin{aligned}
R_*^{\text{single}} = &\ R_*^{(1)}, R_*^{(2)} \\
R_*^{\text{single},(i)} = &\begin{bmatrix} 2W^{(i)}\mathbb{E}[f^{(1)}(v_t,\theta^{(1)})] \\ -A_*^{(i)}\Delta^{(i)}\ B_i\mathbb{E}[f^{(1)}(v_t,\theta^{(1)})]\otimes\mathbb{E}[f^{(1)}(v_t,\theta^{(1)})] \end{bmatrix} \\
V_*^{\text{single}}(S) = &\lim_{T\to\infty}\text{Var}(\sum_{t\in\mathcal{N}_{\setminus S}}\xi_t)
\end{aligned}
\tag{1.12}
$$

And The asymptotic normality holds:

$$R_{\text{CV}} \to N(0,1).$$

The proof of Theorem 1.2.10 directly follows by following the same steps as Theorem 1.2.2 with additional fact that Assumption (1.2.10) implies the block-diagonal property of V_* as $c_{i,j} \to 0$ for $i = j$ and the identity of each block.

1.3 Monte-Carlo Experiments in Linear Model

In this section we present a simple simulation of instrumental variables (IV) regression models to illustrate the consistency of our cross-validation algorithm of model selection. This example also highlights how GMM-minimand-based model comparison and cross-validation can exhibit different results. The setting is similar to the one on (**?**). Suppose the true data generating process is

$$\boldsymbol{y} = X_1\boldsymbol{\beta^1} + X_2\boldsymbol{\beta^2} + Z_2\boldsymbol{\alpha} + \quad ,$$

where $\boldsymbol{y}$ is a $T \times 1$ vector and X_1 and X_2 are $T \times p_1$ and $T \times p_2$ matrix respectively. X_1 and X_2 are generated from instrumental variables as

$$
\begin{aligned}
X_1 &= Z_1\boldsymbol{\delta^1} + \boldsymbol{\xi}^1, \\
X_2 &= Z_2\boldsymbol{\delta^2} + \boldsymbol{\xi}^2,
\end{aligned}
$$

where Z_1 and Z_2 are $T \times c_1$ and $T \times c_2$ matrix respectively.

We consider a case where we have two candidate models to compare. The first model exploits the explanatory variables X_1 and instrumental variables Z_1.

$$\mathcal{M}^1 : \boldsymbol{y} = X_1\boldsymbol{\beta} + \ ^1,$$
$$\mathbb{E}[Z_1 \ ^1] = 0,$$

whereas the second model employs X_2 and Z_2;

$$\mathcal{M}^2 : \boldsymbol{y} = X_2\boldsymbol{\beta} + \ ^2,$$
$$\mathbb{E}[Z_2 \ ^2] = 0.$$

Each model has different explanatory variables as well as the set of instrumental variables so that two models are non-nested. In addition, there are two important differences between the two candidates. First, the second model can be "misspecified" when $\alpha = 0$, because the instrumental variables Z_2 influences $\boldsymbol{y}$ directly and thus IVs are not independent from ^{2}. When $|\alpha| > 0$ and does not decrease with the number of observations, i.e. $\alpha = 10$, it is globally misspecified, which results in inconsistent estimates of the parameters.

The second difference is that the number of the variables. In the following, we assume that $p^1 \leq p^2$, meaning that the second model has a larger number of explanatory variables. As discussed earlier, this may cause "over-fitting" issue to the estimation even if the model is falsely specified. In such a case, previous literature proposes the ways to penalize the model by the number of parameters ((Andrews, 1999)). We compare the performance of the proposed method with the ones of those existing methods in the later section.

Though this example may seem to be somewhat arbitrary, similar problems arise in many situations when econometric models are compared. Specifically, one model can be flexible (or even "over flexible") but misspecified, while the other is simpler but accurate. Some researchers may not value the simplicity, but they would prefer a "correctly specified" model than misspecified models. For example, think of a case where economists try to explain wage from education and other variables, where education is endogenous and has to be proxied by IVs. The misspecified model includes incorrect IVs that gives bias to the estimate of the coefficient. Even if one model exhibits a good fit to the data, if the

coefficient of interest is not properly estimated, such a model does not serve well for labor economists. In those occasions, our algorithm serves to help researchers to find the most "correct" model. Our method is general enough so that any specification can be compared.

1.3.1 Results

First we consider the case where over-fitting is a concern as the misspecified model has more parameters therefore could exhibit better fit to the data. We compare our methodology in this case to the model selection procedures proposed by (Andrews, 1999) as well as simple GMM comparison as in the previous section. (Andrews, 1999) defines GMM-AIC and GMM-BIC criterion as

$$\text{GMM-AIC: } TQ_T^{(i)}(\theta_T^i) - 2(|c^i| - p^i);$$
$$\text{GMM-BIC: } TQ_T^{(i)}(\theta_T^i) - (|c^i| - p^i)lnT,$$

for $i = 1, 2$. The procedure chooses the model that exhibits smaller value of the criterion.

Figure 1.1 shows the empirical probability of choosing the correctly specified model by cross-validation. One can see that, even when the model 2 has larger number of variables, it chooses the model 1 with very high chance even when the data is limited. When the bias parameter of the model 2 α is as large as 12., it selects the first model with probability 91.2% even when the data size is only 100 and the second model has 9 variables compared to 3 of the first model.

On the other hand, model selection based on in-sample moment performs extremely poorly when the misspecified model has much more variables than the first model. When $p^2 = 9$, even with data size 1600 the accuracy is as bad as 59.1%, only slightly above chance level of 50% (when $\alpha = 12.$). With data size 200, it chooses the second model only for 15.7%, clearly indicating it is subject to over-fitting.

Note that in our setting, GMM-AIC and GMM-BIC exhibit exactly same choice of models as simple GMM based selection. This is due to the unbalance of two terms in the criterion. In our case, the first term is typically on order of more than 10^5, while the second term is no greater than 10^2. Many factors influence the magnitude of the first term, such as

the choice of weighting matrix or number of moment conditions. Our result suggests that while cross validation robustly performs in many situations, performance of GMM based model selection is sensitive to those settings.

We turn to the case where the two models have the same number of parameters, while the second model is misspecified. As the number of parameters is the same across two models, note that GMM, GMM-BIC, and GMM-AIC simply choose the model with smaller GMM minimand. Figure 1.2 compares the performance of cross-validation algorithm and the GMM minimand based model selection when the second model is globally misspecified. The y-axis shows the probability that the correctly specified model is chosen by each algorithm. The result indicates that when overfitting is not a concern, GMM based model selection performs slightly better than cross validation, especially when the data is smaller.

1.4 Nonlinear Experiment: Collusion Detection

In this section, we demonstrate another Monte-Carlo study to show how our algorithm works in a structural estimation incorporating nonlinear and non-nested models. Specifically, we simulate and estimate a variant of a price collusion model suggested by Bresnahan (1987). The goal of our model selection procedure is to detect whether the firms are colluded, or determining the price competitively using the share and price data. The underlying idea is that the prices of the products of colluded firms are determined to maximize the joint profit, while the competitive price should maximizes the profit of individual firms. Therefore, given the same (true) parameters in demand and cost function, the pricing pattern varies according to the collusive structure.

A methodology to study whether collusive behavior exists within a certain industry is by itself an important research topic because ignoring the possibility of collusive pricing may lead to a biased inference of cost estimation, which could be a critical problem for policy implication in applications such as merger analysis.

In the same way as the previous section, we compare the performance of CV-based algorithm to GMM-minimand-based algorithm based on the theory of (Rivers and Vuong,

2002). Note that since the number of parameters in a model does not vary across collusive structure, AIC or BIC adjustment does not influence the model selection criteria. We show that in a realistic sample size, CV performs better than in-sample comparison in many cases.

The shares and prices are simulated from a standard logit demand system and static pricing. We simulate data assuming a certain collusive structure. Then we test if and how often CV algorithm can discover the assumed collusive structure. The estimation process is similar to (Hu et al., 2014).

1.4.1 Model

Assume each firm produces a single product and denote them as $j = 1, ..., J$. The markets are denoted as $t = 1, ..., T$. The demand is assumed to be a simple logit demand specification: the utility of a consumer i purchasing a product j in a market t is expressed as

$$u_{ijt} = X_{jt}\beta + \alpha p_{jt} + \xi_{jt} + \ _{ijt},$$

where X_{jt} is the observed characteristics that influence the demand and ξ_{jt} is the unobserved utility shock . Assuming $_{ijt}$ follows i.i.d type-I extreme value distribution, the share function is

$$D_{jt}(\mathbf{p}_t) = \frac{\exp(X_j\beta + \alpha p_{jt} + \xi_{jt})}{\sum_{j=1}^{J} \exp(X_{j\ t}\beta + \alpha p_{j\ t} + \xi_{j\ t})} M_t,$$

where $\mathbf{p}_t = \{p_{jt}\}_{j=1,...,J}$ is the vectorized prices and M_t is the market size which is known to the researcher. For simplicity, we do not allow random-coefficients ((Berry et al., 1995)) as typically done in applications.

Firms' marginal cost is expressed as

$$MC_{jt} = Y_{jt}\gamma + \lambda_{jt}$$

,where Y_{jt} is the observed characteristics that affect the marginal cost, and λ_{jt} is the i.i.d cost shocks. The profit of each product is

$$\pi_{jt}(\mathbf{p}_t) = (p_{jt} - MC_{jt})D_{jt}(\mathbf{p}_t).$$

We assume that colluded firms jointly maximize their net profit, sum of π_{jt} over j in a group. Define $\mathbf{\Delta}$ as a $J \times J$ matrix of price elasticity of colluded products where the (j, r)th element is

$$\mathbf{\Delta}_{jr} = \begin{cases} -\frac{\partial D_r}{\partial p_j} & \text{if } j \text{ and } r \text{ are colluded} \\ 0 & \text{otherwise.} \end{cases}$$

By solving the first order conditions, the equilibrium prices are determined to satisfy

$$\mathbf{p}_t = (\mathbf{\Delta})^{-1} \mathbf{D}_t - \mathbf{MC}_t,$$

where $\mathbf{D}_t$ and $\mathbf{MC}_t$ are a vectorized representation of $D_{jt}(\mathbf{P}_t)$ and $\{MC_{jt}\}_{j=1,..,J}$ respectively.

1.4.2 Estimation and Model Selection

The parameter estimation under each model follows a standard GMM procedure with instrumental variables. Let Z be instrumental variables that influence the price but are not correlated with the unobserved shocks ξ and λ. Given a model, the parameters are chosen to minimize the GMM objective defined from the moment condition

$$\mathbb{E}[\xi Z] = 0$$
$$\mathbb{E}[\lambda Z] = 0.$$

The instrumental variables Z include (i) own characteristics, (ii) square of own characteristics, (iii) mean of characteristics in a market, and (iv) square of mean characteristics in a market. The weighting matrix is set to be $W = (Z\ Z)^{-1}$.

The candidate models are represented as partitions of firms into price-colluded groups. For instance, if the number of firms is two ($j = 1, 2$), the possible models are either

competitive ($\{\{1\},\{2\}\}$) or collusive ($\{\{1,2\}\}$). If three firms (j=1,2,3), possible models are $\{\{1\},\{2\},\{3\}\}$ (all competitive), $\{\{1\},\{2,3\}\}$, $\{\{1,2\},\{3\}\}$, $\{\{1,3\},\{2\}\}$, and $\{\{1,2,3\}\}$ (all colluded).

1.4.3 Simulation Results

We consider different number of observed markets, $T = \{25, 50, 75, 100\}$, realistic numbers for real world application[3]. We also vary the true value of price coefficient to test the performance with different difficulty of model selection. Along with the data size, the difficulty of model selection depends on how different the observed data would be across different models. In this particular example, the key difference between models is generated from cross price elasticity. When the cross price elasticity is low, competitive price and colluded price do not differ as much, which makes it harder to find the true model. In logit-demand, the cross price elasticity is calculated by multiplying the share of the two products. Thus, lower price coefficient generally makes model selection easer as it increases the realized share, and the cross price elasticity as a result. For each setting, we generate 100 synthetic dataset and perform the model selection in each.

Table 1.1 reports the mean and standard deviation of CV score across true models and candidate models with the price coefficient equals to $-.1$ and $-.3$. The second column represents the true partition of firms, and the third to seventh are the results corresponding to each candidate model. The CV score of the true model is on average smaller than the mis-specified models in any specification. Also, the standard deviation of the score is smaller for the true model. Both mean and standard deviation of the true model decline in the number of observations.

We report the probability that each candidate model is chosen by our algorithm in table 1.2. In each setting, the probability to find the true model increases in the number of markets, which corresponds to our theoretical finding in section 2. For comparison, Table 1.3 presents the same for GMM-minimiand comparison.

[3]For instance, (Nevo, 2001) observes 94 independent markets.

Figure 1.3 compares the performance of our model selection to a simple in-sample GMM fit comparison under different price coefficient. It shows that our CV algorithm generally performs better than in-sample comparison. The difference is particularly large when the true model is partially colluded (second column). As seen in Table 1.3, GMM comparison tends to select all-competitive model in such a case.

1.5 Cross-Validation Approach to MPEC Estimation

In this section, we propose a method to apply cross-validation algorithm when estimation is based on Mathematical Programming of Equilibrium Constraint (MPEC) approach proposed by (Su and Judd, 2012). MPEC approach formulates the estimation as an optimization problem with constraints: The variables of the optimization consists of structural parameters as well as endogenous latent economic variables, and the constraints among the variables represent the equilibrium condition that the economic model requires.

The application of the cross validation procedure to MPEC estimation is not straightforward: If parameters estimated from training data is substituted in a MPEC model with test data directly, the constraints would be not satisfied in general. In such a case, we cannot directly compare GMM objective on test data across models since we also have to consider the violation of constraints as indication of model misfit.

Taking the above discussion into consideration, we propose a modified cross validation procedure. We differentiate the choice variables for the optimization problem into two categories: model variables and observation-specific variables. Model variables are specific to the model, therefore shared across training and test data. Observation-specific variables are latent variables defined on each observation. For instance, in BLP demand estimation example on (Dubé et al., 2012), the price elasticity is a parameter assumed to be constant across observations, thus treated as a model variable. Meanwhile, the unobserved utility shock (ξ_{jt} in their notation) is defined for each datapoint, thus regarded as observation specific.

Our modification is simple. In training data, we jointly choose the model variables and observation-specific variables to optimize the GMM objective function with equilibrium constraints. In test data, we still solve a constrained optimization problem, but only with respect to observation-specific variables while the model variables are set to the estimates from training data. The algorithm is described in detail below and summarized in Algorithm 2.

1.5.1 GMM-MPEC

We first outline the MPEC formulation of parameter estimation. Here we follow the notation of (Su and Judd, 2012) except that we allow some endogenous variables to be observation-specific. Suppose an econometric model $\mathcal{M}_i$ is expressed with the parameter vector $\theta^{(i)}$, a vector of endogenous variables $\sigma^{(i)}$, and endogenous variables that are observation-specific $\eta^{(i)}$, and the equilibrium constraint $h^{(i)}(\theta^{(i)}, \sigma^{(i)}, \eta^{(i)}) = 0$. In MPEC formulation, each model is characterized by a set of moment conditions with equilibrium constraints:

$$\begin{aligned} \mathcal{M}_i \Rightarrow & E[f^{(i)}(v_t, \theta_0^{(i)}, \sigma_0^{(i)}, \eta_0^{(i)})] = 0 \\ & \text{s.t.} \\ & h^{(i)}(\theta_0^{(i)}, \sigma_0^{(i)}, \eta_0^{(i)}) = 0. \end{aligned}$$

Given the observation $\{v_t\}_{t=1,..,T}$, the parameters of each model are estimated via MPEC:

$$(\theta_T^{(i)}, \sigma_T^{(i)}, \eta_T^{(i)}) = \underset{\theta^{(i)}, \sigma^{(i)}, \eta^{(i)}}{\arg\min}\ Q_T^{(i)}(\theta^{(i)}, \sigma^{(i)}, \eta^{(i)}) \tag{1.13}$$

$$\text{s.t.} \tag{1.14}$$

$$h^{(i)}(\theta^{(i)}, \sigma^{(i)}, \eta^{(i)}) = 0. \tag{1.15}$$

where

$$
\begin{aligned}
& Q_T^{(i)}(\theta^{(i)}, \sigma^{(i)}, \eta^{(i)}) \\
= \quad & \frac{1}{T} \sum_{t=1}^{T} f^{(i)}(v_t, \theta^{(i)}, \sigma^{(i)}, \eta^{(i)}) \quad W_T^{(i)} \quad \frac{1}{T} \sum_{t=1}^{T} f^{(i)}(v_t, \theta^{(i)}, \sigma^{(i)}, \eta^{(i)}) \quad .
\end{aligned}
$$

Let $\theta_T^{(i),\mathrm{GMM-MPEC}}$ be the parameters that are solution of Eq. (1.15), and let $\theta_T^{(i),\mathrm{GMM}}$ be the solution of standard GMM (i.e., Eq. (1.1)). Moreover, let

$$
\begin{aligned}
V_T^{(i),\mathrm{GMM-MPEC}}(\theta) = \min_{\sigma^{(i)}, \eta^{(i)}} \; & Q_T^{(i)}(\theta, \sigma^{(i)}, \eta^{(i)}) && (1.16) \\
\text{s.t.} \quad & && (1.17) \\
& h^{(i)}(\theta, \sigma^{(i)}, \eta^{(i)}) = 0, && (1.18)
\end{aligned}
$$

and $V_T^{(i),\mathrm{GMM}}(\theta) = Q_T^{(i)}(\theta)$. The equivalence of GMM and GMM-MPEC implies

$$
\begin{aligned}
\theta_T^{(i),\mathrm{GMM-MPEC}} &= \theta_T^{(i),\mathrm{GMM}} \\
V_T^{(i),\mathrm{GMM-MPEC}}(\theta) &= V_T^{(i),\mathrm{GMM}}(\theta). \qquad (1.19)
\end{aligned}
$$

1.5.2 Cross-Validation in GMM-MPEC Approach

We split the observations in the same way as section 2. The moment on the datapoints S is

$$
\begin{aligned}
& Q_S^{(i)}(\theta^{(i)}, \sigma^{(i)}, \eta^{(i)}) \\
= \quad & \frac{1}{|\mathcal{N}_S|} \sum_{t \in \mathcal{N}_S} f^{(i)}(v_t, \theta^{(i)}, \sigma^{(i)}, \eta^{(i)}) \quad W_S^{(i)} \quad \frac{1}{|\mathcal{N}_S|} \sum_{t \in \mathcal{N}_S} f^{(i)}(v_t, \theta^{(i)}, \sigma^{(i)}, \eta^{(i)}) \quad .
\end{aligned}
$$

We train the model to minimize the moment under equilibrium constraint. The trained model is denoted as

$$
\begin{aligned}
(\theta_S^{(i)}, \sigma_S^{(i)}, \eta_S^{(i)}) = \arg\min_{\theta^{(i)}, \sigma^{(i)}, \eta^{(i)}} \; & Q_S^{(i)}(\theta^{(i)}) \\
\text{s.t.} \quad & \\
& h^{(i)}(\theta^{(i)}, \sigma^{(i)}, \eta^{(i)}) = 0.
\end{aligned}
$$

Once the model is trained, it is validated by the rest of datapoints. Instead of simply evaluating the GMM objective in the validation data at the trained model parameters, observation-specific endogenous variables need to be chosen so that the equilibrium constraints are satisfied. We do so by minimizing the GMM objective subject to equilibrium constraints with respect to η only, while model parameters are fixed at trained value. Formally,

$$Q^{(i)}_{S,\text{valid}} =$$

$$\arg\min_{\eta^{(i)}} \left\{ \frac{1}{|\mathcal{N}_{\setminus S}|} \sum_{t \in \mathcal{N}_{\setminus S}} f^{(i)}(v_t, \theta^{(i)}_S, \sigma^{(i)}_S, \eta^{(i)}) \right\} W^{(i)}_S \left\{ \frac{1}{|\mathcal{N}_{\setminus S}|} \sum_{t \in \mathcal{N}_{\setminus S}} f^{(i)}(v_t, \theta^{(i)}_S, \sigma^{(i)}_S, \eta^{(i)}) \right\}$$

s.t.

$$h^{(i)}(\theta^{(i)}_S, \sigma^{(i)}_S, \eta^{(i)}) = 0.$$

The averaged validation score of each model

$$Q^{(i)}_{\text{valid}} = \frac{1}{{}_rC_k} \sum_{S \subset \{1,2,\ldots,r\}: |S| = r-k} Q^{(i)}_{S,\text{valid}}$$

is compared and the model of smaller averaged validation score is selected.

Remark 1.5.1 (consistency of GMM-MPEC) *From* (1.19) *and Theorem 1.2.1, the consistency of GMM-MPEC with the same assumption on the moment directly follows.*

1.6 Application: Dynamic Demand and Dynamic Pricing Model on Online Retailer Data

In this section, we perform our model selection procedure in a structural model with a real-world dataset. The models we compare are dynamic and static demand and pricing model that are taken from (Conlon, 2012). In particular, we first apply our cross-validation algorithm to test either the state-of-the-art dynamic demand model ((Gowrisankaran and Rysman, 2012)) or the traditional static demand model ((Berry et al., 1995)) has stronger explanatory power in the consumer behavior. To this aim, we use monthly sales and price

data of an online-retail shop. Furthermore, we consider supply side dynamics of pricing that takes the seasonality and consumer skimming into consideration such as (Nair, 2007): We investigate whether or not such a model explains the observed pricing pattern better than traditional static profit maximization model that is based on the consumer model selected in the previous step.

Structural estimation of a dynamic model has been an important frontier in industrial organization, both on demand side and supply side. On demand side, dynamic model of consumer behavior has been widely applied by researchers recently ((Gowrisankaran and Rysman, 2012)). The underlying idea in the dynamic demand model is that consumers are forward-looking regarding the changes in the market such as price and make a dynamic decision by considering the future market state. Such a model is justified by the fact that important parameters such as price elasticity could be severely mis-estimated by ignoring the forward-looking behavior of consumers. Meanwhile, similar mis-estimation would occur if a researcher applies a dynamic model in the case the consumers are in fact myopic. From a market level data, it is not directly visible if consumers are forward-looking or myopic.

Contrary to the demand side, dynamic pricing in supply side has a long history of theoretical studies dating back to (Coase, 1972). Nevertheless, little empirical attention is paid until recent years ((Nair, 2007)). Under certain conditions, firms have the incentive to determine current price by taking its effect on the future profit into consideration. For example, when consumers are heterogeneous in an evaluation of a product, firms are motivated to ”skim” high-evaluation consumers in earlier periods by setting a high price and later lower it. With myopic consumers (as in (Luo, 2015)), the pricing decision boils down to a dynamic programming of a firm in the case of monopoly or a dynamic game between firms in the case of oligopoly. If the consumers are also forward-looking, the pricing boils down to a dynamic game between consumers and firms as studied in (Nair, 2007). In this case, the observed price and demand are interpreted as a result of dynamic equilibrium.

It is not straightforward to infer if the pricing is dynamic or not from the market level data. A declining tendency on the price does not always indicate that firms are

making pricing decision dynamically: If the consumers are heterogeneous in either product evaluation or price sensitivity and leave market after purchase, a myopic optimal price may be decreasing in periods since the remaining consumers are more price elastic.

Applying dynamic pricing model to data generated from myopic pricing would cause a significant bias in the estimates of supply-side parameters such as marginal cost. For instance, a dynamic pricing model may interpret an observed high price in a certain period as a firm sparing some demand for the future, while it is a result of high marginal cost in truth. Therefore, estimation of supply-side model parameters such as marginal cost requires researchers to know if firms are myopic or forward-looking.

As it is important to correctly specify the dynamic feature of the agent's decision making both on demand and supply side, researchers are encouraged to verify whether the decision making is static or dynamic from the data rather than appealing to intuition, desirably based on real-world datasets. Regarding this aspect, we demonstrate our cross-validation algorithm to compare two by two alternative models; dynamic or myopic consumers, and dynamic or myopic firms. The models are estimated via GMM-MPEC. We take a simple dynamic model from (Conlon, 2012).

We perform estimation and model selection on a dataset of price and sales of an online-retailer based in UK. The data is taken from the University of California Irvine (UCI) Machine Learning Repository (henceforth, UCI). UCI repository consists of more than 300 datasets. The data used in this study is available here at `https://archive.ics.uci.edu/ml/datasets/Online+Retail` free of charge. We consider the use of such a publicly available dataset increases a reproducibility of a research process. In machine learning field, researchers are encouraged to compare the performance of a newly proposed model or algorithm to old ones with a publicly available dataset, and the UCI repository is widely used in this aim.

1.6.1 Models

We consider models of 2 by 2 design: static or dynamic demand, static or dynamic pricing. We denote each model as $m \in \{1, 2, 3, 4\}$, where $m = 1, 2$ assume static demand, $m = 3, 4$ assume dynamic demand, $m = 1, 3$ assume static pricing, and $m = 2, 4$ assume dynamic pricing. For simplicity, we assume that the firm and consumers make their purchase decision independently across products. It is entirely possible to test if this assumption is valid or not using our CV algorithm, but we omit it as the main purpose of this section is an illustration of model selection procedure. The consumers are heterogeneous in price sensitivity and the constant term of utility as in random coefficients model. We assume that consumers make a purchase at most once for each product within the considered period. This assumption is justified by the transaction level data. Among all the transactions used in the data, 75.8% of them are made by consumers who purchased the same product only once in the considered period. An alternative approach is to model repeated purchase and inventory behavior explicitly as in (Hendel and Nevo, 2006), but we do not take this path for tractability.

Demand Model

In each period, consumers in the market decide whether to purchase a product or not to maximize their objective function. If the demand is assumed to be static, the objective function is simply the utility function defined below. If the demand is dynamic, the objective function is the infinite-period sum of discounted utility.

Denote products as $j = 1, ..., J$ and period as $t = 1, ..., T$. Consumer i's utility of purchasing a product j at period t is

$$\begin{aligned} u_{ijt} &= \alpha_i^p p_{jt} + \alpha_{ij}^0 + \mathbf{X}_{jt}\boldsymbol{\alpha}^x + \xi_{jt} + \ _{ijt}. \\ &\equiv \delta_{ijt} + \ _{ijt} \end{aligned}$$

where p_{jt} is the price of a product j in period t, $\mathbf{X}_{jt}$ is the observable characteristics, and ξ_{jt} is the i.i.d preference shock, which enters the moment conditions. $_{ijt}$ is the logit error

term that follows type-I extreme value distribution and i.i.d across periods and products. The utility of not purchasing is $u_{i0t} = \ _{i0t}$ as the non-random component is normalized to be zero. The random coefficients follow a normal distribution.

$$\alpha_i^p = \alpha^p + \nu_i^p \rho^p$$
$$\alpha_i^0 = \alpha^0 + \nu_i^0 \rho^0$$

,where (α^p, α^0) are the population mean of the utility coefficients, ν_i^p and ν_i^0 are draws from a standard normal distribution, and (ρ^p, ρ^0) are the standard deviation of the distribution of the random coefficients.

In the static demand model, the consumers simply compare the utility of purchase to non-purchase in each period. Thus the purchase probability is

$$s_{ijt}^m = \frac{\exp(\delta_{ijt})}{\exp(\delta_{ijt}) + 1} \quad \text{for } m = 1, 2.$$

In the dynamic demand model, the consumers make purchase decision by comparing the instant utility to the value of waiting until next period. Let Ω_{ijt}^d be a state space for a consumer i on product j at period t and $W_{ij}(\Omega_{ijt}^c)$ be a value function associated to the state. The Bellman equation is expressed as

$$W_{ij}(\Omega_{ijt}^d) = \max\{u_{ijt}, u_{i0t} + \beta \mathbb{E}[W_{ij}(\Omega_{ijt+1}^d)|\Omega_{ijt}^d]\}.$$

The purchase probability of product j of a consumer i at period t is

$$s_{ijt}^m(\Omega_{ijt}^d) = \frac{\exp(\delta_{ijt})}{\exp(\delta_{ijt}) + \exp(\beta \mathbb{E}[W(\Omega_{ijt+1}^c)|\Omega_{ijt}^c])} \quad \text{for } m = 3, 4.$$

Following (Conlon, 2012), we make an assumption that consumers have perfect foresight over a transition of state Ω_{ijt}^d. Formally,

$$\mathbb{E}[W(\Omega_{ijt+1}^d)|\Omega_{ijt}^d] = w_{ijt+1},$$

where

$$w_{it} = \ln(\exp(\delta_{ijt}) + \exp(\beta w_{it+1}))$$

for all i, j, and t. The second line is a direct consequence of the first line following the argument of (Rust, 1987). An alternative and more popular specification is to assume that consumers form an expectation of the future state by certain functional form, typically an AR(1) regression. Compared to functional assumption perfect foresight reduces the computational burden significantly as it avoids integration over a distribution for calculating expectation (See (Conlon, 2012) for further discussion.) Also, note that by our CV algorithm we can even investigate which of perfect foresight and AR(1) assumption makes the model more accurate, which we believe is an interesting future work.

Finally, for both static and dynamic model let M_{ijt} be the market size of consumers for a product j at period t. Given the consumers purchase the same product at most once, the market size transition for any model $m \in \{1, 2, 3, 4\}$ follows

$$M_{ijt+1}^{(m)} = M_{ijt}(1 - s_{ijt}^{(m)}).$$

Supply Model

We express the marginal cost of product j at period t for the retailer as MC_{jt} where

$$MC_{jt} = \mathbf{Y}_{jt}\boldsymbol{\gamma}_{jt} + \lambda_{jt}.$$

X_{jt}^{cost} is the observable characteristics of the product, and λ_{jt} is the cost shock i.i.d across time and products.

Denote the states of a product j for the retailer at period t as $\mathbf{\Omega}_{jt}^{s}$. $\mathbf{\Omega}_{t}^{s}$ includes the market size of each consumer segment $\{M_{ijt}\}_i$ and the draw of unobserved utility shock, $\{\xi_{ijt}\}_i$ and λ_{jt}. Given the demand system described above, the demand function is written as

$$D_{jt}^{(m)}(p_{jt}, \mathbf{\Omega}_{t}^{s}) = \sum_{r=1}^{R} M_{ijt}^{(m)} s_{ijt}^{(m)}.$$

The instant profit function of a product j at period t is therefore

$$\pi_{jt}(p_{jt}, \mathbf{\Omega}^s_{jt}) = D_{jt}(p_{jt}, \mathbf{\Omega}^s_{jt})(p_{jt} - MC_{jt}),$$

In static pricing model, $m = 1, 3$, the retailer simply chooses the price to maximize the myopic profit:

$$p^m_{jt} = \arg\max_{p_{jt}} \pi_{jt}(p_{jt}) \forall j, t$$
$$\text{for } m = 1, 3.$$

In dynamic pricing model ($m = 2, 4$), the retailer maximizes the net profit over time with discounting. The discounting factor β is assumed to be same with consumers. The retailer determines the price after observing the realization of the shocks, $\{\xi_{ijt}\}_i$ and λ_{jt}. The value function of a product j is expressed as

$$V_j(\Omega^s_t) = \mathbb{E} \quad \max_{p_{jt}} \quad \pi_{jt} + \beta V_j(\Omega^s_{t+1}) \quad \mathbf{\Omega}^f_t, p_{jt} \quad ,$$

where the expectation is over the unobserved cost shock in the next period, λ_{jt+1}. The optimal price is determined as

$$p^m_{jt} = \arg\max_{p_{jt}} \pi_{jt}(p_{jt}) + \beta \mathbb{E}[V_j(\Omega^s_{t+1})|\Omega^s_t]$$
$$\text{for } m = 2, 4.$$

Similar to the demand side, we assume that the retailer has a perfect information on the transition of the error draw.

Equilibrium

This section describe the equilibrium condition for each model. When consumers and firms are both static ($m = 1$), the equilibrium price and demand are the standard one as in many models such as (Berry et al., 1995). When consumers are static but firms are dynamic ($m = 2$), pricing can be seen as a single agent dynamic optimization problem with continuous choice variable p_{jt}. Similarly, when consumers are dynamic but firms are static

($m = 3$), consumers solve a single agent dynamic optimization problem. The consumers problem is an optimal stopping problem as the choice is the timing of purchase. When both consumers and the retailer are both dynamic ($m = 4$), we assume their behavior is at Markov Perfect Nash Equilibrium (MPNE) where consumers' and retailer's prediction of the value function matches to the realization.

1.6.2 Data

We obtain our data from UCI machine Learning Repository. The UCI Machine Learning Repository maintains more than three hundreds datasets that are intensely used by machine learning community for empirical investigation and comparison of algorithms. When researchers propose a new model or algorithm in machine learning field, a common practice is to test its performance on the dataset in this repository. Such a culture gives a thorough idea on the practical performance of existing models and algorithms. Moreover, it helps a new researcher replicate the results on the existing papers.

The dataset we utilize in this study is the online retail data created by (Chen et al., 2012), posted on UCI Machine Learning Repository in November 2015. The data is publicly available at `https://archive.ics.uci.edu/ml/datasets/Online+Retail`. The information about the data source is provided by the authors as follows: "The online retailer under consideration is a UK-based and registered non-store business with some 80 members of staff. The company was established in 1981 mainly selling unique all-occasion gifts. For years in the past, the merchant relied heavily on direct mailing catalogs, and orders taken over phone calls. It was only 2 years ago that the company launched its own web site and shifted completely to the web. Since then the company has maintained a steady and healthy number of customers. The company also uses Amazon.co.uk to market and sell its products."

The data include all the transactions occurred on this retailer from December 2010 to December 2011. Each transaction information includes quantity, unit price, consumer ID, and country. We dropped any sales to outside UK. The majority of the sales is inside UK

and non-UK sales has only limited amount (approximately 20%.) Since our purpose is to demonstrate application of CV model selection to static and dynamic models, we aggregate the data into a monthly sales of each product so that the data format follows typical market level data and we can apply commonly used economic models. The monthly sales is simply a sum of the quantity sold in a particular month. The monthly price is calculated as the average of the price of transaction occurred in each month weighted by the quantity. We omitted the products that have any zero sales in the considered months from the data.

On the top of price and sales data, the author hand-coded product category and subcategory based on the description of products. The categories include *Children*, *Decoration*, or *Kitchen*. The number of products as well as basic statistics are summarized in table 1.4. Figure 1.4 shows the average of monthly price and quantity sold in each category. It shows that the dynamics is heterogeneous across categories. For instance, the price of products in *Gift* and *Decoration* show tendency to decline over periods, while *Home and Garden* or *Candle* show more fluctuation.

1.6.3 Estimation and Model Selection

We implement model selection for the demand side and supply side sequentially. First we test if the demand is static or dynamic. Subsequently, we test if the pricing is static or dynamic, assuming the demand model chosen in the previous step. The endogenous variables such as the market size M_{ijt} and the share s_{ijt} are estimated in the demand side, and imported over to the supply side estimation. Importantly, we do not have to specify the pricing model on estimation of demand side by virtue of perfect foresight assumption. We treat the data in each category independently.

We adapt 3-fold cross validation, $(k, r) = (1, 3)$. Because the data has a panel structure of products and periods, either the product-wise or period-wise split is possible. We adopt split based on products. That is, we split the products into three groups, and use two of them to estimate a model and use the last one for validation.

MPEC formulation

To estimate each model by GMM-MPEC, we formulate the estimation as a minimization problem of GMM objective with equilibrium constraints based on the model described above. Under the assumptions we impose, the equilibrium constraints are convex and mostly either linear or quadratic. This fact ensures that we are able to find an optimal solution of the estimation problem.

First we describe the MPEC formulation of demand models. For the static demand model ($m = 1, 2$), the set of constraints are

$$\begin{aligned} s_{ijt}^m &= \frac{\exp(\delta_{ijt})}{\exp(\delta_{ijt}) + 1} \\ D_{jt}^m &= \sum_i M_{ijt}^m s_{ijt}^m \\ \delta_{ijt} &= \alpha_i^p p_{jt} + \alpha_{ij}^0 + \mathbf{X}_{jt}\boldsymbol{\alpha}^x + \xi_{jt} \\ \alpha_i^p &= \alpha^p + \nu_i^p \rho^p \\ \alpha_i^0 &= \alpha^0 + \nu_i^0 \rho^0 \\ M_{ijt}^m &= M_{ijt-1}^m (1 - s_{ijt}^m), \end{aligned} \tag{1.20}$$

for all (i, j, t).

For dynamic demand model, the constraints are similar except the consumers compare the purchase utility to the value of waiting until next period.

$$\begin{aligned} s_{ijt}^m &= \frac{\exp(\delta_{ijt})}{\exp(\delta_{ijt}) + \exp(\beta w_{it+1})} \\ w_{it} &= \ln(\exp(\delta_{it}) + \exp(\beta w_{it+1})) \\ D_{jt} &= \sum_i M_{ijt}^m s_{ijt}^m \\ \delta_{ijt} &= \alpha_i^p p_{jt} + \alpha_{ij}^0 + \mathbf{X}_{jt}\boldsymbol{\alpha}^x + \xi_{jt} \\ \alpha_i^p &= \alpha^p + \nu_i^p \rho^p \\ \alpha_i^0 &= \alpha^0 + \nu_i^0 \rho^0 \\ M_{ijt}^m &= M_{ijt-1}^m (1 - s_{ijt}^m) \end{aligned} \tag{1.21}$$

for all (i, j, t).

The model parameters to estimate are $\theta^d = (\alpha^p, \alpha^0, \rho^p, \rho^0)$. The data to input are the realized demand D_{jt}, the observed price p_{jt}, and the random draws ν_i^p and ν_i^0. The predicted share s_{ijt}, the market size of each consumer type M_{ijt}, and the error draw ξ_{jt} are the endogenous variables. In the dynamic demand model, the value function w_{ijt} is also observation-specific endogenous variable to choose for the optimization.

We define the supply side estimation problem by the first order condition and the Bellman equation. By abusing notation, let $D_{jt}^m(p)$ as a demand function with respect to price in model m. The supply side equilibrium constraints of static pricing model is that the observed prices are chosen to maximize the instant profit:

$$\begin{aligned} D_{jt}^m &= \sum_i M_{ijt}^m s_{ijt} \\ MC_{jt}^m &= X_{jt}^s \gamma + \lambda_{jt} \\ p_{jt} &= \arg\max_p [D_{jt}^m(p)(p - MC_{jt}^m)] \end{aligned} \tag{1.22}$$

for all (i, j, t).

Instead of the third line above, the dynamic pricing model includes Bellman equation:

$$\begin{aligned} D_{jt}^m &= \sum_i M_{ijt}^m s_{ijt} \\ MC_{jt} &= X_{jt}^s \gamma + \lambda_{jt} \\ p_{jt} &= \arg\max_p [D_{jt}^m(p)(p - MC_{jt}) + \beta V_{jt+1}(\Omega_{jt+1}^s)] \\ V_{jt}(\Omega_{jt}^s) &= \max_p [D_{jt}^m(p)(p - MC_{jt}) + \beta V_{jt+1}(\Omega_{jt+1}^s)]. \end{aligned} \tag{1.23}$$

The model parameters to estimate is $\theta^s = \gamma$. MC_{jt}, λ_{jt}, and the value function are observation-specific endogenous variables. M_{ijt} and s_{ijt} are estimated in the demand side as endogenous variables.

In both static and dynamic model, the constraint includes the retailer's optimization problem. We convert it to the first order condition when solving for the estimation. The details are in the Appendix.

The GMM objective is a function defined by moment conditions

$$\mathbb{E}[\xi Z] = 0$$
$$\mathbb{E}[\lambda Z] = 0,$$

where Z is the instrumental variables. It includes category and subcategory dummies, period dummy, and the market size of consumer segments $\{M_{it}\}_i$. The market size information is correlated with price because it relates to the price elasticity. Since we assume that the unobserved shocks are not serially correlated, the market size at period t is not correlated with the shocks in the same period. Further detail of the setting for estimation is described in the Appendix.

1.6.4 Results

Table 1.5 presents the cross validation score of each model. The second from the last column shows the demand model selected by CV. The last column exhibits the selected pricing model. One can see that the selected model varies across categories. On demand side, the data on *Children Decoration*, and *Kitchen* are explained better by the static model, while the dynamic model is preferred on other categories. On supply side, static pricing explained the data of *Crafts*, *Decoration*, and *Personal Item* better.

The result of model selection is difficult to interpret. One could try to provide some intuition: For instance, the products that fits static demand model better may be the ones that consumers cannot make a consumption plan. On products where the retailer engages in static pricing, it may be due to certain circumstance that researchers do not observe, such as a contract with wholesaler or limitation of inventory. However, prior to observing the result of cross validation, it is hard to make an reliable and scientific argument and justification for any model to be realistic.

The difficulty of interpretation in turn suggests that it is impractical for researchers to assume a certain model beforehand. Selecting a structural model based on intuition may severely bias the inference. To see the problem, 1.6 shows the estimated price coefficient in each category in different specification. While in some cases two models exhibit fairly

similar result, in some cases such as *Candle* or *Party* the result is largely different. Therefore, we recommend that researchers cross validate their models whenever possible, unless they have a strong reason to believe in certain model.

1.7 Conclusion

In this paper, we have proposed a cross-validation approach to model selection when models are estimated via GMM criterion. Cross-validation procedure can be readily implemented in any existing economic models without much extra work for researchers. We have proved its asymptotic consistency, and Monte-Carlo experiments in both linear and non-linear model confirm that cross-validation outperforms in-sample comparison that economists traditionally practice.

We also proposed a way to apply cross-validation when models are estimated through MPEC. As its real-world application, we adapt our CV based model selection to test dynamic demand model and dynamic pricing model in an online-retailer data. We find a quite diverse result across product categories. Unexpectedly, even on the same retailer it is not consistent whether a dynamic model is preferred or not. As the implication of structural estimation largely depends on the assumed model, this result suggests that economists should cross-validate their structural models rather than appealing to for reliability of their inference.

Algorithm 1 (k, r)-Cross Validation on GMM

1: Input: Models $\{\mathcal{M}_i\}$, data $\{v_t\}_{t=1,\ldots,T}$.

2: **for** each model $\mathcal{M}_i$ **do**

3: **for** each training data $\{v_t\}_{t\in N_S}$ **do**

4: Estimate model parameters as

$$\theta_S^{(i)} = \underset{\theta^{(i)}\in\Theta^{(i)}}{\arg\min}\, Q_S^{(i)}(\theta^{(i)})$$

5: Calculate the score $Q_{S,\text{valid}}^{(i)}(\theta_S^{(i)})$

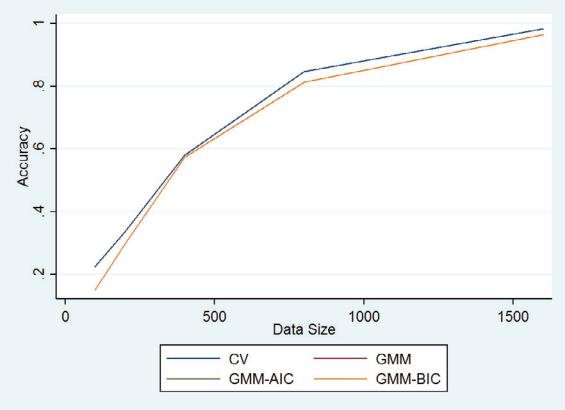
6: **end for**

7: Calculate the average score

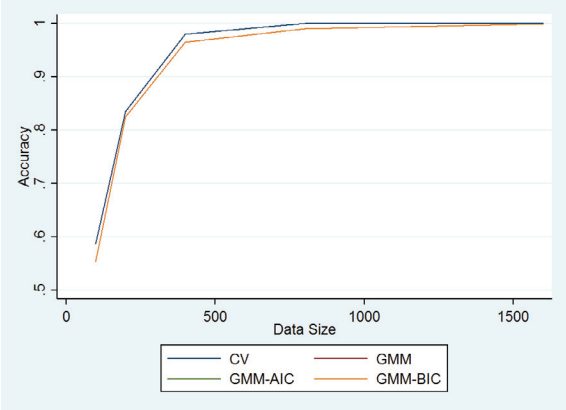
$$Q_{\text{valid}}^{(i)} = \frac{1}{{}_rC_k} \sum_{S\subset\{1,2,\ldots,r\}:|S|=r-k} Q_{S,\text{valid}}^{(i)}(\theta_S^{(i)})$$

8: **end for**

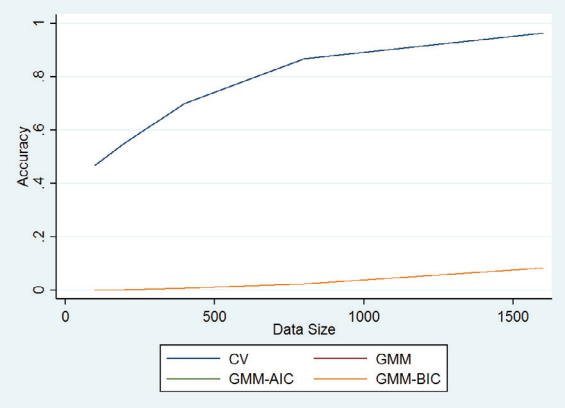
9: Find the best model that exhibits the smallest $Q_{\text{valid}}^{(i)}$.



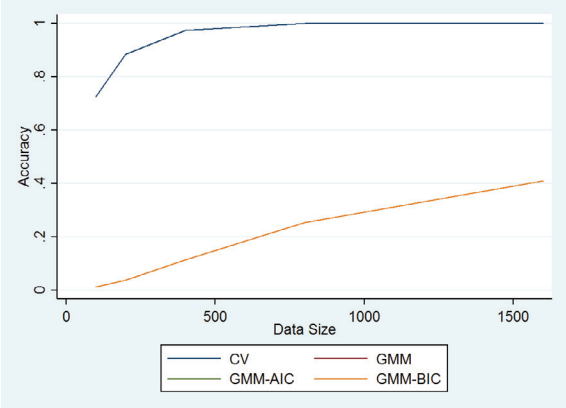
(a) $p^1 = 3$,$p^2 = 5$,$\alpha = 3$.



(b) $p^1 = 3$,$p^2 = 5$,$\alpha = 7$.



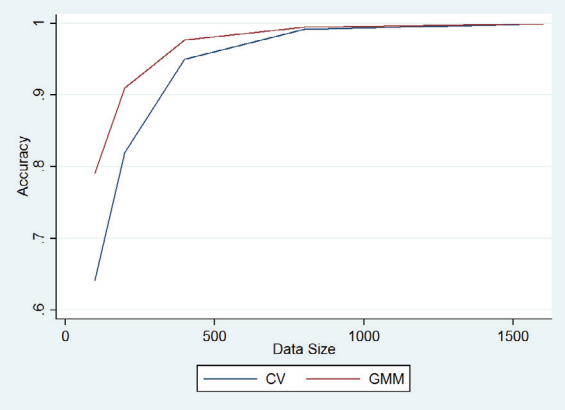
(c) $p^1 = 3$,$p^2 = 9$, $\alpha = 3$.



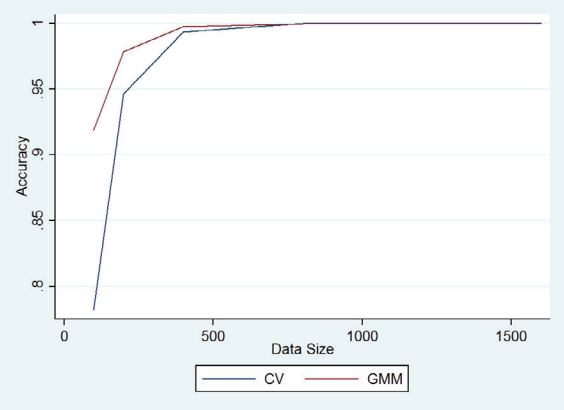
(d) $p^1 = 3$,$p^2 = 9$, $\alpha = 7$.

Figure 1.1.: The accuracy of model selection when $p^1 < p^2$.

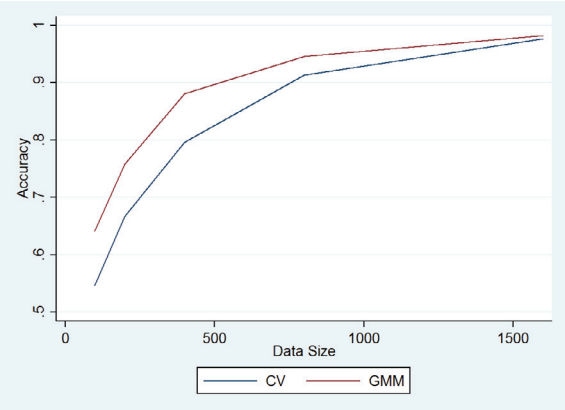
Note: The y-axis is the probability that the correctly specified model (model 1) is chosen by each procedure. The number of instruments is set to be $c^1 = c^2 = 10$. The cross-validation is 2-folds, i.e. $r = 2$. The weighting matrix is set to be identity matrix.



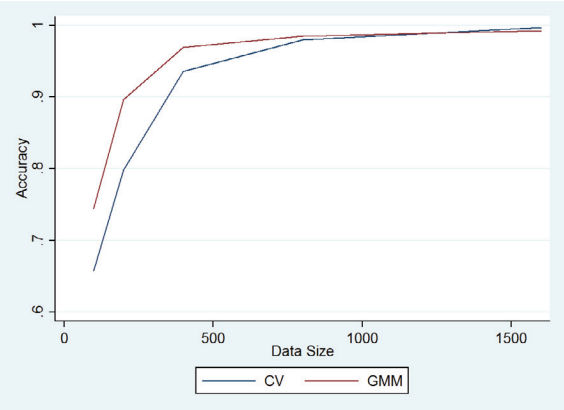
(a) $p^1 = p^2 = 3$,$\alpha = 7$.



(b) $p^1 = p^2 = 3$,$\alpha = 12$.



(c) $p^1 = p^2 = 7$, $\alpha = 7$.



(d) $p^1 = p^2 = 7$, $\alpha = 12$.

Figure 1.2.: The accuracy of model selection when $p^1 = p^2$.

Note: The y-axis is the probability that the correctly specified model (model 1) is chosen by each procedure. The number of instruments is set to be $c^1 = c^2 = 10$. The cross-validation is 2-folds, i.e. $r = 2$. The weighting matrix is set to be identity matrix.

Table 1.1.: The validation Score of CV. Average of 100 iterations (standard deviation in the bracket).

		$\alpha = -.1$				
		Candidate Model				
Number of Market	**True Model**	$\{1,2,3\}$	$\{1,2\}\{3\}$	$\{1\}\{2,3\}$	$\{1,3\}\{2\}$	$\{1\}\{2\}\{3\}$
25	$\{1,2,3\}$	**1.175**	23.732	28.822	28.799	30.709
		(1.570)	(44.224)	(89.920)	(56.014)	(60.764)
	$\{1,2\}\{3\}$	41.026	**1.022**	27.687	22.208	8.799
		(61.458)	(1.049)	(54.927)	(29.543)	(13.536)
	$\{1\}\{2\}\{3\}$	25.441	10.115	8.536	8.659	**0.912**
		(26.184)	(17.287)	(9.610)	(8.646)	(1.021)
50	$\{1,2,3\}$	**0.233**	6.892	6.505	5.708	6.686
		(0.181)	(10.774)	(5.495)	(5.073)	(7.486)
	$\{1,2\}\{3\}$	9.890	**0.314**	5.328	6.143	2.466
		(11.450)	(0.207)	(3.391)	(5.696)	(1.781)
	$\{1\}\{2\}\{3\}$	10.050	2.552	3.764	2.808	**0.274**
		(10.486)	(2.366)	(10.730)	(2.180)	(0.215)
75	$\{1,2,3\}$	**0.144**	3.436	3.470	3.272	3.580
		(0.094)	(2.328)	(2.232)	(1.865)	(2.415)
	$\{1,2\}\{3\}$	5.736	**0.170**	3.284	3.315	1.367
		(4.349)	(0.114)	(2.456)	(1.864)	(0.867)
	$\{1\}\{2\}\{3\}$	6.651	1.710	1.800	2.046	**0.190**
		(5.723)	(1.334)	(1.082)	(3.060)	(0.152)
100	$\{1,2,3\}$	**0.084**	2.314	2.475	2.371	2.374
		(0.046)	(1.195)	(1.716)	(1.418)	(1.424)
	$\{1,2\}\{3\}$	4.050	**0.124**	2.289	2.384	0.951
		(3.121)	(0.071)	(1.754)	(1.785)	(0.671)
	$\{1\}\{2\}\{3\}$	4.463	1.314	1.266	1.357	**0.124**
		(2.282)	(1.761)	(0.791)	(1.017)	(0.081)
		$\alpha = -.3$				
25	$\{1,2,3\}$	**1.139**	2.531	2.355	1.906	1.704
		(2.269)	(3.149)	(3.209)	(2.087)	(1.771)
	$\{1,2\}\{3\}$	8.223	**1.174**	3.528	4.746	1.646
		(13.529)	(1.405)	(3.663)	(8.142)	(2.184)
	$\{1\}\{2\}\{3\}$	10.684	3.272	3.175	4.273	**1.190**
		(16.949)	(4.735)	(3.757)	(9.404)	(1.373)
50	$\{1,2,3\}$	**0.281**	0.651	0.661	0.640	0.643
		(0.256)	(0.439)	(0.442)	(0.416)	(0.461)
	$\{1,2\}\{3\}$	1.713	**0.319**	0.970	1.165	0.396
		(1.448)	(0.229)	(0.742)	(0.947)	(0.278)
	$\{1\}\{2\}\{3\}$	3.628	0.998	1.056	1.041	**0.365**
		(15.427)	(1.742)	(2.198)	(1.949)	(0.607)
75	$\{1,2,3\}$	**0.159**	0.387	0.387	0.426	0.356
		(0.110)	(0.233)	(0.257)	(0.431)	(0.214)
	$\{1,2\}\{3\}$	1.096	**0.210**	0.623	0.574	0.238
		(0.684)	(0.164)	(0.370)	(0.384)	(0.150)
	$\{1\}\{2\}\{3\}$	1.303	0.504	0.467	0.464	**0.169**
		(1.023)	(0.336)	(0.288)	(0.322)	(0.117)
100	$\{1,2,3\}$	**0.103**	0.261	0.255	0.277	0.258
		(0.060)	(0.134)	(0.135)	(0.165)	(0.124)
	$\{1,2\}\{3\}$	0.743	**0.134**	0.419	0.414	0.157
		(0.393)	(0.079)	(0.244)	(0.246)	(0.094)
	$\{1\}\{2\}\{3\}$	0.937	0.317	0.335	0.329	**0.108**
		(0.464)	(0.170)	(0.243)	(0.253)	(0.068)

Table 1.2.: The Model Selection Probability with CV.

		$\alpha = -.1$					
		Candidate Model					
Number of Market	**True Model**	$\{1,2,3\}$	$\{1,2\}\{3\}$	$\{1,3\}\{2\}$	$\{1\}\{2,3\}$	$\{1\}\{2\}\{3\}$	**true**
25	$\{1,2,3\}$	**0.99**	0.00	0.00	0.01	0.00	0.99
	$\{1,2\}\{3\}$	0.00	**0.95**	0.00	0.00	0.05	0.95
	$\{1\}\{2\}\{3\}$	0.00	0.01	0.00	0.00	**0.99**	0.99
50	$\{1,2,3\}$	**1.00**	0.00	0.00	0.00	0.00	1.00
	$\{1,2\}\{3\}$	0.00	**0.99**	0.00	0.00	0.01	0.99
	$\{1\}\{2\}\{3\}$	0.00	0.00	0.00	0.00	**1.00**	1.00
75	$\{1,2,3\}$	**1.00**	0.00	0.00	0.00	0.00	1.00
	$\{1,2\}\{3\}$	0.00	**1.00**	0.00	0.00	0.00	1.00
	$\{1\}\{2\}\{3\}$	0.00	0.00	0.00	0.00	**1.00**	1.00
100	$\{1,2,3\}$	**1.00**	0.00	0.00	0.00	0.00	1.00
	$\{1,2\}\{3\}$	0.00	**1.00**	0.00	0.00	0.00	1.00
	$\{1\}\{2\}\{3\}$	0.00	0.00	0.00	0.00	**1.00**	1.00
		$\alpha = -.3$					
25	$\{1,2,3\}$	**0.62**	0.04	0.04	0.11	0.19	0.62
	$\{1,2\}\{3\}$	0.00	**0.62**	0.01	0.03	0.34	0.62
	$\{1\}\{2\}\{3\}$	0.00	0.07	0.05	0.09	**0.79**	0.79
50	$\{1,2,3\}$	**0.77**	0.05	0.04	0.10	0.04	0.77
	$\{1,2\}\{3\}$	0.00	**0.64**	0.01	0.00	0.35	0.64
	$\{1\}\{2\}\{3\}$	0.00	0.01	0.04	0.04	**0.91**	0.91
75	$\{1,2,3\}$	**0.78**	0.05	0.07	0.06	0.04	0.78
	$\{1,2\}\{3\}$	0.00	**0.57**	0.02	0.00	0.41	0.57
	$\{1\}\{2\}\{3\}$	0.00	0.05	0.01	0.03	**0.91**	0.91
100	$\{1,2,3\}$	**0.82**	0.04	0.09	0.02	0.03	0.82
	$\{1,2\}\{3\}$	0.00	**0.64**	0.00	0.00	0.36	0.64
	$\{1\}\{2\}\{3\}$	0.00	0.04	0.02	0.01	**0.93**	0.93

Table 1.3.: The Model Selection Probability with GMM.

		$\alpha = -.1$					
		Candidate Model					
Number of Market	**True Model**	$\{1,2,3\}$	$\{1,2\}\{3\}$	$\{1,3\}\{2\}$	$\{1\}\{2,3\}$	$\{1\}\{2\}\{3\}$	**true**
	$\{1,2,3\}$	**0.99**	0.01	0.00	0.00	0.00	0.99
25	$\{1,2\}\{3\}$	0.00	**0.95**	0.00	0.00	0.05	0.95
	$\{1\}\{2\}\{3\}$	0.00	0.02	0.01	0.02	**0.95**	0.95
	$\{1,2,3\}$	**1.00**	0.00	0.00	0.00	0.00	1.00
50	$\{1,2\}\{3\}$	0.00	**0.92**	0.02	0.00	0.06	0.92
	$\{1\}\{2\}\{3\}$	0.00	0.00	0.00	0.03	**0.97**	0.97
	$\{1,2,3\}$	**1.00**	0.00	0.00	0.00	0.00	1.00
75	$\{1,2\}\{3\}$	0.00	**0.97**	0.00	0.00	0.03	0.97
	$\{1\}\{2\}\{3\}$	0.00	0.00	0.00	0.00	**1.00**	1.00
	$\{1,2,3\}$	**1.00**	0.00	0.00	0.00	0.00	1.00
100	$\{1,2\}\{3\}$	0.00	**0.96**	0.00	0.00	0.04	0.96
	$\{1\}\{2\}\{3\}$	0.00	0.01	0.00	0.00	**0.99**	0.99
		$\alpha = -.3$					
	$\{1,2,3\}$	**0.61**	0.10	0.09	0.09	0.11	0.61
25	$\{1,2\}\{3\}$	0.00	**0.66**	0.00	0.02	0.32	0.66
	$\{1\}\{2\}\{3\}$	0.00	0.06	0.05	0.05	**0.84**	0.84
	$\{1,2,3\}$	**0.69**	0.10	0.13	0.07	0.01	0.69
50	$\{1,2\}\{3\}$	0.01	**0.54**	0.01	0.03	0.41	0.54
	$\{1\}\{2\}\{3\}$	0.00	0.02	0.04	0.05	**0.89**	0.89
	$\{1,2,3\}$	**0.77**	0.09	0.07	0.06	0.01	0.77
75	$\{1,2\}\{3\}$	0.00	**0.44**	0.00	0.00	0.56	0.44
	$\{1\}\{2\}\{3\}$	0.00	0.03	0.00	0.06	**0.91**	0.91
	$\{1,2,3\}$	**0.77**	0.05	0.11	0.06	0.01	0.77
100	$\{1,2\}\{3\}$	0.00	**0.39**	0.01	0.01	0.59	0.39
	$\{1\}\{2\}\{3\}$	0.00	0.01	0.00	0.02	**0.97**	0.97

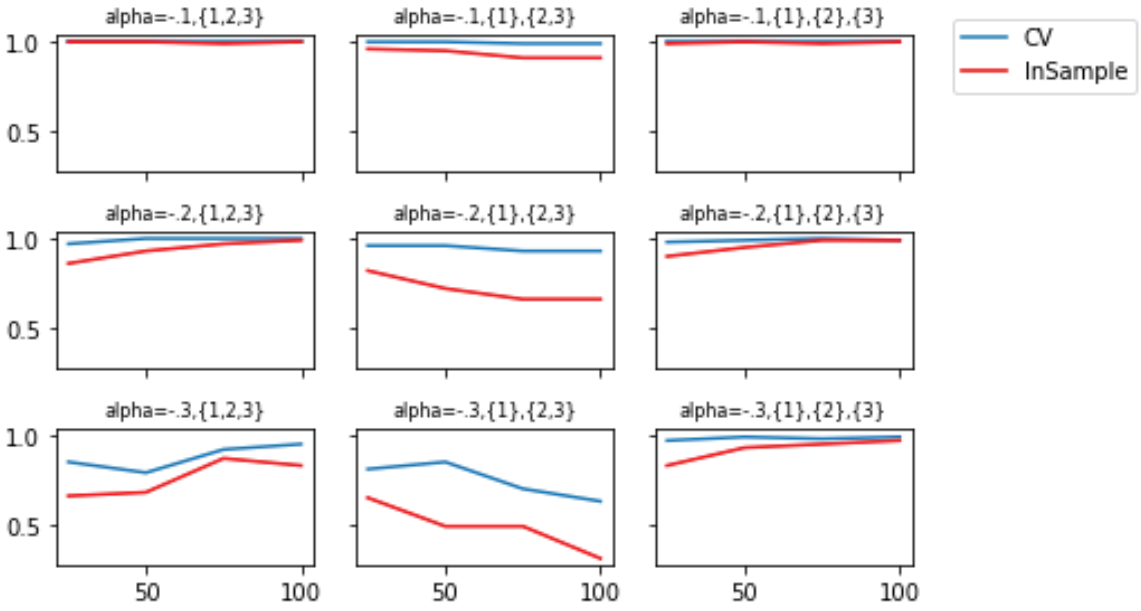


Figure 1.3.: The choice probability of true model on CV and GMM model selection.

Algorithm 2 (k, r)-Cross Validation on GMM-MPEC

1: Input: Models $\{\mathcal{M}_i\}$, data $\{v_t\}_{t=1,\ldots,T}$.

2: **for** each model $\mathcal{M}_i$ **do**

3: **for** each training data $\{v_t\}_{t\in N_S}$ **do**

4: Estimate model parameters as

$$(\theta_S^{(i)}, \sigma_S^{(i)}, \eta_S^{(i)}) = \underset{\theta^{(i)},\sigma^{(i)},\eta^{(i)}}{\arg\min}\ Q_S^{(i)}(\theta^{(i)}, \sigma^{(i)}, \eta^{(i)})$$
$$\text{s.t. } h(\theta^{(i)}, \sigma^{(i)}, \eta^{(i)}) = 0.$$

5: Calculate the score as

$$Q_{S,\text{valid}}^{(i)}(\theta_S^{(i)}) = \min_{\eta^{(i)}} Q_{\setminus S}(\theta_S^{(i)}, \sigma_S^{(i)}, \eta^{(i)})$$
$$\text{s.t. } h(\theta_S^{(i)}, \sigma_S^{(i)}, \eta^{(i)}) = 0.$$

6: **end for**

7: Calculate the average score

$$Q_{\text{valid}}^{(i)} = \frac{1}{{}_rC_k} \sum_{S\subset\{1,2,\ldots,r\}:|S|=r-k} Q_{S,\text{valid}}^{(i)}(\theta_S^{(i)})$$

8: **end for**

9: Find the best model that exhibits the smallest $Q_{\text{valid}}^{(i)}$.

Table 1.4.: Summary of Online-Retail Data

Category	Example of Products	# Products	Ave. Unit Price (USD)	Ave. Monthly Sales (Thousand)
Candle	Candles, Candle Holder, Candle Plate	77	1.944	0.232
Children	Baby Bib, Doll, Stationery Set	175	4.122	0.148
Crafts	Knitting, Patches, Flannel, Sketchbook	38	2.694	0.214
Decoration	Photo frame, Flower, Decorative Signs	153	2.454	0.1954
Gift	Gift boxes, Tape, Message cards	65	0.7881	0.207
Home and Garden	Lamp,Cushion,Bath Salt	199	4.342	0.196
Kitchen	Mug, Tea Set, Lunch box	247	3.352	0.189
Party	Balloons,Napkins, Paper cup	75	2.432	0.197
Personal	Umbrella, Ring, Shopping bag	109	2.864	0.159

Figure 1.4.: The price and quantity dynamics of online retail data in each category.

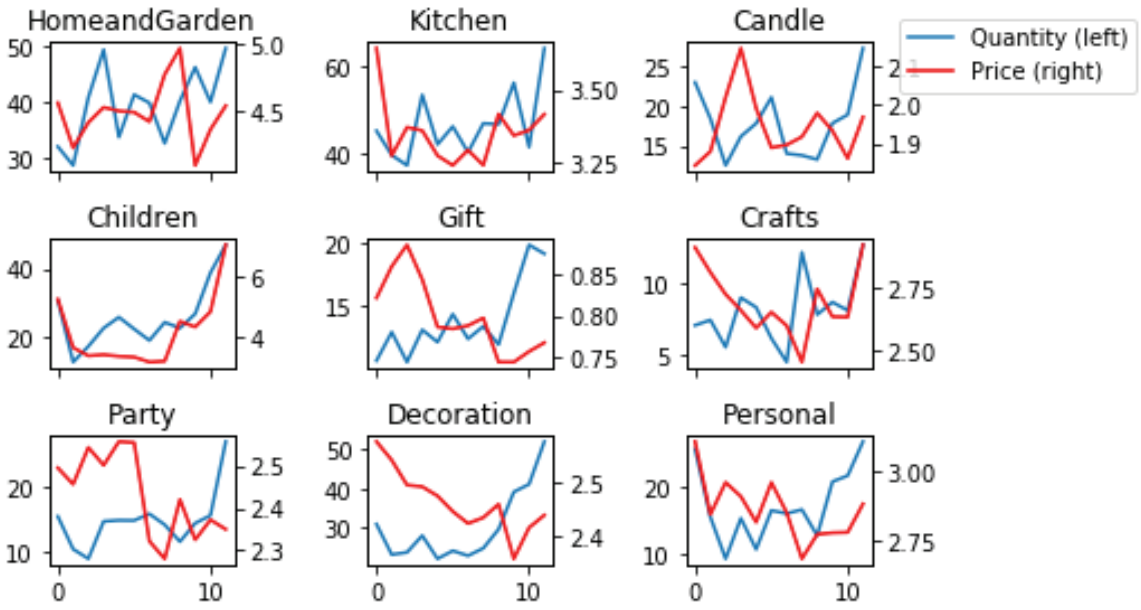


Table 1.5.: CV score in different categories

	Demand		Pricing		Selected Model	
Category	Static	Dynamic	Static	Dynamic	Demand	Pricing
Candle	.00683	.00651	.015302	.011402	Dyn	Dyn
Children	.00913	17.9	.982	.376887	Stat	Dyn
Crafts	.00847	.00655	.003628	.004258	Dyn	Stat
Decoration	.00162	.00163	.000454	.000644	Stat	Stat
Gift	.00328	.00277	.000288	.000119	Dyn	Dyn
Home and Garden	.00177	.00109	.053322	0.022654	Dyn	Dyn
Kitchen	.00152	.00158	.002165	.000763	Stat	Dyn
Party	.00795	.00310	.016513	.002486	Dyn	Dyn
Personal Item	.00305	.00193	.003356	0.003443	Dyn	Stat

Table 1.6.: Estimated price coefficient in different categories

Category	Static model	Dynamic model
Candle	-6.52515	-1.71062
Children	-0.01343	-0.01835
Crafts	-2.12241	-0.66246
Decoration	-1.26782	-0.83267
Gift	-3.28775	-3.42115
Home and Garden	-0.1841	-0.4384
Kitchen	-0.56447	-0.53658
Party	-5.37787	-1.18728
Personal Item	-0.78579	-0.83831

2. TWO-STAGE ALGORITHM FOR DISCRIMINATION-FREE MACHINE LEARNING

2.1 Introduction

Algorithmic decision making process based on machine learning now affects many aspects of our lives. Emails are spam-filtered by classifiers, images are automatically tagged and sorted, and news articles are clustered and ranked. These days, even decisions regarding individual people are being made algorithmically. For example, computer-generated credit scores are popular in many countries, and job interviewees are sometimes evaluated by assessment algorithms[1]. However, a potential loss of transparency, accountability, and fairness arises when decision making is conducted on the basis of past data. For example, if a dataset indicates that specific groups based on sensitive variables (e.g., gender, race, and religion) are of higher risk in receiving loans, direct application of machine learning algorithm would highly likely result in loan applicants on those groups being rejected.

This could be viewed as an algorithmic version of *statistical discrimination*. Statistical discrimination has been an important problem for economists both theoretically and empirically ((Coate and Loury, 1993);(Arrow, 1998);(Altonji and Pierret, 2001);(Fang and Moro, 2011)). In the upcoming big data era, this problem could arise severer than ever. When decision is made from many variables, the difficulty is that removing the sensitive variable from the dataset is not a sufficient solution. This problem is long known as *disparate impact*, a notion that was born in the 1970s. The U.S. Supreme Court ruled that the hiring decision at the center of the Griggs v. Duke Power Co. case [2] was illegal because it disadvantaged an application of to a certain race, even though the decision was not explicitly determined based on the basis of race. Duke Power Co. was subsequently

[1] https://www.hirevue.com/

[2] Griggs v. Duke Power Co. 401 U.S. 424 (1971).

forced to stop using test scores and diplomas, which are highly correlated with race, in its hiring decisions.

The issue of disparate impact is particularly critical when big data is available. Machine learning algorithms utilize thousands of variables, each of which may be correlated with the sensitive variable to some extent. As a result, information of the sensitive variable can be easily recovered even if the variable itself is not included in the input of the machine. Moreover, it is extremely difficult for human to check the influence of variables on the prediction. Unlike the case of Duke Power Co., it is impossible for a human judge to determine discriminative effect of each variable one by one. As a result, we may statistically discriminate a certain group even without noticing it.

The potential economic impact of statistical discrimination can be extremely considering how rapidly algorithmic decision making is prevailing in economic situations. Given the importance of the problem, it is desirable to invent a methodology to eliminate disparate impact from algorithmic decision makings. To do so, it requires an algorithmic approach since achieving it manually is impossible.

In this paper, we propose a new fair algorithm that prevents disparate impact inspired by two-stage least square regression. Though some literature have studied disparate impact in the context of fairness-aware machine learning, there are three major limitations on the existing algorithms intended to alleviate disparate impact:[3]

- Most of the existing algorithms are built for classification tasks and cannot deal with regression tasks. While classification is very important, there are tasks that require continuous target variables, such as salaries quoted in a job offer and penalties of criminals. Unfortunately, only a few algorithms such as (Calders et al., 2013; Berk et al., 2017a) are able to handle regression.

- Existing algorithms cannot deal with numerical (continuous) sensitive variables. Although most sensitive variables, such as gender, race, and religions are binary or categorical (polyvalent), some sensitive variables are naturally dealt with in terms

[3]More detailed discussion of existing algorithms is in Appendix.

of numerical values. For example, the Age Discrimination Act [4] in the U.S. prohibits discrimination in hiring, promotion, and compensation on the basis of age for workers age 40 or above; here, age is a sensitive variable that is naturally dealt with numerically.

- Direct application of a fair algorithm could lead to reverse discrimination. To see this, let us take the example of income prediction in the *A*dult dataset[5]((Zliobaite et al., 2011)). In the Adult dataset , women on average have lower incomes than men. However, women in the dataset work fewer hours than men per week on average. A fairness-aware classifier built on the top of this dataset, which equalizes the wage prediction of women and men, leads to a reverse discrimination that makes the salary-per-hour of men smaller than that of women. Such discrimination can be avoided by introducing explanatory variables and this allows us to make a difference on the basis of the explanatory variables. In fact, as in the case of Griggs v. Duke Power Co., promoting decisions that cause disparate impacts is not allowed because they are not based on a reasonable measure of job performance, which implies (in some cases) decisions can be fair if they are of reasonable explanatory variables. Unfortunately, most of the existing studies cannot utilize explanatory variables.

Inspired by the econometrics literature, we propose a two-stage discrimination remover (2SDR) algorithm (Section 2.3). The algorithm consists of two stages. The first removes disparate impact, and the second is for prediction. The first stage can be considered to be a data transformation that makes the linear classifiers of the second stage fair.

We showed that 2SDR is a fair algorithm that (i) performs quite well in not only regression tasks but also classification tasks and (ii) is able to utilize explanatory variables to improve estimation accuracy. Moreover (iii), it reduces discrimination bias in numeric sensitive variables, which enables us to avoid other classes of discrimination, such as age discrimination (Center., 1975).

[4]The United States Civil Rights Center (1975).

[5]https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

Section 2.2 defines the problem of the prediction with potential discrimination. Section 2.3 introduces our 2SDR algorithm. Theoretical property of 2SDR is analyzed in Section 2.4. We verified the practical utility of 2SDR by using real-world datasets in Section 2.5. Section 2.6 concludes the paper.

2.2 Problem

Each vector in this paper is a column vector and is identified as a $d \times 1$ matrix where d is the dimension of the vector. Let n be the number of datapoints. The i-th datapoint is comprised of a tuple (s_i, x_i, z_i, y_i), where

- $s_i \in \mathbb{R}^{d_s}$ is the "sensitive" variables of d_s dimensions that requires special care (e.g., sex, race, and age).
- $x_i \in \mathbb{R}^{d_x}$ is the normal non-sensitive variables of d_x dimensions. The difficulty in fairness-aware machine learning is that x_i is correlated with s_i and requires to be "fairness adjusted".
- $z_i \in \mathbb{R}^{d_z}$ is the set of explanatory variables of d_z dimensions that either are not independent of s_i, or not to be adjusted for other reasons. Note that z_i can be blank (i.e., $d_z = 0$) when no explanatory variable is categorized in.
- y_i is the target variable to predict. In the case of classification, $y_i \in \{0, 1\}$, whereas in the case of regression, $y_i \in \mathbb{R}$.

Note that, unlike most existing algorithms, we allow s_i to be continuous.

Unlike economic research, the goal of machine learning is to provide a prediction. We try to find a function $\hat{y}(s, x, z)$ that calculates an estimate of y from the observed data s, x, z.

First, we estimate a function $\hat{y}(s, x, z)$ using the training data. The objective is given a data (s, x, z) out of the training data, the prediction $\hat{y}(s, x, z)$

$\hat{y}(s, x, z)$ is also supposed to comply with some fairness criteria, which we discuss in the next section.

A fairness-aware algorithm outputs $\hat{y}(s, x, z)$, which is an estimator of y that complies with some fairness criteria, which we discuss in the next section. We also use $\mathbf{Y} \in \mathbb{R}^{n\times 1}, \mathbf{X} \in \mathbb{R}^{n\times d_x}, \mathbf{Z} \in \mathbb{R}^{n\times d_z}, \mathbf{S} \in \mathbb{R}^{n\times d_s}$ to denote a sequence of n datapoints. Namely, the i-th rows of $\mathbf{S}, \mathbf{X}, \mathbf{Z}$, and $\mathbf{Y}$ are s_i , x_i , z_i , and y_i , respectively.

2.2.1 Fairness criteria

This section discusses fairness criteria that a fairness-aware algorithm is expected to comply with. We consider group-level fairness in the sense of preventing disparate impact ((Commission., 1979)), which benefits some group disproportionally. For ease of discussion, we assume $d_s = 1$ and s is a binary[6] or real single variable. Note that our method (Section 2.3) is capable of dealing with (i) multiple sensitive variables, (ii) continuous s. Let (s, x, z, y) be a sample from the target dataset to make a prediction. Let $\hat{y} = \hat{y}(s, x, z)$ be an estimate of y that an algorithm outputs. For binary s and $\hat{y}$, the P%-rule ((Commission., 1979; Zafar et al., 2017b)) is defined as

$$\min_{p} \quad \left(\frac{\mathbb{P}[\hat{y} = 1|s = 1]}{\mathbb{P}[\hat{y} = 1|s = 0]}, \frac{\mathbb{P}[\hat{y} = 1|s = 0]}{\mathbb{P}[y = 1|s = 1]} \right) \geq \frac{p}{100}. \tag{2.1}$$

The rule states that each group has a positive probability at least p% of the other group. The 100%-rule implies perfect removal of disparate impact on group-level fairness, and a large value of p is preferred.

For binary s and continuous $\hat{y}$, an natural measure that corresponds to the p%-rule is the mean distance (MD) (Calders et al., 2013), which is defined as:

$$\left| \mathbb{E}[\hat{y}|s = 1] - \mathbb{E}[\hat{y}|s = 0] \right|, \tag{2.2}$$

which is a non-negative real value, and a MD value close to zero implies no correlation between s and y. Moreover, Calders et al. (Calders et al., 2013) introduced the area under the receiver operation characteristic curve (AUC) between $\hat{y}$ and s:

$$\frac{\sum_{i\in\{1,2,\ldots,n\}:s_i=1} \quad \sum_{j\in\{1,2,\ldots,n\}:s_j=0} \mathrm{I}[\hat{y}_i > \hat{y}_j]}{n_{s=1} \times n_{s=0}}, \tag{2.3}$$

[6]Although there are several possible definitions, it is not very difficult to extend a fairness measure of binary s to one of a categorical s.

where $\mathrm{I}[x]$ is 1 if x is true and 0 otherwise, and $n_{s=1}$ (resp. $n_{s=0}$) is the number of datapoints with $s = 1$ (resp. $s = 0$), respectively. The AUC takes value in $[0, 1]$ and is equal to 0.5 if s shows no predictable effect on y.

Moreover, for continuous s, we use the correlation coefficient (CC) $|\mathrm{Cov}s\hat{y}|$ between s and $\hat{y}$ as a fairness measure. Note that, when s is binary, the correlation is essentially equivalent to MD (Eq. (2.2) up to a normalization factor.

2.3 Proposed Algorithm

Here, we start by reviewing the idea of the two-stage least squares (2SLS), a debiasing method that is widely used in statistics, econometrics, and many branches of natural science (Section 2.3.1). Inspired by 2SLS, we describe the two-stage discrimination remover (2SDR) for fairness-aware classification and regression (Section 2.3.2). Section 2.3.3 compares 2SDR with existing data preprocessing methods.

2.3.1 Two-stage least squares (2SLS)

Consider a linear regression model

$$y_i = x_i \, \beta + \ _i,$$

where the goal is to predict $y_i \in \mathbb{R}$ from variables $x_i \in \mathbb{R}^{d_x}$. If the noise $\ _i$ is uncorrelated with x_i, an ordinary least square $\hat{\beta}_{\mathrm{OLS}} = (X \ \ X)^{-1} X \ \ Y$ consistently estimates β. However, the consistent property is lost when x_i is correlated with $\ $: Namely, it is well-known ((Wooldridge, 2013)) that, under mild assumption

$$\hat{\beta}_{\mathrm{OLS}} \xrightarrow{p} \beta + \frac{\mathrm{Cov}x\epsilon}{\sigma_x^2},$$

where $\mathrm{Cov}x\epsilon$ is the covariance between x and $\ .$ σ_x^2 is the variance of x_i, and the arrow $\xrightarrow{p}$ indicates a convergence in probability. To remove the bias term, one can utilize a set of

additional variables z_i that are (i) independent of $\ _i$, and (ii) correlate with x_i. The crux of 2SLS is to project the columns of X in the column space of Z:

$$\hat{X} = Z(Z\ \ Z)^{-1}Z\ \ X$$
$$\hat{\beta}_{\text{2SLS}} = (\hat{X}\ \ \hat{X})^{-1}\hat{X}\ \ Y. \tag{2.4}$$

Unlike the OLS estimator, the 2SLS estimator consistently estimates β. That is,

$$\hat{\beta}_{\text{2SLS}} \xrightarrow{p} \beta.$$

Note that the exogenous control variables are included both in X and Z if they exist.

2.3.2 Proposed algorithm: 2SDR

The idea of our algorithm is inspired by 2SLS described above. Intuitively, in the first stage of 2SLS, the variation of $\hat{X}$ represents the variation of X that is explained by the instrumental variables Z. Meanwhile, the residual of the first stage, $X - \hat{X}$, should capture all the variation of X that is orthogonal to Z. Now if we replace Z with sensitive variables $\mathbf{S}$, the residual still contains the information of X that is useful to predict y, but the correlation with s is removed. Thus, if we use this residual instead of $\hat{X}$ in the second stage, the resulting prediction is not influenced by the correlation between S and X and therefore free from disparate impact. Note that in the first stage, one can add more control variables to $\mathbf{S}$ as long as they are not correlated with S, or disparate impact due to them are acceptable. One potential example of such a variable is high school GPA on college admission: even if difference across gender or race exist, we may not need to adjust it if it is self responsibility.

Formally, our case considers a prediction problem with a fairness constraint (Section 2.2). That is, to estimate the relationship

$$y_i = x_i\ \beta + \ _i,$$

subject to fairness criteria that urges an estimator $\hat{y}_i$ to be uncorrelated to s_i (Section 2.2.1).

Algorithm 3 2-Stage Discrimination Remover (2SDR).

1: Input: Second stage algorithm $f(x, z)$.
2: Using training data $(\mathbf{S}_{\text{train}}, \mathbf{X}_{\text{train}}, \mathbf{Z}_{\text{train}}, \mathbf{Y}_{\text{train}})$:
3: $\hat{\mathbf{B}}_s \leftarrow ([\mathbf{S}_{\text{train}}, \mathbf{Z}_{\text{train}}] \ \ [\mathbf{S}_{\text{train}}, \mathbf{Z}_{\text{train}}])^{-1}[\mathbf{S}_{\text{train}}, \mathbf{Z}_{\text{train}}] \ \ \mathbf{X}_{\text{train}}$.
4: $\mathbf{U}_{\text{train}} \leftarrow \mathbf{X}_{\text{train}} - [\mathbf{S}_{\text{train}}, \mathbf{Z}_{\text{train}}]\hat{\mathbf{B}}_s$.
5: Train the function f with $(\mathbf{U}_{\text{train}}, \mathbf{Z}_{\text{train}})$.
6: **for** each data point (s_i, x_i, z_i, y_i) in testdata **do**
7: Predict $u_i \ \leftarrow x_i \ - [s_i, z_i] \ \ \hat{\mathbf{B}}_s$.
8: Predict $\hat{y}_i \leftarrow f(u_i, z_i)$.
9: **end for**

The main challenge here is that x_i is correlated with the sensitive variable s_i, and thus, simple use of the OLS estimator yields a dependency between $\hat{y}_i$ and s_i. To resolve this issue, we use $\mathbf{U} = \mathbf{X} - [\mathbf{S}, \mathbf{Z}]([\mathbf{S}, \mathbf{Z}] \ \ [\mathbf{S}, \mathbf{Z}])^{-1}[\mathbf{S}, \mathbf{Z}] \ \ \mathbf{X}$, which is the residual of $\mathbf{X}$ regressed on $\mathbf{S}$ and $\mathbf{Z}$ and is free from the effect of $\mathbf{S}$, for predicting $\mathbf{Y}$. In the second stage, we use $\mathbf{U}$ and $\mathbf{Z}$ to learn an estimator of $\mathbf{Y}$ by using an off-the-shelf regressor or classifier. The entire picture of the 2SDR algorithm is summarized in Algorithm 3.

One big advantage of our algorithm is that one may use any algorithm in the second stage, though we mainly intend a linear classifier or regressor for the reason discussed in Section (Theorem 2.4). Following the literature of machine learning, we learn the first and the second stage with the training dataset, and use them in the testing data set.

2.3.3 Comparison with other data preprocessing methods

The first stage of 2SDR (Line 3 of Algorithm 3) learns a linear relationship between $\mathbf{S}$ and $\mathbf{X}$. This stage transforms each datapoint by making the second stage estimator free from the disparate impact, so one may view 2SDR as a preprocessing-based method that changes the data representation. This section compares 2SDR with existing methods that transform a dataset before classifying or regressing it. At a word, there are two classes of

data transformation algorithm: An algorithm of the first class utilizes the decision boundary and intensively resamples datapoints close to the boundary (Kamiran and Calders, 2010). Such an algorithm performs well in classifying datasets, but its extension to a regression task is not straightforward. An algorithm of the second class successfully learns a generic representation that can be used with any classifier or regressor afterward (Zemel et al., 2013; Feldman et al., 2015). Such an algorithm tends to lose information at the cost of generality: the method proposed by Zemel et al. (Zemel et al., 2013) maps datapoints into a finite prototypes, and the one in Feldman et al. (Feldman et al., 2015) conducts a quantile-based transformation, and loses the individual modal structures of the datapoints of $s = 0$ and $s = 1$. As a result, these methods tend to lose estimation accuracy. Moreover, its extension to a numeric s is non-trivial. The first stage in our method can be considered to be a minimum transformation for making linear regression fair and preserves the original data structure. Section 2.5 compares the empirical performance of 2SDR with those of Zemel et al. and Feldman et al. (Zemel et al., 2013; Feldman et al., 2015).

2.4 Analysis

This section analyses 2SDR. We first assume the linearity between $\mathbf{S}$ and $\mathbf{X}$ in the first stage, and derive the asymptotic independence of $\mathbf{U}$ and $\mathbf{S}$ (Theorem 2.4.1). Although such assumptions essentially follow the literature of 2SLS and are reasonable, regarding our aim of achieving fairness, a guarantee for any classes of distribution on x and s is desired: Theorem 2.4.2 guarantees the fairness with a very mild assumption when the second stage is a linear regressor.

Assumption 2.4.1 *Assume the following data generation model where datapoints are i.i.d. drawn:*

$$y_i = x_i \, \beta + \ _i \tag{2.5}$$

and

$$x_i \ = s_i \, \mathbf{B}_s + \eta_i \tag{2.6}$$

where $_i \in \mathbb{R}$ *and* $\eta_i \in \mathbb{R}^{d_x}$ *are mean-zero random variables independent of* s_i*. Moreover, the covariance matrix of* x *is finite and full-rank* [7].

The following theorem states that under Assumption 2.4.1, u is asymptotically independent of s.

Theorem 2.4.1 (Asymptotic fairness of 2SDR under linear dependency) *Let* (s_i, x_i, z_i, y_i) *be samples drawn from the same distribution as the training dataset, and* u_i *is the corresponding residual learnt from the training distribution. Under Assumption 2.4.1,* u_i *is asymptotically independent of* s_i*. Moreover, if* z_i *is independent of* s_i*,* $\hat{y}_i$ *is asymptotically independent of* s_i*.*

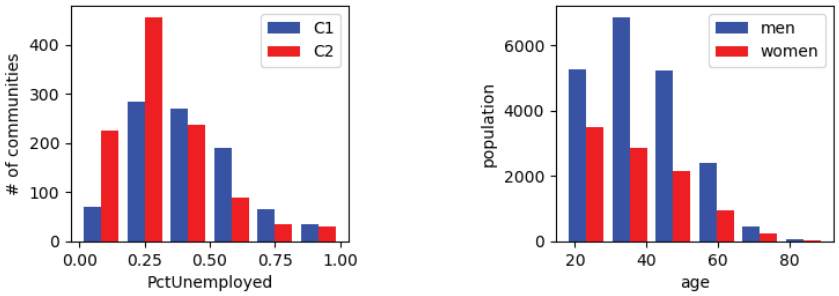
Proof Under Assumption 2.4.1, it is well known (e.g., Thm 5.1 in ((Wooldridge, 2013))) that the first-stage estimator is consistent. That is, $\hat{\mathbf{B}}_s \xrightarrow{p} \mathbf{B}_s$ as $n \to \infty$, from which we immediately obtain $u_i \xrightarrow{p} \eta_i$. By the assumption that η_i is independent of s_i, u_i is asymptotically independent of s_i. The independence of $\hat{y}_i$ and s_i follows from the fact that $\hat{y}_i$ is a function of u_i and z_i that are asymptotically independent of s_i. ■

From Theorem 2.4.1, we see that 2SDR combined with any classifier or regressor in the second-stage is fair (i.e., achieves a p%-rule for any $p < 100$ (resp. any MD > 0) in classification (resp. regression) with a sufficiently large dataset. Essentially, Theorem 2.4.1 states that if the relation between u and s is linear, the first-stage OLS estimator is able to learn the relationship between them, and as a result u is asymptotically equivalent to η, which is the fraction of u that cannot be explained by s.

Heteroskedasticity in x: As long as Assumption 2.4.1 holds, x is asymptotically independent of s. However, some of the assumptions may not hold for some variables in a dataset. In particular, Eq. (2.6) implies that x is linear to s, and thus, the distribution of x conditioned on $s = 1$ and $s = 0$ is identical after correcting the bias $\mathbb{E}[x|s = 1] - \mathbb{E}[x|s = 0]$ [8]. Figure 2.1 shows some variables where the distribution of x is very different among $s = 1$

[7] Note that this is a sufficient condition for the "no perfect collinearity" condition in Wooldridge (Wooldridge, 2013).

[8] For the ease of discussion, let s be binary value here.



(a) Distribution of PctUnemployed in the C&C dataset

(b) Distribution of age in the Adult dataset

Figure 2.1.: The difference of distribution in characteristics in sensitive characteristics.

The first histogram (Figure (a)) shows the percentage of people in the labor force and unemployed (PctUnemployed) in each community in the C&C dataset, where the horizontal axis is PctUnemployed and vertical axis is the number of corresponding communities. The communities are categorized into the ones with a large portion of black people (C1) and the others (C2). One can see that PctUnemployed in C2 is sharply centered around 0.25, whereas the value in C1 shows a broader spectrum: As a result the variance of PctUnemployed is greatly differ among the two categories. The second histogram (Figure (b)) shows the number of people of different age in the Adult dataset, where the horizontal axis is the age and the vertical axis is the number of people. One can see that not only the variances but also the form of distributions are different between women and men, as majority of the women in the dataset are of the youngest category. The details of these datasets are provided in Section 2.5.

and $s = 0$. Taking these variables into consideration, we would like to seek some properties that hold regardless of the linear assumption in the first stage. The following theorem states that 2SDR has a plausible property that makes $\hat{y}$ fair under very mild assumptions.

Theorem 2.4.2 (Asymptotic fairness of 2SDR under general distributions) *Assume that each training and testing datapoint is i.i.d. drawn from the same distribution. Assume that*

the covariance matrix of x and s are finite and full-rank. Assume that the covariance matrix between x and s is finite. Then, the covariance vector $\mathrm{Cov}su \in \mathbb{R}^{d_s \times d_x}$ converges to 0 in probability as $n \to \infty$, where 0 denotes a zero matrix.

Proof Let (s, x, z, y) be a sample from the identical distribution. The OLS estimator in the first stage is explicitly written as

$$\hat{\mathbf{B}}_s = (\mathbf{S}_{\text{train}}\mathbf{S}_{\text{train}})^{-1}\mathbf{S}_{\text{train}}\mathbf{X}_{\text{train}},$$

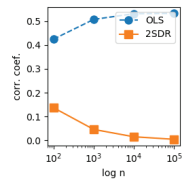
which, by the law of large numbers, converges in probability to $\mathrm{Cov}^{-1}(s, s)\mathrm{Cov}sx$, where $\mathrm{Cov}^{-1}(s, s) \in \mathbb{R}^{d_s \times d_s}$ is the inverse of the covariance matrix of s, and $\mathrm{Cov}sx \in \mathbb{R}^{d_s \times d_x}$ is the covariance matrix between s and x. Then,

$$\begin{aligned}
\mathrm{Cov}su &= \mathrm{Cov}sx - \mathrm{Cov}s\hat{\mathbf{B}}_s\, s \\
&\xrightarrow{p} \mathrm{Cov}sx - \mathrm{Cov}ss\mathrm{Cov}^{-1}(s, s)\mathrm{Cov}sx \\
&= \mathrm{Cov}sx - \mathrm{Cov}sx = 0. \qquad (2.7)
\end{aligned}$$

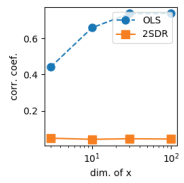
■

Asymptotic fairness of regressor: Notice that a linear regressor in the second stage outputs $\hat{y}$ as a linear combination of the elements of u and z. Theorem 2.4.2 implies that a regressor is asymptotically fair in the sense of MD (for binary s) or correlation coefficient (for continuous s). Unfortunately, it does not necessarily guarantee a fair classification under heteroskedasticity: A linear classifier divides datapoints into two classes by a linear decision boundary (i.e. $\hat{y}$ is whether a linear combination of u and z is positive or negative), and no correlation between u, z and s does not necessarily implies no correlation property between $\hat{y}$ and s. Still, later in Section 2.5 we empirically verify the fairness property of 2SDR in both classification and regression.

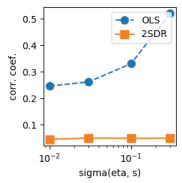
Generalization and finite-time analysis: The analysis in this section is very asymptotic and lacks a finite time bound. As OLS is a parametric model, the standard central limit theorem can be applied to obtain the asymptotic properties of the 2SDR estimator: Like the 2SLS estimator, the 2SDR estimator is expected to converge at a rate of $O(1/\sqrt{n})$.



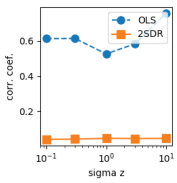
(a) CC as a func. of n



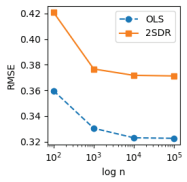
(b) CC as a func. of d_x



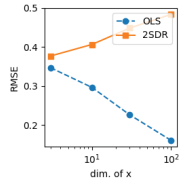
(c) CC as a func. of $\sigma_{\eta s}$



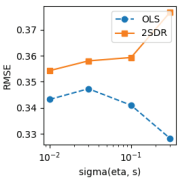
(d) CC as a func. of σ_s



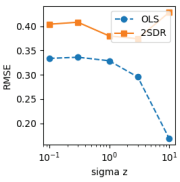
(e) RMSE as a func. of n



(f) RMSE as a func. of d_x



(g) RMSE as a func. of $\sigma_{\eta s}$



(h) RMSE as a func. of σ_s

Figure 2.2.: Performance of the algorithm with different parameters.

Correlation coefficient (CC) with different parameters (Figures (a)-(d)). Figure (a) is the result with different datasize, Figure (b) is the result with different dimension of x, Figure (c) is the result with different strength of correlation between x and s, and Figure (d) is the result with different variance of s. One can see that with sufficient large n ($n \geq 1,000$), 2SDR has consistently removes correlation between s and $\hat{y}$. Figures (e)-(h) shows the root mean square error (RMSE) with the same setting as Figures (a)-(d), where RMSE is defined as the squared empirical mean of $(\hat{y} - y)^2$. The larger d_x, $\mathrm{Cov}xs$, or σ_s is, the gap of RMSE between 2SDR and OLS is larger. This is because (i) the correlation between x and s causes a disparate impact of OLS, (ii) whereas 2SDR, which keeps $\hat{y}$ fair, forces a large bias correction when the correlation is large. For each set of parameters the result is averaged over 100 independent runs.

2.5 Experiments

In the previous section, we provided results suggesting that 2SDR achieves fairness in an asymptotic sense. To verify the actual performance of 2SDR, we conducted computer simulations. We first describe its results for a synthetic dataset (Section 2.5.1), and then describes its results for five real-world datasets (Section 2.5.2). Our simulation was implemented in Python by using the scikit-learn library[9]. Each of the simulations took from several seconds to several minutes on a modern PC.

2.5.1 Synthetic dataset

This section compares 2SDR with the standard OLS estimator on synthetically-generated datasets. Each data point (s_i, x_i, z_i, y_i) was generated from the following process, which is the standard assumption in the two-stage regression problem (Section 2.3.1):

$$y_i = x_i\ \beta_x + z_i\ \beta_z + \tag{2.8}$$

$$x_i = s_i\ \beta_s + \eta_i \tag{2.9}$$

$$z_i \sim N(0, \sigma_z) \tag{2.10}$$

$$\sim N(0, \sigma\ ) \tag{2.11}$$

$$(\eta_i, s_i) \sim N\left(0, \begin{pmatrix} \sigma_\eta & \sigma_{\eta s} \\ \sigma_{\eta s} & \sigma_s \end{pmatrix}\right).$$

Obviously, x_i and s_i are correlated, and thus, a naive algorithm that tries to learn Eq. (2.8) suffers a disparate impact, whereas 2SDR tries to untangle this dependency by learning the relationship (2.9) in the first stage. Unless specified, we set each parameters as follows: $d_x = d_z = 5$ and $d_s = 1$. $\sigma\ = 3.0$. σ_η, σ_z, and σ_s are diagonal matrices with each diagonal entry is 1.0, and $\sigma_{\eta s}$ is a matrix with each entry is 0.3. Each entry of β_x and β_z are 0.5, and each entry of β_s is 0.2. The number of datapoint n is set to $1,000$, and $2/3$ (resp. $1/3$) of the datapoints are used as training (resp. testing) datasets, respectively. Figure 2.2 shows the correlation coefficient as as measure of fairness and root mean squared error (RMSE)

[9]http://scikit-learn.org/

as a measure of prediction power for various values of parameters. 2SDR is consistently fair regardless of the strength of the correlation between x and s.

2.5.2 Real-world datasets

This section examines the performance of 2SDR in real-world datasets. The primary goal of Section 2.5.2 and 2.5.2 are to compare the results of 2SDR with existing results. We tried to reproduce the settings of existing papers (Calders et al., 2013; Feldman et al., 2015) as much as possible. Section 2.5.2 provides the results with numerical s. In the Appendix, we provide additional results for other datasets and other settings such as multiple sensitive variables case and nonlinear machines.

We conducted a set of simulations with four datasets: Namely, The Adult dataset, the Community and Crime (C&C) dataset, the Compas dataset, and the German dataset. Unordered categorical attributes are expended into dummies. Adult, Compas, and German are classification datasets (i.e., $y = \{-1, +1\}$), whereas C&C is a regression dataset. Unless explicitly described, we only put the intercept attribute (i.e., a constant 1 for all datapoints) into z. We used OLS in each attribute of the first stage, and OLS (resp. the Ridge classifier) in the regression (resp. classification) of the second stage. Note that the ridge classifier is a linear model that imposes l2-regularization to avoid very large coefficients, which performs better when the number of samples is limited. For binary s, (i) we removed the attributes of variance conditioned on $s = 0$ or $s = 1$ being zero because such a attribute gives a classifier information that is very close to s, and (ii) we conducted a variance correction after the first stage that makes the variance of U conditioned on $s = 1$ and $s = 0$ identical.

Regression results for C&C dataset

We first show the results of a regression on the Communities and Crime[10] dataset that combines socio-economic data and crime rate data on communities in the United States. The Community and Crime (C&C) dataset involves 101 attributes and $1,994$ datapoints.

Following (Calders et al., 2013), we made a binary attribute s as to the percentage of black population, which yielded 970 instances of $s = 1$ with a mean crime rate $y = 0.35$ and $1,024$ instances of $s = 0$ with a mean crime rate $y = 0.13$. Note that these figures are consistent with the ones reported in Calders et al. (Calders et al., 2013). Table 2.1 shows the results of the simulation. At a word, 2SDR removes discrimination while minimizing the increase of the root mean square error (RMSE). One can see that in the sense of RMSE, OLS and SEM-MP (Calders et al., 2013) perform the best, although these algorithms do not comply with the two fairness criteria. On the other hand, 2SDR and SEM-S (Calders et al., 2013) comply with the fairness criteria, and with 2SDR performing better in the sense of regression among the two algorithms. Furthermore, we put two attributes ("percentage of divorced females" and "percentage of immigrants in the last three years") into explanatory attributes z, whose results are shown as "2SDR with explanatory attrs" in Table 2.1. One can see that the RMSE of 2SDR with these explanatory attributes is significantly improved and very close to OLS.

Classification result with Adult and German datasets

This section shows the result of classification with the Adult and German datasets. The adult dataset is extracted from the 1994 census database, where the target binary attribute indicates whether each person's income exceeds 50,000 dollars or not. German is a dataset that classifies people into good or bad credit risks [11]. The Adult dataset involves 49 attributes and $45,222$ datapoints, whereas the German dataset involves 47 attributes and $1,000$ datapoints.

[10] http://archive.ics.uci.edu/ml/datasets/communities+and+crime

[11] https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

Table 2.1.: Regression Results.

Algorithm	MD	AUC	RMSE
OLS	0.22	0.85	0.14
2SDR	0.02	0.48	0.18
2SDR with explanatory attrs	0.12	0.69	0.15
SEM-S	0.01	0.50	0.20
SEM-MP	0.17	0.76	0.14

Note: The scores are averaged result over 10-fold cross validation ((Calders et al., 2013)). The results of SEM-S and SEM-MP are the ones reported in Calders et al. (Calders et al., 2013). "2SDR with explanatory attrs" shows the result of 2SDR with two explanatory attributes ("FemalePctDiv","PctImmigRecent"]) added to z. A smaller MD indicates better fairness, and an AUC close to 0.5 indicates a very fair regressor. Smaller RMSE indicates better regression accuracy.

Following Zemel et al. and Feldman et al. (Zemel et al., 2013; Feldman et al., 2015), we used sex (resp. age) in the Adult (resp. German) datasets. Age in the German dataset is binarized into Young and Old at the age of 25 (Calders et al., 2013). Some sparse attributes in Adult are summarized to reduce dimensionality (Zafar et al., 2017a).

Let us compare the results shown in Table 2.2 with the ones reported in previous papers. In a nutshell, 2SDR, which complies with the 80%-rule, outperforms two data preprocessing methods on the Adult dataset, and performs as well as them on the German dataset: Zemel et al. (Zemel et al., 2013) reported that their data transformation combined with a naive Bayes classifier has $\sim 80\%$ (resp. $\sim 70\%$) accuracy on the Adult (resp. German) datasets. Moreover, Feldman et al. (Feldman et al., 2015) reported that their data transformation combined with a Gaussian Naive Bayes classifier had accuracy of $79 \sim 80\%$ (resp. $70 \sim 76\%$) on the Adult (resp. German) datasets. The method by Zemel et al.

Table 2.2.: Classification results for the Adult and German dataset.

Adult dataset			**German dataset**		
Algorithm	P%-rule	Accuracy	Algorithm	P%-rule	Accuracy
OLS	0.30	0.84	OLS	0.47	0.73
2SDR	0.83	0.82	2SDR	0.81	0.73

Note: The column "Accuracy" presents the classification accuracy. Unlike OLS, which does not take fairness into consideration, 2SDR complies with the 80%-rule. In German dataset, the result is averaged over 100 random splits over the training and testing datasets, where two-thirds of the datapoints are assigned to the training dataset at each split.

(Zemel et al., 2013) coarse-grains the data by mapping them into a finite space, which we think the reason why its performance is not as good as ours. Meanwhile, the quantile-based method by Feldman et al. (Feldman et al., 2015) performed impressively well in the German dataset but not very well in the Adult dataset: In the Adult dataset, it needed to discard most of the attributes that are binary or categorical, which we consider as the reason for the results.

Numeric s

Next, we considered numeric sensitive attributes. Table 2.3 shows the accuracy and correlation coefficient in OLS and 2SDR. On the C&C dataset, 2SDR reduced correlation coefficient (CC) with a minimum deterioration to its RMSE. In other words, 2SDR was a very efficient at removing the correlation between $\hat{y}$ and s.

Table 2.3.: Results in the case s is median income (C&C) or age (Adult and German).

Algorithm (Dataset)	CC	Accuracy	RMSE
OLS (C&C)	0.50	-	0.14
2SDR (C&C)	0.04	-	0.17
OLS (Adult)	0.22	0.84	-
2SDR (Adult)	0.07	0.83	-
OLS (German)	0.11	0.76	-
2SDR (German)	0.05	0.75	-

Note: Note that age was not binarized in the result of this table.

2.6 Conclusion

We studied indirect discrimination in classification and regression tasks. In particular, we studied a two-stage method to reduce disparate impact. Our method is conceptually simple and has a wide range of potential applications. Unlike most of the existing methods, our method is general enough to deal with both classification and regression with various settings. It lies midway between a fair data preprocessing and a fair estimator: It conducts a minimum transformation so that linear algorithm in the second stage is fair. Extensive evaluations showed that our method complied the 80%-rule the tested real-world datasets.

The following are possible directions of future research:

- **Other criteria of fairness:** While the disparate impact considered in this paper is motivated by the laws in the United States, the notion of fairness is not limited to disparate impact (Berk et al., 2017b). To name a few studies, the equalized odds condition (Hardt et al., 2016) and disparate mistreatment (Zafar et al., 2017a) have been considered. Extending our method to other criteria of fairness would be interesting.

- **Non-Linear second stage:** In this study, we restricted the second-stage algorithm to be linear. The main reason for doing so is that the first stage in 2SDR is designed to remove the correlation between $\hat{y}$ and s, which is very suitable to linear algorithms (Theorem 2.4.1 and 2.4.2). We have also conducted some experiment with generalized linear model in the second stage (Section E), where we observed a inferior fairness than a linear model. Extending our work to a larger class of algorithms would boost the accuracy of 2SDR on some datasets where non-linearity is important.

- **Economic impact of fairness constraint:** Fairness adjustment may influence the incentive of the agents. In hiring decision, the Similar to affirmative action,

3. SO YOU THINK YOU ARE SAFE? IMPLICATIONS OF QUALITY UNCERTAINTY IN SECURITY SOFTWARE

3.1 Introduction

Over the past few years, the importance of information security has become increasingly apparent not only for organizations but also societies. Security software, which serves as the front line of defense against cyber threats, has been widely adopted and has become essential to users. However, despite the availability and variety that security software has to offer, its ability to protect the user is still far from perfect. For example, only about a half of 47 major antivirus software in 2014 could detect new threats on the release date while 10% of them still failed to detect threats a year after the release (Vigna, 2014). This lack of quality issue is especially important as end-users tend to have limited knowledge regarding information security (**?**, e.g.,)]katz2005effect, albrechtsen2007qualitative and thus falsely believe in the quality of the security software they adopt. This "false sense of security" among end-users that they tend to overestimate the quality of the security software they adopt has been documented in several studies (**?**, e.g.,)]guo2013humanfactor,mcafee2013smallbus. An extreme example that illustrates this behavior is the android application named "Virus shield," which was sold as a security application for $3.99 in the Google Play store and became the top-selling application with more than 30,000 downloads and 5-star ratings in only a week. However, a few days later, a security expert discovered that it actually does absolutely nothing (Andow et al., 2016).

Because of this false sense of security, an individual who adopts security protection might wrongly alter her behavior by embracing a higher level of risk (such as downloading or executing files from unknown sources more promptly), thus somewhat offsetting the level of protection she obtains by adopting the security product. This risk compensa-

tion behavior is similar to the well-known "Peltzman Effect," which was introduced to the literature in a study of the effectiveness of the seatbelt regulation. The study found that seatbelt induces drivers to drive less safely (Peltzman, 1975), which may lead to the increase in the number of non-fatal accidents (Cohen and Einav, 2003). The risk implication because of information asymmetry between the perception and the reality of the quality of the security software can be extensive, as highlighted in the prior works (**?**, e.g.,)]christin2012s,warkentin2012areyousure. Yet, most game theoretic models in the domain of information security do not account for information asymmetry or its consequence in their analysis e.g. (Arora et al., 2006; **?**). Their welfare analyses are conducted assuming that consumer expectations are accurate. What do inaccurate consumer expectations mean for social welfare? This question is the key focus of our study.

Our model incorporates several distinctive features regarding consumer behavior. First, we allow users to receive information regarding the quality of security software before making purchasing decisions. We assume that ill-informed consumers receive biased information regarding the software quality while the well-informed consumers, on the other hand, receive the information pertaining to the true quality of security software. Each consumer is unaware whether she belongs to the ill- or well-informed segment when she makes her purchasing decisions under this uncertainty. Following that, after making a purchasing decision, each consumer decides on the extent of engaging in activities that create value in their eyes yet potentially harm them. Finally, consumers realize their utility that is dependent on the risky behavior they exerted, the quality of the product consumed, and their own preference. In this manner, we account for the uncertainty in quality as well as risk compensation behavior in the consumer's utility function, an aspect which is a novel feature of our setup. In such a market, we study the implications of a monopolistic vendor offering to sell a product by choosing its price and quality.

Our study yields particularly interesting insights into welfare implications. First, although the amount of bias (i.e., the difference between the true quality of security software and that of consumer perception) may appear to have a negative impact on society, we find that social welfare could actually increase as the amount of bias increases. Second, in

some circumstances, society is better off even without the security software in the market– it is because the negative impact of over-estimation outweighs the benefit of adopting such software. Third, social welfare is not maximized even when consumers know about the proportion of well- and ill-informed consumers. We provide insights into these seemingly counterintuitive results.

In the next section, we review the literature that relates to our paper. In Section 3.3, we describe the formulation and basic elements of the quantitative model we propose in this study. Section 3.4 analyzes the existence of equilibrium and related observations in the model and subsequently describes the implication of the welfare parameters. We then generalize our model by providing several alternative specifications in Section 3.5. Finally, in Section 3.6, we discuss our findings and conclude our research with managerial implications, contributions, limitations, and future research avenues.

3.2 Literature Review

In this section, we survey the literature in four different streams related to our study. First, we present a survey of prior literature that discusses the differences between perception and reality. Second, we review the literature in the domain of an individual's risk compensation behavior. Third, we explore previous literature that studies the implication of product quality uncertainty. Lastly, we survey the literature on the economics of information security.

3.2.1 Perception versus Reality

The difference between perception and reality is one of the classical topics that has been widely discussed in the philosophy literature e.g.,(Sellars et al., 1963; **?**; **?**). The implications of such differences have been studied in various domains to explain a wide range of phenomena. Examples include the relationships between environmental measures and physical activity in medicine (Kirtland et al., 2003); the concept of disavowal in psychology (Basch, 1983); and the failures in financial report auditing (Chenok, 1994). In

the specific context of information security, the difference between perception and reality corresponds to users' inability to accurately estimate the level of protection they obtain by adopting security technologies. In fact, (Chellappa and Pavlou, 2002) use a survey to show that consumers' perceived information security is not necessarily the same as the objective assessment of potential threats and that this false perception can significantly influence consumer trust in electronic commerce transactions. Furthermore, many industry-based studies find that the gap between consumers' perception and reality tends to make users overly optimistic and creates a phenomenon called "false sense of security" among end-users (Guo, 2013), small business owners (Ragan, 2013), and non-IT executives (Dipietro, 2014). (Hui, 2010) captures this behavior in a laboratory experiment and concludes that a strong security software brand could induce users to overestimate the level of security they would attain from using the software, especially among users with low levels of knowledge about information security. Despite several empirical studies suggesting that a gap exists between consumers' perception and reality, most of the prior works in the context of information security that utilize a game-theoretic model do not model the difference explicitly. Regarding the theoretical modeling of this issue and studying the implications, we are only aware of the advance selling context where it has been done so e.g.,(Xie and Shugan, 2001; **?**). A distinctive feature of our model is that we consider the variation between the realized and expected utilities when studying the implication of information asymmetry in the information security context. Information asymmetry can manifest in terms of risk compensation behavior or a lemon market-like situation. The next two subsections survey the previous literature related to those issues.

3.2.2 Risk Compensation Behavior

This stream of research analyzes how perception and reality not being identical translates into risk compensation behavior by the consumers. The well-known and controversial[1] Peltzman effect (Peltzman, 1975) demonstrates that drivers tend to embrace greater

[1] A few other papers argue against the Peltzman effect e.g.,(Graham and Garber, 1984).

accident risk because they feel safer when wearing a seat belt. He concludes that although the regulation could reduce the risk of death from an accident, compared with an unregulated market, this reduction is offset by the fact that drivers tend to embrace greater accident risk with the presence of seat belts. Many follow-up studies have shown similar behaviors in other contexts. For example, (Rudin-Brown and Jamson, 2013) examine Munich taxi drivers with and without anti-lock braking systems (ABS) and posit that drivers who operate ABS-equipped vehicles are more likely to create traffic conflicts. In addition, (Prasad and Jena, 2014) invoke the Peltzman effect to explain why some health care interventions, which seem noble, fail to yield their intended benefits. (Vrolix, 2006) provides a comprehensive review of related literature and concludes that the magnitude of such risk compensation behavior varies depending on the context. For example, even though the number of accident may increase because of the risk compensation behavior in the classic case of seat belt, the number of fatal accident may decrease because of the seat belt.[2]

The potential issue of risk compensation behavior in the context of information security has been recently raised in the research community (Christin, 2011). However, we are only aware of one prior study in this area which conducts a laboratory experiment to show that users tend to ignore security advice and open themselves to unknown risk when incentives exist to encourage such behavior (Christin et al., 2012). (Warkentin et al., 2012) also argue in their study that consumers in information security markets are likely to exhibit risk compensation behavior. They propose several potential research methodologies for behavioral researchers to further study this topic. To the best of our knowledge, we are the first to incorporate risk compensation behavior into the model of consumer and provide insights into how risk compensation behavior affects both individual and social welfare.

[2]Note that the Peltzman effect in the prior literature accounts jointly for both the direct effect (i.e., engaging in risky behavior) as well as the indirect effect (i.e., learning that occurs from engaging in risky behavior). For instance, (Pope and Tollison, 2010) study the behavior of NASCAR drivers. Here, an accident caused by one driver could also affect other drivers. Hence, their behavior may reflect the indirect effect as drivers may drive more cautiously when other drivers are reckless. Even in this context, the Peltzman effect is recognized.

3.2.3 Implication of Quality Uncertainty

While the previous subsection reviews the literature that focuses on the effect of perception versus reality on the demand side, this subsection surveys papers that examine the supply side implications of information asymmetry. Note that one main reason for the difference between perception and reality is that consumers face uncertainty about quality. Regarding market implications in the face of quality uncertainty, one of the seminal papers is (Akerlof, 1970), which investigates the second-hand automobile market. He concludes that such uncertainty can push good quality products out of the market and collapse the market as a result. Other areas of management have also observed this phenomenon, including finance e.g.,(Beatty and Ritter, 1986); accounting e.g.,(DeAngelo, 1981); operation management e.g.,(Lim, 2000); and information systems (Dimoka et al., 2012). A number of papers have followed up on potential avenues to overcome the problems in Akerlof's lemon market. One such idea is to build a reputation system e.g.,(Resnick et al., 2000), and papers have demonstrated support for such a system e.g.,(Gefen et al., 2003; Ba and Pavlou, 2002). Others have also considered governmental interventions but have concluded that reducing uncertainty by imposing government regulation alone might not be effective e.g.,(Hoffer and Pratt, 1987). As we demonstrate later, quality uncertainty does indeed play an important role in moderating the welfare implications in the information security market.

3.2.4 Economics of Information Security

In this subsection, we review analytical research in the information security context that is closely related to our study. Particularly, we survey two substreams of prior literature that share key modeling details with our work, including papers that study the market for security software and a firm's decisions on product quality, and the implications of policymaker and government intervention. However, this paper is significantly different from the prior works surveyed in this subsection as we allow users to be uncertain about

product quality. We are not aware of any previous work in this area that has incorporated this aspect of information uncertainty into an analytical model.

The market for information security software has been modeled and analyzed in the literature. Its unique characteristics (e.g., the market is highly competitive yet the coverage is low) have been established by (Dey et al., 2012), who model the market for security software in the presence of hackers, different types of attacks, and network effects; and argue that these elements contribute to the uniqueness of the market. In addition to the characteristics of the market, the welfare implications of the market entities has also been studied in a scenario where software patch availability is restricted (Kannan et al., 2016). In such a case, the vendor can strategically choose the price and maintenance decisions to take advantage of the presence of the hacker in the market. Furthermore, a firm's decision on product quality has also been studied in many contexts. For example, (August and Tunca, 2006) incorporate network externalities to analyze different patching policies to manage network security. They show that patching policy is not a one-size-fits-all approach in the sense that the optimal policy differs based on context (e.g., proprietary software vs. freeware, patching cost, and security risk) and that using the right user incentive can significantly improve software generated value and firm profits. In addition, the trade-offs between tolerating illegal software usage and enjoying positive network effect from higher number of users has been analytically analyzed (Lahiri, 2012). The results demonstrate that the conventional wisdom, which suggests that companies could benefit from the illegal distribution of their software product due to positive network externalities, might not be true when patching is also considered. Moreover, (Arora et al., 2006) build an economics model based on a firm's trade-off decision between selling error-prone software early and the cost of fixing it later. They show that the firm has incentives to release software with more bugs early when the market is sufficiently large, in contrast with the case of manufacturers of physical goods. Our model is constructed based on key modeling details proposed by the literature in this substream. Meanwhile, we incorporate consumers' risk compensation behavior into the model and allow the firm to make decisions based on the presence of quality uncertainty among consumers.

The second substream of literature examines the implications for policymaker and government intervention in the context of information security. On the one hand, the intervention has been shown to benefit society. For example, (Kannan and Telang, 2005) conclude that the case where companies sell a subscription for software vulnerability disclosures almost always underperforms the approach where such disclosures are provided for free by government-subsidized entities. On the other hand, such an intervention has also been shown in other contexts to be suboptimal. For instance, (Png and Wang, 2009) find that enforcement by the government against attackers is less effective compared with educating end-users, especially when attacks are targeted. Furthermore, different interventions can lead to different outcomes. For instance, (Chen and Png, 2003) study several cases of government policy on copyright enforcement and find that the case where a government subsidizes a legitimate purchase leads to higher social welfare compared with the case of a fine for piracy or a tax on copying medium. Given the conflicting findings from the previous literature, our work provides insights into the welfare implications for policymaker intervention in the presence of quality uncertainty about security software.

In summary, the literature in subsection 3.2.1 has demonstrated that a large portion of security software users tend to be overly optimistic about the level of security their software offers. In addition, previous works in subsection 3.2.2 and 3.2.3 have shown that risk compensation behavior, particularly drawing upon the Peltzman effect, implies that this uncertainty could adversely affect the firm's decision and social welfare. However, much of the literature in the economics of information security has not considered modeling this aspect. Our paper utilizes a game-theoretic model to provide insights into the resulting welfare implications. We are among the first to provide a formal analysis of the implications of consumer uncertainty in the context of the security software market.

3.3 Model

We begin by highlighting two key features of our model. The first feature is that we capture the consumers' uncertainty regarding the nature of the security software by

assuming that they cannot directly observe the software's true quality. Instead, they observe a signal that contains imperfect information of the quality. Furthermore, they are aware that the signal may be noisy. The signals lead to beliefs about quality and play an important role in determining whether the consumers purchase the software and also how they generate value from the software. The second key feature is that, after consumers decide whether or not to purchase the security product, they engage in value-adding but risky activities based on their perception of the software quality. The main intention of this feature is not only to capture the Peltzman-like effect in the information security context but also to analyze the effect of misperception on the welfare parameters.

We study such behaviors of consumers in a monopolistic market with the vendor choosing the security quality q and price p. Our analysis of the monopolistic market[3] can be justified as follows. Information security can be considered as an information good. As (Jones and Mendelson, 2011) note, the markets for information goods tend to result in a monopolistic market. Further, ours is one of the first papers in the information security context to analyze welfare implications caused by consumer uncertainty. A monopolistic model is useful in providing insights about the trade-offs that are germane to this setup. Next, we describe the two primary players in our models, the vendor and the consumers.

3.3.1 Vendor

The way we model the vendor is fairly standard and is described first. In this market, the monopolistic vendor realizes a demand $D(p, q)$, which varies with the price p and quality q chosen by the vendor. We ignore the fixed cost of producing the security software and consider the marginal cost of producing additional copies to be zero. However, the software quality is a consequence of the maintenance effort, such as virus signatures that need to be identified in antivirus software. For this purpose, we assume the cost function $c(q)$ is strictly convex, strictly increasing in quality, and zero when the quality is zero. Also, we

[3] Although we recognize that in reality, the market for security software typically involves a number of vendors (e.g., antivirus, backup software, host-based intrusion prevention systems), a market also exists with a limited number of vendors (e.g., data-loss prevention, security compliance).

assume $c\,(0) = 0$ and $\lim_{q\to\infty} c\,(q) = \infty$. The vendor's choice is to maximize the profit $\pi = p\ D(p,q) - c(q)$. Therefore, the vendor chooses the optimal price and quality to maximize its profit[4]:

$$(p^*, q^*) = \arg\max_{p,q} \pi.$$

3.3.2 Consumers

Our modeling of consumers is distinctive and is described below. For ease of understanding, we separate the explanation of our modeling into several steps as follows:

Heterogeneous consumer utility function: The consumer utility function involves three parameters x, q, and β in the form $u(x; q, \beta)$. The parameter $\beta \in (0, 1)$ captures consumer heterogeneity regarding valuations that consumers obtain from engaging in digital activities. It is assumed to follow a distribution whose pdf is $f(\beta)$. Each consumer is aware of her own β. The parameter x is the amount of value-adding but risky behavior that the consumer engages in. For example, the recent articles about White House aides using the not-so-secure Confide app to engage in potentially private conversations believing in the "military-grade security" encryption, even though it may not be so (Newman, 2017) illustrates the risky behavior some consumers might engage in. Eventually, we will endogenize x. The term q is the quality of the security software employed. If the consumer does not purchase the software, we assume that she generates the utility $u(x; 0, \beta)$.

Structure of the quality signal: Let us assume that consumers, independent of their β, receive a signal $\tilde{q}$ that contains imperfect information of the quality.[5] For simplicity, we assume that there are only two types of signals: accurate signals and biased signals. For instance, a consumer may receive an accurate evaluation if she consumes information from credible technical reports about security software quality but may receive a biased signal from paid reviews or sponsored reports that exaggerate the quality of the security

[4]In our single period model, we do not explicitly allow the reputation to affect the firm's choice of p and q. However, by incorporating a separate reputation cost that is a function of quality, we can demonstrate that the results remain qualitatively similar.

[5]Assuming uncertainty on product quality is common in the literature in the economics of information. For a comprehensive literature review, see (Stiglitz, 2000).

software. Let consumers who receive the accurate signal $\tilde{q} = q$ be referred to as well-informed consumers, and those who receive the biased signal $\tilde{q} = g(q; s)$ be referred to as ill-informed consumers. Here, s represents the amount of bias. If $Pr(\tilde{q}|q)$ represents the distribution of the signal $\tilde{q}$ conditional on the true quality q, the consumer infers $Pr(q|\tilde{q})$ when making the purchasing decisions. For our analysis, we assume:

$$Pr(\tilde{q}|q) = \begin{cases} \sigma & \text{if } \tilde{q} = g(q; s) \\ 1 - \sigma & \text{if } \tilde{q} = q \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

where σ represents the probability of observing a biased signal.

Assumptions regarding the biased signal ($g(q; s)$): First of all, we assume that s is a common knowledge. Second, we assume $g(q; s) > q$, which conveys the meaning that we only focus on overestimating consumers. Note that the ill-informed consumers' misestimation is a function of an exogenous parameter s which captures the amount of bias. Therefore, a large s means a large difference between perception and reality, i.e., $\frac{\partial}{\partial s}(g(q; s) - q) > 0$ (obviously, it implies that $\frac{\partial g(q;s)}{\partial s} > 0$). Moreover, $g(q; 0) = q$, i.e., if there is no bias, the perceived quality is identical to the true quality of the security software. Finally, we assume $\frac{\partial g(q;s)}{\partial q} > 0$, meaning that the biased signal takes larger value with the larger actual quality.

Redefining some variables: Later in Section 3.5 of the paper, we allow for the generalization of more than two types of signals. Therefore, to facilitate those discussions, we redefine some variables. The total mass of consumers is normalized to one. Each consumer belongs to a segment $t \in \{1, 2\}$, where $t = 1$ represents ill-informed consumers who receive a biased signal, and $t = 2$ represents well-informed consumers who receive an accurate signal. Let $\tilde{q}_t$ be the signal observed by group t, i.e., $\tilde{q}_1 = g(q; s)$ and $\tilde{q}_2 = q$. Recall that the probability a consumer observes a biased signal is σ. We denote the proportion of consumers of a group t as σ_t, where $\sigma_1 = \sigma$ and $\sigma_2 = 1 - \sigma$. These are assumed to be known to the vendor. For ease of readability, we interchangeably use σ with σ_1 and $(1 - \sigma)$ with σ_2.

Getting to $Pr(q|\tilde{q})$: Generally speaking, as mentioned earlier, consumers would infer $Pr(q|\tilde{q})$ for decision-making. To capture the reality of the uncertainty of the market, we assume that consumers are not only unaware of true quality but also of how likely their signal is to be biased. Similar issue has been recently highlighted in the context of news. For example, anecdotal evidence has suggested that consumers are unable to properly evaluate the reliability of the media source e.g.,(Swartz and della Cava, 2016; Silverman and Singer-Vine, 2016). Relatedly, a formal research article has also reached the same conclusion (Wineburg and McGrew, 2016). Therefore, we allow consumers to believe in a data generating process, $\hat{Pr}_t(\tilde{q}|q)$, which may differ across segments and also differs from $Pr(\tilde{q}|q)$ as defined earlier in Equation 3.1. Using Bayesian updates, $\hat{Pr}_t(q|\tilde{q}) = \frac{\hat{Pr}_t(\tilde{q}|q)\hat{Pr}_t(q)}{\hat{Pr}_t(\tilde{q})}$. Assuming non-informative prior, we define:

$$\hat{Pr}_t(q|\tilde{q}) = \begin{cases} r_t & \text{if } q = g^{-1}(\tilde{q}; s) \\ 1 - r_t & \text{if } q = \tilde{q} \\ 0 & \text{otherwise.} \end{cases}$$

Note that r_t is the subjective probability of the signal being biased. Also with this construct, r_t can be interpreted as the chance that a consumer in segment t believes that a signal she receives is biased, which may or may not be the true probability σ. Intuitively, from the perspective of consumers in segment t who observe signal $\tilde{q}_t$, there are two possibilities: the signal they observe may be the true quality, or it is biased. In the former case, the true quality is simply $q = \tilde{q}_t$. In the latter case, $g(q; s) = \tilde{q}_t$, thus the true quality is $q = g^{-1}(\tilde{q}_t; s)$. In the context involving two types, true quality for an ill-informed consumer is $\tilde{q}_1 = g(q; s)$ as the signal indicates, or it could be $g^{-1}(\tilde{q}_1; s) = g^{-1}(g(q; s); s) = q$ if the signal is biased. Similarly, for a well-informed consumer, the true quality can be either $\tilde{q}_2 = q$ or $g^{-1}(\tilde{q}_2; s) = g^{-1}(q; s)$. To summarize the different quality levels, Figure 3.1 illustrates the notations of product qualities in our study. The arrows marked horizontally demonstrate the possible quality levels that the consumer suspects she is in. Note that because $g(q; s)$ is strictly monotone in s, $\frac{\partial g^{-1}(q;s)}{\partial s} < 0$. Note that we construct the main model such that the amount of bias (s) has only one value for all consumers. When we

generalize our model in Section 3.5.1, we allow each segment of consumers to have its own s.

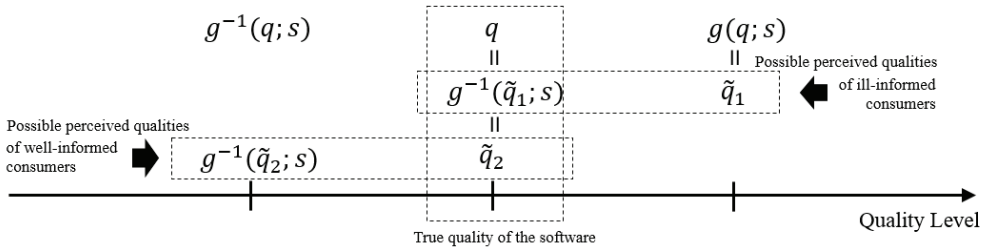


Figure 3.1.: Notations of perceived quality by the two consumer segments

Expected utility based on belief on quality: Given the belief on quality derived in the previous section, a consumer's perceived expected utility is simply

$$\tilde{U}_t(x;\tilde{q}_t,\beta) = (1-r_t)u(x;\tilde{q}_t,\beta) + r_t u(x;g^{-1}(\tilde{q}_t),\beta). \tag{3.2}$$

Optimally perceived risky behavior + Optimally perceived expected utility: Based on the perceived expected utility from Equation 3.2, we compute the optimally perceived risky behavior as $x_t^* = \arg\max_x \tilde{U}_t(x;\tilde{q}_t,\beta)$ and the corresponding expected utility is simply defined as $\tilde{U}_t^*(\beta,q)$. A consumer purchases the product if and only if $\tilde{U}_t^*(\beta,q) - p > U_{no}(\beta) \equiv \max_x u(x;0,\beta)$, where U_{no} is the utility if the consumer does not purchase the product.

Properties of the realized utility function: Because the perceived utility might be different from the realized utility, we next define the properties of the *realized* utility function for a consumer, which we earlier denoted as $u(x;q,\beta)$.

1. We assume that $u(x;q,\beta)$ is strictly concave and continuous in x (i.e., the payoff is concave with respect to the risky behavior) and continuous and strictly increasing in q and β (i.e., the payoff increases as the product quality increases or a consumer possesses higher β).

2. For any β, x, and q, we assume that $\frac{\partial^2 u}{\partial\beta\partial x} > 0$, i.e., a consumer with higher β generates a larger marginal utility from the extra level of risky behavior.

3. Let $\frac{\partial^2 u}{\partial q \partial x} > 0$ for any x and β. This characterization is consistent with the perception that a consumer can enjoy more risky behavior with less security concern when the quality of security software is higher.

Using these results, we can establish that:

Lemma 3.3.1 *For a given β, the perceived expected utility is higher for the ill-informed consumer than for the well-informed one. Hence, the market share from the well-informed consumers is smaller than or equal to that from ill-informed ones. That is, $\tilde{U}_1 > \tilde{U}_2$ and $\beta_2^* \geq \beta_1^*$.*

Model Specialization

For the sake of tractability in our analysis, we impose additional assumptions on the consumer's perceived expected utility. First, we assume that the probability density function $f(\beta)$ follows the uniform distribution. Second, we assume that the perceived expected utility at the chosen amount of risky behavior x is functionally separable as:

$$\tilde{U}_t^*(\beta, q) \;=\; \tilde{U}_t(x_t^*, \tilde{q}_t, \beta) \;=\; \beta m(q) n_t(s, r_t), \tag{3.3}$$

where $m(q)$ is strictly concave in q, and $n_t(s, r_t)$ is a function of s and r_t that is different depending on the consumer segment t. Finally, we assume that $m(q = 0) = 0$.[6] Here, the $m(q)$ term captures how the quality of the security product affects consumers' perceived expected utility. This effect is similar to the Akerlof's lemon market effect (Akerlof, 1970), which we will explain in detail at the end of section 3.4.1. In the meantime, the term $n_t(s, r_t)$ captures the "Peltzman-like" effect for consumers in segment t. Specifically, $n_t(s, r_t)$ endogenizes, within the perceived expected utility expression, the consumers' risky behavior as a function of the bias that the consumers have regarding the software quality and the likelihood that a biased signal is received. The separability assumption delivers two important dimensions. First, it improves the tractability of our model. Second, more importantly, it helps us to separately study the impact of the two effects on consumers'

[6]Other assumptions such as $m(q) \geq 0$ if $q \geq \hat{q}$ and vice-versa will generate similar results.

perceived expected utility. Additionally, our main insights remain intact even though the separability assumption is relaxed, which we will demonstrate in Section 3.5.3. Based on the assumptions mentioned here and the assumptions discussed earlier, we have:

Lemma 3.3.2 *$m(q)$ is a continuous, differentiable, and strictly increasing function of q. $n_t(s, r_t)$ is a continuous and differentiable function of s and r_t. $\frac{\partial n_1(s,r_1)}{\partial s} > 0$; $\frac{\partial n_2(s,r_2)}{\partial s} < 0$; and $\frac{\partial n_t(s,r_t)}{\partial r_t} \leq 0$ for $t = \{1, 2\}$ but the inequality is strict only if $s > 0$. The utility from not purchasing the security software, $U_{no}(\beta) = 0$ for any β.*

The lemma shows that $m(q)$ is a well-behaved function of q. In addition, the change in $n_t(s, r_t)$ with respect to s and r_t allows us to observe that the perceived expected utility of ill- (well-) informed consumers increases (decreases) with respect to s, and that the perceived expected utility decreases with respect to r_t. Before we progress forward, Table 3.1 summarizes the variables we have defined so far.

3.4 Equilibrium Results

We solve this two-stage game by computing the Subgame Perfect Nash Equilibrium using backward induction. Recall that the first stage is the vendor's decision on quality and price, and that the second stage is the consumers' purchasing and usage behaviors.

3.4.1 Consumers' Actions: Second Stage

Consumers choose whether to buy security software or not based on the price and perceived quality. Fix consumer segment $t = \{1, 2\}$. A consumer purchases the product if $\beta m(q) n_t(s, r_t) - p > 0$ but not otherwise. If β_t^* represents the indifferent consumer in segment t, only consumers with $\beta \in (\beta_t^*, 1)$ purchase the product (because the consumer utility is monotonically increasing with β). Based on the expected utility function, $\beta_t^* = \min\{\frac{p}{m(q)n_t(s,r_t)}, 1\}$. By Lemma 3.3.1, $\beta_2^* > \beta_1^*$.

If the consumer's heterogeneity parameter β is kept constant, the perceived expected utility when she is ill-informed will be higher than that when she is well-informed. As a

Table 3.1.: List of variables in the model

$u(x; q, \beta)$	realized consumer utility
x	amount of risky behavior
q	quality of the security software
β	consumer heterogeneity
t	consumer group. 1 for ill-informed and 2 for well-informed.
$\tilde{q}_t$	quality signal for consumer group t
$g(q; s)$	quality signal for ill-informed consumer ($\tilde{q}_1$)
s	amount of bias in quality signal
σ_t	proportion of the market with consumer group t
r_t	subjective probability of the signal being biased for group t
$\tilde{U}_t^*$	maximized perceived expected utility of consumer group t
$m(q)$	quality term in U_t^*
$n_t(s, r_t)$	information uncertainty term in U_t^*
p	price of the security software
π	profit of the vendor

result, the cutoff β_t^* to purchase the product is lower for the ill-informed. Following that, the total demand function can be defined as:

$$D(p, q) = \sum_{t=1}^{2} \sigma_t \left(1 - \beta_t^*\right),$$

which the vendor uses to maximize profit by choosing p and q. Two scenarios are possible with respect to demand. One scenario is that only the ill-informed consumers ($t = 1$) will

purchase the product; this happens when $\beta_2^* = 1$. In the other scenario, both consumer segments purchase the product. In order to characterize the equilibrium, we define

$$W_t = \frac{\sum_{t=1}^{t} \frac{\sigma_t}{n_t(s,r_t)}}{\sum_{t=1}^{t} \sigma_t}, \text{ and} \tag{3.4}$$

$$W(s, r_1, r_2, \sigma) = W_t | t \in \arg\max_{\tau} \sum_{t=1}^{\tau} \sigma_t \quad W_\tau^{-1} \quad . \tag{3.5}$$

We refer to the term $W(s, r_1, r_2, \sigma)$ as the *aggregated distrust factor* and will interpret it in the following Lemma. Given the definition, we characterize the equilibrium of the game as follows.

Proposition 3.4.1 *The optimal vendor profit and optimal quality are non-zero, finite, unique, and continuous in all the parameters s, r_t, and σ_t. The equilibrium price is:*

$$p^* \in \frac{m(q^*)}{2W(s, r_1, r_2, \sigma)},$$

and it is unique iff $W(s, r_1, r_2, \sigma)$ is a singleton. If $W(s, r_1, r_2, \sigma) = W_1$, only the ill-informed consumers are served; if $W(s, r_1, r_2, \sigma) = W_2$, both well- and ill-informed consumers are served. The implicit function that finds the unique optimal quality is

$$\frac{m(q^*)}{4} \sum_{t=1}^{t} \sigma_t \quad W_t^{-1} - c(q^*) = 0 \text{ if } W_t \in W(s, r_1, r_2, \sigma).$$

This proposition highlights the benefit of defining the aggregated distrust factor. It shows that in our framework, the equilibrium behavior of the vendor can be simply expressed as a function of the aggregate distrust factor. Importantly, any parameters that define the information structure of the quality only influence the vendor through this factor.

Next, before we understand how the equilibrium changes with various exogenous parameters, we focus on explaining the aggregate distrust factor term, $W(s, r_1, r_2, \sigma)$. The explanation is clearer if one understands how the exogenous parameters affect $W(s, r_1, r_2, \sigma)$.

Lemma 3.4.1 *$W_t \in W(s, r_1, r_2, \sigma)$ is non decreasing in r_t and non increasing in σ. With respect to s, the function decreases when $W(s, r_1, r_2, \sigma) = W_1$ but may increase or decrease when $W(s, r_1, r_2, \sigma) = W_2$.*

Note specifically how the variation is with respect to the bias. First, consider the case where only the ill-informed consumers serve the market. As the bias increases, $n_1(s, r_1)$ increases, and so the trust that ill-informed consumers place in the software increases. Hence, a larger market share of ill-informed consumers purchases the product. Since this term $n_1(s, r_1)$ enters the W_1 inversely as follows $W_1 = \frac{1}{n_1(s,r_1)}$, we claim that the term $W(s, r_1, r_2, \sigma)$ accounts for the distrust. Next, after observing that $W_2 = \frac{\sigma_1}{n_1(s,r_1)} + \frac{\sigma_2}{n_2(s,r_2)}$ when both consumer segments purchase the product, we extend our interpretation to this case also. Here, the notion of trust is weighted in proportion to consumer segments, thus the term aggregate distrust factor. Because the factor $n_2(s, r_2)$ makes well-informed consumers more cautious, the distrust factor only increases when both consumer segments are present. Relatedly, if ill-informed consumers dominate the market (i.e., $\sigma \to 1$), the term $W(s, r_1, r_2, \sigma)$ decreases with respect to s. On the opposite end, if the well-informed consumers dominate the market (i.e., $\sigma \to 0$), the term $W(s, r_1, r_2, \sigma)$ increases.

Explaining the effects of r_t and σ on $W(s, r_1, r_2, \sigma)$ is straightforward. A higher r_t implies that consumers are more suspicious of the observed quality. Therefore, the distrust factor $W(s, r_1, r_2, \sigma)$ naturally increases as a consequence. When σ increases, a larger proportion of consumers perceive overestimated quality. With debiasing, it implies that consumers as a whole will be more trustful of the product. Hence, the value of the distrust factor decreases. Now, we use the results of these sensitivity analyses to consider the effect of the exogenous parameters on the equilibrium.

Theorem 3.4.1 *When the parameters change, the following hold at the equilibrium.*

1. *If r_t increases (equivalently, decreases): the quality and the profit decrease (increase); if the price is a singleton, it also decreases (increases).*

2. *If σ increases (decreases): the quality and the profit increase (decrease); if the price is a singleton, it also increases (decrease).*

3. *If s increases: the profit, quality, and price* may *decrease.*

The implication of the change in the amount of bias, s, on the vendor's profit is specifically insightful. The conventional wisdom usually suggests that the vendor always benefits

from the presence of information asymmetry. Some even suspect that the vendor promotes uncertainty for its own benefit. In contrast with popular belief, we do not find it to be necessarily true. Under certain circumstances, such as the case where the population of the ill-informed consumers is sufficiently large, the vendor indeed enjoys higher profit as the bias increases. However, under other circumstances, such as the case where ill-informed consumers are not sufficiently sensitive to the bias, the vendor could actually be worse off when the bias increases. We next explain how price, quality, and profits change with bias.

Note that the effect of s on price, quality, and profits can be either positive or negative. If the vendor serves only ill-informed consumers (i.e., $W(s, r_1, r_2, \sigma) = W_1$), the effect of increasing s is equivalent to the effect of increasing σ since it only increases the overall consumers' perception on software quality. Therefore, at the margin, not only the vendor's choice of quality and price, but the vendor's profits also increase. This corresponds to the conventional wisdom mentioned in the previous paragraph. However, when both segments of consumers are in the market, an increase in s may no longer be straightforward as in the previous scenario. Consider specifically when the portion of ill-informed consumers is sufficiently low. Then, an increase in s negatively affects consumers' average perception instead (because of the debiasing). Therefore, at the margin, the value for the vendor from increasing the quality decreases – leading to a decrease in equilibrium price, quality, and profits.

Specifically with respect to the variation of quality, we wish to highlight the Akerlof's lemon market-like effect ((Akerlof, 1970)) that can occur. When a large number of well-informed consumers exist in the market, they lose trust in the software quality because of bias. This tends to decrease the $W(s, r_1, r_2, \sigma)$ term, and so the vendor's incentive to provide high quality decreases. It is the equivalent of the vendor offering only "lemons." The implication of the "Akerlof-like effect" is clearer with regard to the welfare implications.

The other effects identified in the theorem are fairly straightforward. Note that an increase in consumer suspicion (r_t) negatively affects the equilibrium price and quality. At the margin, the value from increasing quality decreases. This also means that the consumer surplus that can be extracted as profit decreases. On the other hand, an increase in the

proportion of ill-informed consumers (σ) has the opposite effect. When σ increases, more consumers observe positively biased software quality. Therefore, the vendor's marginal value from increasing the quality is higher. As a result, the software quality is higher and the price is also higher. For the same reason, the profit is also higher.

3.4.2 Welfare Implications

In this subsection, we analyze the implication of information uncertainty in consumer surplus and social welfare. For consumer segment t, the consumer surplus is defined as:

$$CS = \sum_{t=1,2} \sigma_t \left(\int_0^{\beta_t^*(p,q)} U_{no}(\beta) + \int_{\beta_t^*(p,q)}^{1} (u(x_t^*; q, \beta) - p)\, d\beta \right) .$$

Social welfare is the sum of the vendor profit, $\pi(p^*, q^*)$, and the consumer surplus, CS:

$$SW = CS + \pi(p^*, q^*).$$

Next, we examine the welfare implications at the individual consumer level before aggregating. We begin by considering the variation with respect to the amount of bias s. We can separate the variation into three terms, where each corresponds to a different effect. Specifically, the variation of consumer utility with respect to s can be expressed as follows:

$$\frac{\partial u(x_t^*; q, \beta)}{\partial s} = \underbrace{\frac{\partial u(x_t^*; q, \beta)}{\partial x}}_{\text{Peltzman-like effect}} \frac{\partial x_t^*}{\partial s} + \underbrace{\frac{\partial u(x_t^*; q, \beta)}{\partial q}}_{\text{Akerlof-like effect}} \frac{\partial q^*}{\partial s} + \underbrace{\frac{\partial u(x_t^*; q, \beta)}{\partial x} \frac{\partial q^*}{\partial s}}_{\text{Interaction effect}} \frac{\partial x_t^*}{\partial q} \quad (3.6)$$

In the first term, the Peltzman-like effect captures the consumer engaging in a suboptimal level of risky behavior because of quality misperception. For the well-informed (equiv. ill-informed), that effect is positive (resp. negative). As regards the second part of the first term, $\frac{\partial x_t^*}{\partial s}$, notice that increasing s decreases (resp. increases) x. Hence, the first term is always negative, independent of the consumer segment.

The second term accounts for the Akerlof-like effect. The part, $\frac{\partial q^*}{\partial s}$, corresponds to change in software quality because of bias. It can be both positive and negative, depending on whether the aggregated distrust factor W increases or decreases with respect to s. The

first part, however, is always positive. As a result, the second term is overall positive if and only if $\frac{\partial q^*}{\partial s}$ is positive.

As evident, the last term captures the interaction of the two effects and includes three parts. The last part $\frac{\partial x_t^*}{\partial q}$ is always positive. The other two parts correspond to the previous two paragraphs. The overall term captures how the consumers' choice of risky behavior is affected by the change in security software quality caused by the Akerlof-like effect and ends up changing the Peltzman-like effect as a result. For instance, if the quality increases as s increases (i.e., $\frac{\partial q^*}{\partial s} > 0$), the perceived quality is further higher, prompting the consumers to choose a larger x_t^*. Consequentially, the realized utility of ill-informed consumers decreases because their choice of x_t^* becomes further away from the true optimal point, which they would have chosen had they know the true quality of the security software. For the well-informed ones, the same logic leads to increased realized utility. Note that a similar set of insights can also be obtained for other exogenous parameters, except for some slight changes. For example, the first term is zero when considering the variation with respect to the proportion of ill-informed consumer σ. These effects when combined across consumers provide insights about consumer surplus.

Next, we investigate the change in social welfare, and also the vendor profit, with respect to s. In that regard, recall that there is a case where the equilibrium price is a doubleton. For the sake of simplicity, we present here the change in the social welfare with

respect to the change of s when the price is a singleton. Note that qualitatively similar results can also be established when the price is not a singleton.

$$\begin{aligned}
\frac{\partial SW}{\partial s} &= \sum_{t=1,2} \sigma_t \int_{\beta_t^*}^{1} \underbrace{\frac{\partial u(x_t^*; q^*, \beta)}{\partial x_t^*}}_{\text{Peltzman-like effect}} \frac{\partial x_t^*}{\partial s} d\beta \\
&+ \underbrace{\frac{\partial q^*}{\partial s}}_{\text{Akerlof-like effect}} \sum_{t=1,2} \sigma_t \int_{\beta_t^*}^{1} \frac{\partial u(x_t^*; q^*, \beta)}{\partial q^*} d\beta - c\left(q^*\right) \\
&+ \sum_{t=1,2} \sigma_t \int_{\beta_t^*}^{1} \underbrace{\frac{\partial q^*}{\partial s} \frac{\partial u(x_t^*; q^*, \beta)}{\partial x_t^*}}_{\text{Second order effect}} \frac{\partial x_t^*}{\partial q^*} d\beta \\
&- \underbrace{\sum_{t=1,2} \sigma_t \left(\frac{\partial \beta_t^*}{\partial s} + \frac{\partial \beta_t^*}{\partial q^*} \frac{\partial q^*}{\partial s} + \frac{\partial \beta_t^*}{\partial p^*} \frac{\partial p^*}{\partial s} \right) u(x_t^*; q^*, \beta_t^*)}_{\text{Demand change}}
\end{aligned} \tag{3.7}$$

Compare with individual utility, two additional components enter the equation in studying the variation of the social welfare with respect to s. One is the marginal cost term interacting with the Akerlof-like effect. That is because when the quality changes, the cost incurred by the vendor also changes. The second change is the shift in demand. When s changes, both due to the Peltzman-like effect and Akerlof-like effect, the purchase decisions of the consumers change. With s increasing, ill-informed consumers tend to purchase more, while well-informed consumers purchase less. When quality increases, both consumer segments tend to purchase the product more. The demand change aggregates these effects.

Theorem 3.4.2 *Social welfare can increase with the amount of bias.*

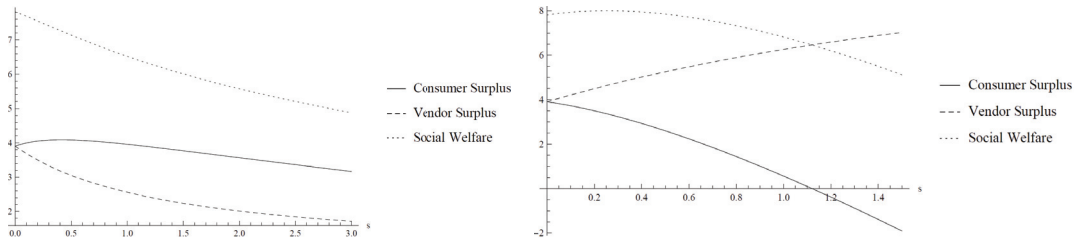
The Peltzman-like effect, as pointed out earlier, negatively impacts social welfare. From the traditional perspective, the Akerlof-like effect can lead to market failures; hence, by extrapolation, one may interpret its effect as also decreasing social welfare. For these reasons, we expect the bias to have a negative impact on consumer and social welfare metrics. In that regard, the theorem may seem counterintuitive.

Our analysis finds that the bias may have a positive impact on social welfare because of the Akerlof-like effect. It allows ill-informed consumers to further overestimate the

quality, which decreases the aggregated distrust factor. This in turn leads to well-informed consumers also purchasing more. Consequently, the vendor has an incentive to marginally improve the software quality. Thus, the social welfare may improve. Additionally, the second order effect term identified in Equation 3.7 may contribute positively to social welfare also – for example, if the proportion of ill-informed consumers is small and the Akerlof-like effect is positive. When both these effects are combined, bias creates a positive impact on social welfare.

Since we do not capture the welfare parameters in reduced form expressions, our next objective is to provide insights based on some numerical simulations. Figure 3.2 plots the changes to vendor profit, consumer surplus, and social welfare with respect to the bias for two specific values of σ. One can see that $s = 0$ yields the optimal social welfare when σ is small (in Figure 3.2(a)), but that is not the case when σ is large (in Figure 3.2(b)). These results can be readily understood from Equations 3.6 and 3.7. Note that when s increases, ill-informed consumers contribute to social welfare through the Akerlof-like effect by exhibiting higher trust in the quality but undermine it because of the Peltzman-like effect by over-exerting risky behavior. When σ is large as in Figure 3.2(b), it corresponds to more consumers receiving (upwardly) a biased signal. For that case, when s is relatively small, the Akerlof-like effect dominates, and the increment in s only improves the welfare. In contrast, when σ is small as in Figure 3.2(a), most of the consumers observe an unbiased signal. They debias the signal even more if s increases, which harms the social welfare through the Akerlof-like effect. In addition, their choice of risky-behavior also becomes further from optimal. Thus, in the case where σ is small, a larger s does society no good.

In addition to the change in social welfare with respect to s, Figure 3.3 shows the variation of social welfare when both s and r change. Recall that r represents the subjective probability of the signal being biased. When r is small, consumers are naive in the sense that they believe the signal that they observe. As a result, they do not put much weight on debiasing the signal. In such a case, the change in s mostly influences ill-informed consumers, and thus the increase in s can raise the social welfare because the Akerlof-like effect is dominant. On the other hand, with a large r, consumers are suspicious of the



(a) (b)

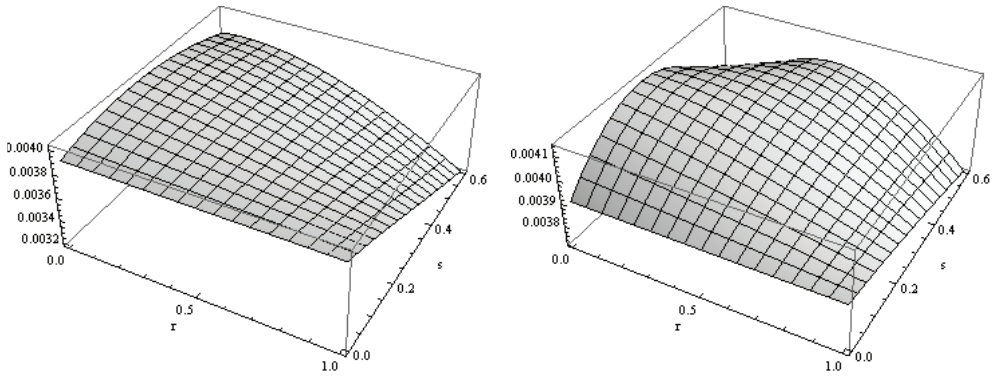
Figure 3.2.: The changes in consumer surplus, the vendor surplus, and the social welfare with respect to the amount of bias.

The utility function is $u(x;\beta,q)=\sqrt{\beta qx}-\lambda x$ and the quality perception is $\tilde{q}_1=q(s+1)$. The cost function is $c(q)=kq^2$. $\sigma=.2$ and $r_1=r_2=.5$ for the first and $\sigma=.6$ and $r_1=r_2=.2$ for the second graph. Other parameters are set to be $k=.1$ and $\lambda=.05$.

signal that they observe and put more weight on debiasing. Hence, an increase in s mostly affects well-informed consumers and thus fails to improve social welfare. In addition, we also observe that at the point where r is equal to σ, social welfare does not increase when s increases. The following theorem proves another associated result.

Theorem 3.4.3 *There exist scenarios where social and consumer welfare are higher without security software in the market.*

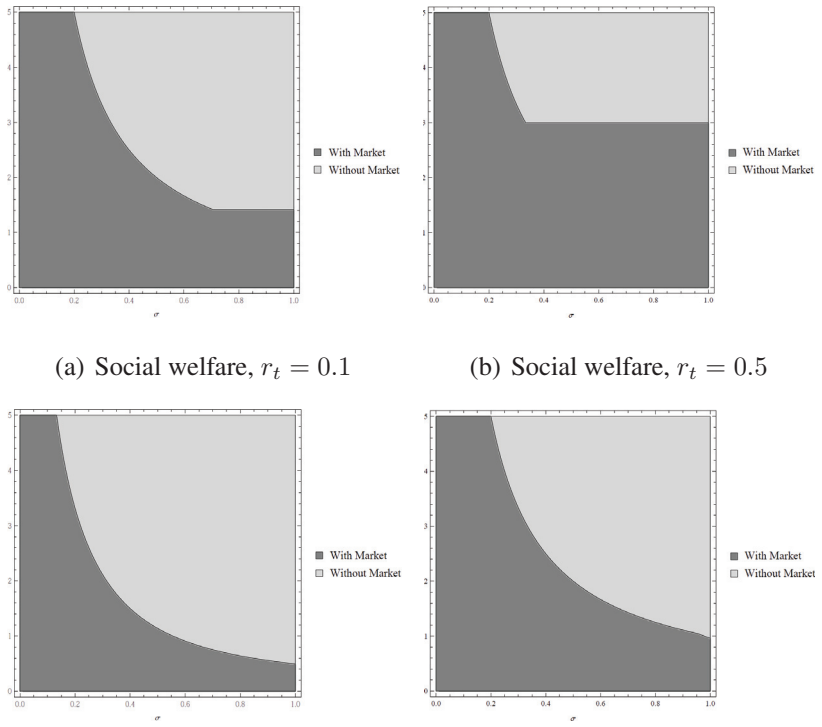
This is an interesting result. In the previous explanations, we explained how the Akerlof-like effect moderates the negative impact on social welfare because the value of the aggregated distrust factor decreases. Also, we explained how having some bias can improve social welfare. However, when both the amount of bias and the proportion of overestimating consumers are large, social welfare can be worse than without any market. Figure 3.4 illustrates the same phenomenon. The first two graphs represent social welfare, while the last two graphs represent consumer surplus. The dark gray area corresponds to the case where the welfare parameters are higher if security software exists, while the light



(a) (b)

Figure 3.3.: The social welfare evaluated with parameters s and r.

Note: $r_1 = r_2 = r$ is assumed. The utility function, the quality perception, and the cost function are the same as fig 3.2. $\sigma = .5$ for (a) and $\sigma = .9$ for (b). Other parameters are set equal to fig 3.2.



(a) Social welfare, $r_t = 0.1$ (b) Social welfare, $r_t = 0.5$

(c) Consumer surplus, $r_t = 0.1$ (d) Consumer surplus, $r_t = 0.5$

Figure 3.4.: The parameter region where the social welfare is larger with the market or without the market.

Note: The utility function is $u(x;\beta,q) = \sqrt{\beta q x} - \lambda x$. The quality perception is $\tilde{q}_1 = q(s+1)$. The cost function is $c(q) = kq^2$. The other parameters are set to be $k = .1, \lambda = .05$

gray region represents the case where the welfare parameters are actually lower if there is security software in the market.[7]

[7]One might wonder whether the results from Theorem 3.4.3 hold if we model a strategic hacker. Let $e(q, D(p,q))$ be hackers' effort level, which is a function of the quality of security software and the mass of protected consumers. For simplicity, assume linear relationship $e(q, D(p,q)) = \alpha q + \beta D(p,q)$. As the consumer and social welfare are continuous functions, it is obvious that there exists some small α and β that the existence in Theorem 7 still holds. This may change the parameter region of which such a case arises, but we can still show that such a case exists.

3.5 Generalizations

In this section, we relax several assumptions from the main model and discuss results.

3.5.1 Multiple Consumer Segments

In our main model, we simplify consumers' observation of security software quality by assuming only two types of signals, biased and accurate signals. As a result, consumers belong to one of the two segments, ill- and well-informed. In this subsection, we relax such an assumption by allowing more than two types of signals. Let there be T consumer segments each identified by $t = 1, \ldots, T$ and occurring in proportions σ_t, where $\sum_{t=1}^{T} \sigma_t = 1$. Define $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_T\}$. The quality perceived by consumers in each segment is $\tilde{q}_t = g_t(q; s_t)$. Let $\mathbf{s} = \{s_\tau\}_{\tau=1,\ldots,T}$ be a vector of the amount of bias each segment observes. For ease of representation, assume that the consumer segments are sorted in terms of bias. Specifically, we assume that a larger t implies less bias. By construction, $g_t(q; s_t) > g_t(q; s_t)$ for $\forall t < t$. Let $\mathbf{r_t} = \{r_t^\tau\}_{\tau=1,\ldots,T}$ be a vector of the weight that consumer segment t puts on believing that he belongs to segment τ and, in that case, the debiased quality is $g_\tau^{-1}(\tilde{q}_t; s_\tau)$. Therefore, $\tilde{U}_t^* = \max_x \sum_{\tau=1}^{T} r_t^\tau u(x; g_\tau^{-1}(\tilde{q}_t; s_\tau), \beta)$. As in our main model, we further assume that the expected perceived utility can be separated into two terms: $\tilde{U}_t^* = \beta m(q) n_t(\mathbf{s}, \mathbf{r_t})$. Here, $n_t(\mathbf{s}, \mathbf{r_t})$ is smaller for larger t (i.e., $n_t(\mathbf{s}, \mathbf{r_t}) > n_t(\mathbf{s}, \mathbf{r_t}) \forall t < t$), which is consistent with the influence of the information uncertainty being smaller for more well-informed consumers. Similar to Equations 3.4 and 3.5, we define:

$$W_t = \frac{\sum_{t=1}^{t} \frac{\sigma_t}{n_t(\mathbf{s}, \mathbf{r_t})}}{\sum_{t=1}^{t} \sigma_t}, \text{ and}$$

$$W(\mathbf{s}, \{\mathbf{r}_t\}_{t=1,\ldots,T}, \boldsymbol{\sigma}) = W_t | t \in \arg\max_{\tau} \sum_{t=1}^{\tau} \sigma_t \quad W_\tau^{-1} \quad .$$

The main difference is that instead of point values, we now use vectors as parameters of the functions. The equilibrium price similarly is:

$$p^* \in \frac{m(q^*)}{2W(\mathbf{s}, \{\mathbf{r}_t\}_{t=1,\ldots,T}, \boldsymbol{\sigma})},$$

and it is unique iff $W(\mathbf{s}, \{\mathbf{r}_t\}_{t=1,...,T}, \boldsymbol{\sigma})$ is a singleton. If $W(\mathbf{s}, \{\mathbf{r}_t\}_{t=1,...,T}, \boldsymbol{\sigma}) = W_c$, then only consumer segments $t = 1, ..., c$ purchase the product but not the segments $t = c+1, ..., T$. As before, however, the optimal vendor profit and optimal quality are non-zero, finite, unique, and continuous in all the parameters.

We can interpret $W(\mathbf{s}, \{\mathbf{r}_t\}_{t=1,...,T}, \boldsymbol{\sigma})$ as before. The term W_c expresses the weighted average of $\frac{1}{n_t\,(s,r_t\,)}$, where the weights correspond to the proportions of consumer segments in the market. The term $(\ \sum_{t=1}^{c} \sigma_t\ )$ represents the market size of the consumers of the segment $t = 1, ..., c$. The variations of the aggregated distrust factor, $W(\mathbf{s}, \{\mathbf{r}_t\}_{t=1,...,T}, \boldsymbol{\sigma})$, with respect to the exogenous parameters are also mostly the same. Hence, the results in the main model hold regarding the price, quality, vendor profit, and welfare metrics with respect to exogenous parameters even with multiple consumer segments.

3.5.2 Ill-Informed Consumers Underestimate Software Quality

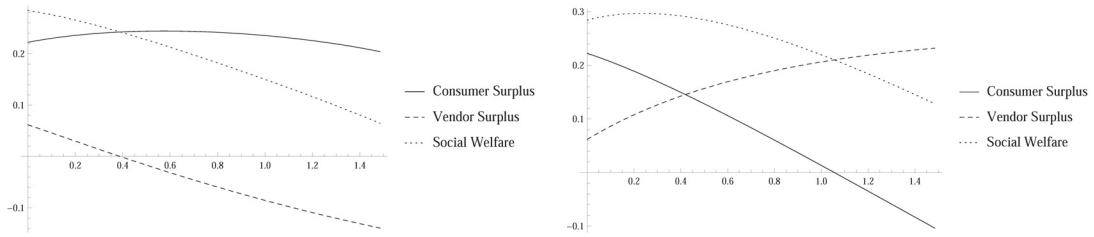
Earlier, we assume that ill-informed consumers observe a signal with only a positive bias (i.e., they are overly optimistic about the security protection from the software). In this subsection, we allow the bias to be negative (i.e., ill-informed consumers are now pessimistic about the quality of the security software and are not appreciative of the software). Suppose ill-informed consumers perceive the quality of the software to be worse than its actual quality. In other words, $\tilde{q}_1 = g(q; s) < q = \tilde{q}_2$. Let $\frac{\partial g}{\partial s} < 0$.

Note that, as before, the consumers try to debias the perceived quality of the security software. However, in this case, they suspect that the actual quality may be higher, as opposed to lower, than the signal $\tilde{q}_t$ they receive. As a result of debiasing upward instead of downward, the well-informed consumers exhibit higher perceived expected utility than the ill-informed ones. Symmetric to the main model, if the bias size s becomes larger, well-(ill-)informed consumers' perceived expected utility shifts higher (lower). In the abstract sense, similar to the main model, it results in one segment exhibiting larger demand while the demand is smaller for another segment. For these reasons, all of our main results remain qualitatively similar to the results generated by the main model.

3.5.3 Perceived Expected Utility is Non-separable

In the main model, we assumed that the perceived expected utility can be separated into the quality term $m(q)$ and the information uncertainty term $n_t(s, r_t)$ for analytical tractability. In this subsection, we relax that assumption. However, because the problem becomes intractable, we use numerical simulation to show that the main results continue to hold.

For this subsection, we consider the utility $u(x; q, \beta) = \ln(\beta x) - \frac{\lambda x}{q}$. The first term characterizes decreasing marginal utility from engaging in risky behavior, while the cost from the risk is linearly increasing. Then, the optimal expected utility is $EU_t^* = \ln\left(\frac{\beta}{\lambda}\frac{\tilde{q}\cdot g^{-1}(\tilde{q};s)}{r\tilde{q}+(1-r)g^{-1}(\tilde{q};s)}\right) - 1$. Obviously, the function is non-separable, and so we do not have a term equivalent to the aggregated distrust factor. Therefore, we consider numerical simulation and find that the results are consistent with the main model. Figure 3.5 plots the variation of the welfare metrics with respect to the amount of bias s. The results are consistent with Theorem 3.4.2 as the social optimal is not at the point where $s = 0$ and the social welfare can be increased with s.



(a) (b)

Figure 3.5.: The changes in consumer surplus, the vendor surplus, and the social welfare with respect to the amount of bias.

Note: The utility function is $u(x; \beta, q) = \ln(\beta q x) - \frac{\lambda x}{q}$ and the quality perception is $\tilde{q}_1 = q(s + 1)$. The cost function is $c(q) = \frac{kq^2}{2}$. $\sigma = .2$ and $r_1 = r_2 = .5$ for the first and $\sigma = .6$ and $r_1 = r_2 = .2$ for the second graph. The other parameters are set to be $k = .03$ and $\lambda = 1$.

3.5.4 Endogenous choice of r_t

In the main model, we assumed that the parameter r_t, the weight that consumers place on the debiased quality, is an exogenous parameter. In this subsection, we discuss an extension where r_t is endogenized. There are two possible entities in the model that can influence r_t: the vendor and the consumers. The vendor can possibly influence perceptions through warning messages on the software or through articles accessed by the different consumer segments. On the consumer side, it is possible that consumers somehow learn about the scenario regarding their beliefs.

If the vendor can manipulate r_t, it is obvious that the vendor always prefers $r_t = 0$. That is, the vendor wants consumers to believe naively in the perceived quality. On the other hand, if every consumer myopically chooses r_t, then we can prove that she will find it optimal to set $r_t = \sigma$, independent of the segment she belongs to. In other words, consumers weigh their possible biased observation as being equal to the true probability. The welfare implications of these cases are not ex ante clear. When $r_t = \sigma$ as opposed to $r_t = 0$, consumer loss from the Peltzman-like effect is smaller. However, $r_t = \sigma$ leads to lower quality product compared with the case where $r_t = 0$ because the vendor has to account for consumer suspicion of the software – the suspicion can be attributed to the Akerlof-like effect. To study this further, we conducted a numerical analysis using the same utility function as specified in Figure 3.4.

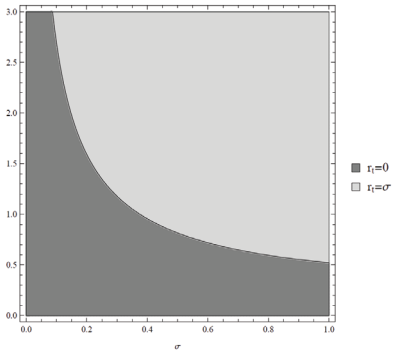


Figure 3.6.: The parameter region where the social welfare is larger with $r_t = 0$ or $r_t = \sigma$.

Note: The utility function is $u(x; \beta, q) = \sqrt{\beta q x} - \lambda x$. The cost function is $c(q) = \frac{kq^2}{2}$ and the quality perception is $\tilde{q}_1 = q(s + 1)$. Other parameters are set to be $k = .1, \lambda = .05$

Figure **??** demonstrates the social welfare comparison between the two cases. If the bias or the proportion of ill-informed consumers is sufficiently large, the society is better off when $r_t = \sigma$ compared with $r_t = 0$. However, if both uncertainty parameters are small, $r_t = 0$ is better for social welfare. Interestingly, this result demonstrates that the case where consumers know the probability of the signal being biased and set $r_t = \sigma$ accordingly is not always the social optimal.

3.6 Discussions and Conclusions

In the information security context, decisions are often made without a clear understanding of the expected losses (because consumers often do not know the probability and/or the value of the loss) suffered from breaches. The popular press as well as security researchers have even documented this lack of clarity. Yet, prior works on the economics of information security have not studied the welfare implications of such uncertainty with respect to losses. In this paper, we study how bias in consumers' estimation of software quality impacts welfare outcomes. We do so by developing a game theoretic model. Our model has two unique features. First, we model how some users observe signals regarding

software quality with a positive bias. Second, all users – including the well- and ill-informed consumers – engage in risky but value adding behavior based on their perception of the software's quality. We compute the equilibrium of our game and, based on that, develop insights.

Our paper is the first to demonstrate an interesting dynamic between two distinctive effects – the Peltzmann- and Akerlof-like effects. The Akerlof-like effect, which occurs when consumer perception of product quality is uncertain, causes the market to be unsustainable by deincentivizing the vendor in the market to improve quality. The Peltzman-like effect drives consumers, who tend to overestimate the software quality, to engage in more risky behavior. While those two effects seem to be both harmful to the consumer and social welfare, the interaction between them can create surprising results such as when the social welfare improves with bias. The reason is that the upward-bias of quality perception due to the Peltzman-like effect serves as a beneficial tool. It allows consumers to form a trust in the quality of the security software and increases product demand. Consequently, the vendor is encouraged to invest in improving the quality, which otherwise might not have occurred in a market dominated by the Akerlof-like effect alone. Thus, in cases where the loss from suboptimal consumer behavior can be offset by gain from higher quality, the larger bias benefits the consumer/social welfare.

In addition, our research yields many important practical implications at multiple levels. At the individual consumer level, our paper highlights the Peltzman effect in the context of information security. It models the consumers' false perception of the software quality and the consequently potentially dangerous impact on those purchasing security software. Thereby, we account for possible differences between realized and perceived utilities. We wish to highlight this difference as many firms in reality continue to believe that simply adopting security software automatically yields a higher level of protection. When users are overly optimistic about the software's quality, the loss from the Peltzman-like effect may be more severe than having no protection at all. In this regard, we emphasize the importance of treating information security holistically by also investing in educating end-users in addition to implementing the security software. Lastly, for a policymaker, our

model shows that simply reducing bias might not always benefit society. On the other hand, it could actually harm social welfare by collapsing the market. However, a scenario where an increase in bias increases social welfare will not occurs if policymakers can educate consumers about the informational structure of bias.

Given that this is the first paper to have studied the welfare implications of information uncertainty about losses, we have considered a rather simplistic setting. There are several ways in which the model can be extended. An obvious extension is to consider a more competitive market, but we believe that the Peltzmann- and Akerlof-like effects will continue to impact welfare. Another extension is to study the issue in the presence of negative network externality – an aspect which has been considered in many recent papers using game theory to study information security issues. The welfare implications of quality uncertainty in the presence of a negative network effect of information security is more difficult to predict. Hence, it could be another potential avenue for future research.

Bibliography

Akerlof, G. A. (1970). The market for lemons: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.

Altonji, J. G. and Pierret, C. R. (2001). Employer learning and statistical discrimination. *The Quarterly Journal of Economics*, 116(1):313–350.

Andow, B., Nadkarni, A., Bassett, B., Enck, W., and Xie, T. (2016). A study of grayware on google play. In *Security and Privacy Workshops (SPW), 2016 IEEE*, pages 224–233. IEEE.

Andrews, D. W. K. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica*, 67(3):543–563.

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (May 23, 2016). *Machine Bias: There's software used across the country to predict future criminals.*

Arora, A., Caulkins, J. P., and Telang, R. (2006). Research note-sell first, fix later: Impact of patching on software quality. *Management Science*, 52(3):465–471.

Arrow, K. J. (1998). What has economics to say about racial discrimination? *The journal of economic perspectives*, 12(2):91–100.

August, T. and Tunca, T. I. (2006). Network software security and user incentives. *Management Science*, 52(11):1703–1720.

Ba, S. and Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS Quarterly*, 26(3):243–268.

Basch, M. F. (1983). The perception of reality and the disavowal of meaning. *The annual of psychoanalysis*, 11:125–153.

Beatty, R. P. and Ritter, J. R. (1986). Investment banking, reputation, and the underpricing of initial public offerings. *Journal of financial economics*, 15(1):213–232.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017a). A convex framework for fair regression. In *Fairness, Accountability, and Transparency in Machine Learning (FATML)*.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2017b). Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint arXiv:1703.09207*.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890.

Berto Villas-Boas, S. (2007). Vertical relationships between manufacturers and retailers: Inference with limited data. *The Review of Economic Studies*, 74(2):625–652.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 23*.

Bonnet, C. and Dubois, P. (2010). Inference on vertical contracts between manufacturers and retailers allowing for nonlinear pricing and resale price maintenance. *The RAND Journal of Economics*, 41(1):139–164.

Bresnahan, T. (1987). Competition and collusion in the american automobile industry: The 1955 price war. *Journal of Industrial Economics*, 35(4):457–82.

Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. (2013). Controlling attribute effect in linear regression. In *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*, pages 71–80.

Center., T. U. S. C. R. (November 28, 1975). *Age Discrimination Act.*

Chellappa, R. K. and Pavlou, P. A. (2002). Perceived information security, financial liability and consumer trust in electronic commerce transactions. *Logistics Information Management*, 15(5/6):358–368.

Chen, D., Sain, S. L., and Guo, K. (2012). Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19(3):197–208.

Chen, Y.-n. and Png, I. (2003). Information goods pricing and copyright enforcement: Welfare analysis. *Information Systems Research*, 14(1):107–123.

Chenok, P. B. (1994). Perception vs. reality. *Journal of Accountancy*, 177(1):47.

Christin, N. (2011). Network security games: combining game theory, behavioral economics, and network measurements. In *Decision and Game Theory for Security*, pages 4–6. Springer.

Christin, N., Egelman, S., Vidas, T., and Grossklags, J. (2012). It's all about the benjamins: An empirical study on incentivizing users to ignore security advice. In *Financial Cryptography and Data Security*, pages 16–30. Springer.

Coase, R. H. (1972). Durability and monopoly. *The Journal of Law and Economics*, 15(1):143–149.

Coate, S. and Loury, G. C. (1993). Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240.

Cohen, A. and Einav, L. (2003). The effects of mandatory seat belt laws on driving behavior and traffic fatalities. *Review of Economics and Statistics*, 85(4):828–843.

Commission., T. U. S. E. E. O. (March 2, 1979). *Uniform guidelines on employee selection procedures*.

Conlon, C. T. (2012). A dynamic model of prices and margins in the lcd tv industry. *unpublished*.

DeAngelo, L. E. (1981). Auditor size and audit quality. *Journal of accounting and economics*, 3(3):183–199.

Dey, D., Lahiri, A., and Zhang, G. (2012). Hacker behavior, network effects, and the security software market. *Journal of Management Information Systems*, 29(2):77–108.

Dimoka, A., Hong, Y., and Pavlou, P. A. (2012). On product uncertainty in online markets: Theory and evidence. *MIS Quarterly*, 36(2):395–426.

Dipietro, B. (2014). Survey roundup: False sense of security? Online; posted 14-November-2014.

Dubé, J.-P., Fox, J. T., and Su, C.-L. (2012). Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica*, 80(5):2231–2267.

Fang, H. and Moro, A. (2011). Theories of statistical discrimination and affirmative action: A survey. In *Handbook of social economics*, volume 1, pages 133–200. Elsevier.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 259–268.

Fish, B., Kun, J., and Lelkes, Á. D. (2015). Fair boosting: a case study. In *2nd Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML 2015)*.

Gautier, E. and Kitamura, Y. (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica*, 81(2):581–607.

Gefen, D., Karahanna, E., and Straub, D. W. (2003). Trust and tam in online shopping: an integrated model. *MIS Quarterly*, 27(1):51–90.

Goh, G., Cotter, A., Gupta, M. R., and Friedlander, M. P. (2016). Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2415–2423.

Gowrisankaran, G. and Rysman, M. (2012). Dynamics of consumer demand for new durable goods. *Journal of political Economy*, 120(6):1173–1219.

Graham, J. D. and Garber, S. (1984). Evaluating the effects of automobile safety regulation. *Journal of Policy Analysis and Management*, 3(2):206–224.

Guo, K. H. (2013). Revisiting the human factor in organizational information security management. *ISACA Journal*, 6:1–5.

Hall, A. (2005). *Generalized Method of Moments*. Advanced Texts in Econometrics. OUP Oxford.

Hall, A. R. and Inoue, A. (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics*, 114(2):361–394.

Hall, A. R. and Pelletier, D. (2011). Non-Nested Testing in Models Estimated via Generalized Method of Moments. Technical Report 2.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323.

Hendel, I. and Nevo, A. (2006). Measuring the implications of sales and consumer inventory behavior. *Econometrica*, 74(6):1637–1673.

Hoffer, G. E. and Pratt, M. D. (1987). Used vehicles, lemons markets, and used car rules: Some empirical evidence. *Journal of Consumer Policy*, 10(4):409–414.

Hu, W.-M., Xiao, J., and Zhou, X. (2014). Collusion or competition? interfirm relationships in the chinese auto industry. *The Journal of Industrial Economics*, 62(1):1–40.

Hui, W. (2010). Brand, knowledge, and false sense of security. *Information Management & Computer Security*, 18(3):162–172.

Jones, R. and Mendelson, H. (2011). Information goods vs. industrial goods: Cost structure and competition. *Management Science*, 57(1):164–176.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 325–333.

Kamiran, F. and Calders, T. (2010). Classification with no discrimination by preferential sampling. In *The annual machine learning conference of Belgium and The Netherlands (BENELEARN)*.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012a). Enhancement of the neutrality in recommendation. In *Proceedings of the 2nd Workshop on Human Decision Making in Recommender Systems, Dublin, Ireland, September 9, 2012*, pages 8–14.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012b). Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II*, pages 35–50.

Kamishima, T., Akaho, S., Asoh, H., and Sato, I. (2016). Model-based approaches for independence-enhanced recommendation. In *IEEE International Conference on Data*

Mining Workshops, ICDM Workshops 2016, December 12-15, 2016, Barcelona, Spain., pages 860–867.

Kannan, K., Rahman, M., and Tawarmalani, M. (2016). Economic and policy implications of restricted patch distribution. *Management Science*, 62(11):3161–3182.

Kannan, K. and Telang, R. (2005). Market for software vulnerabilities? think again. *Management Science*, 51(5):726–740.

Kirtland, K. A., Porter, D. E., Addy, C. L., Neet, M. J., Williams, J. E., Sharpe, P. A., Neff, L. J., Kimsey, C. D., and Ainsworth, B. E. (2003). Environmental measures of physical activity supports: perception versus reality. *American journal of preventive medicine*, 24(4):323–331.

Lahiri, A. (2012). Revisiting the incentive to tolerate illegal distribution of software products. *Decision Support Systems*, 53(2):357–367.

Lee, R. S. (2013). Vertical integration and exclusivity in platform and two-sided markets. *The American Economic Review*, 103(7):2960–3000.

Lim, W. S. (2000). A lemons market? an incentive scheme to induce truth-telling in third party logistics providers. *European Journal of Operational Research*, 125(3):519–525.

Luo, R. (2015). The operating system network effect and carriers dynamic pricing of smartphones.

Nair, H. (2007). Intertemporal price discrimination with forward-looking consumers: Application to the us market for console video-games. *Quantitative Marketing and Economics*, 5(3):239–292.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342.

Newman, L. H. (2017). Encryption apps help white house staffers leak—and maybe break the law. Online; posted February-2017.

Peltzman, S. (1975). The effects of automobile safety regulation. *The Journal of Political Economy*, 84(4):677–725.

Png, I. P. and Wang, Q.-H. (2009). Information security: Facilitating user precautions vis-à-vis enforcement against attackers. *Journal of Management Information Systems*, 26(2):97–121.

Pope, A. T. and Tollison, R. D. (2010). "rubbin'is racin"': evidence of the peltzman effect from nascar. *Public Choice*, 142(3-4):507–513.

Prasad, V. and Jena, A. B. (2014). The peltzman effect and compensatory markers in medicine. *Healthcare*, 2(3):170–172.

Ragan, S. (2013). Mcafee and office depot study indicates dissonance between respondents' perception and reality. Online; posted 31-October-2013.

Resnick, P., Kuwabara, K., Zeckhauser, R., and Friedman, E. (2000). Reputation systems. *Communications of the ACM*, 43(12):45–48.

Ristanoski, G., Liu, W., and Bailey, J. (2013). Discrimination aware classification for imbalanced datasets. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 1529–1532.

Rivers, D. and Vuong, Q. (2002). Model selection tests for nonlinear dynamic models. *Econometrics Journal*, 5(1):1–39.

Rudin-Brown, C. and Jamson, S. (2013). *Behavioural Adaptation and Road Safety: Theory, Evidence and Action*. CRC Press.

Rust, J. (1987). Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society*, pages 999–1033.

Sellars, W., Sellars, W., and Sellars, W. (1963). *Science, perception and reality*. Routledge & Kegan Paul London.

Silverman, C. and Singer-Vine, J. (2016). Most americans who see fake news believe it, new survey says. Online; posted December-2016.

Smith, R. J. (1992). Non-nested tests for competing models estimated by generalized method of moments. *Econometrica*, 60(4):973–980.

Stiglitz, J. E. (2000). The contributions of the economics of information to twentieth century economics. *The Quarterly Journal of Economics*, 115(4):1441–1478.

Su, C.-L. and Judd, K. L. (2012). Constrained optimization approaches to estimation of structural models. *Econometrica*, 80(5):2213–2230.

Swartz, J. and della Cava, M. (2016). The fake web: why we're so apt to believe fake news, apps and reviews. Online; posted December-2016.

Vigna, G. (2014). Antivirus isn't dead, it just can't keep up. Technical report, Technical Report. Lastline Labs.

Vrolix, K. (2006). Behavioural adaptation, risk compensation, risk homeostasis and moral hazard in traffic safety, literature review. Hasselt University. RA-2006-95.

Warkentin, M., Crossler, R. E., and Malimage, N. (2012). Are you sure you are safe?: Perceived security protection as an enabler of risky it behavior. In *Proceedings of the 2012 International Federation of Information Processing (IFIP)*.

Wineburg, S. and McGrew, S. (2016). Evaluating information: The cornerstone of civic online reasoning. Technical report, Stanford History Education Group, Robert R. McCormick Foundation.

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach*. South-Western, Cengage Learning, 5th edition.

Xie, J. and Shugan, S. M. (2001). Electronic tickets, smart cards, and online prepayments: When and how to advance sell. *Marketing Science*, 20(3):219–243.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, pages 2450–2473.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2017a). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1171–1180.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2017b). Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, pages 962–970.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 325–333.

Zliobaite, I., Kamiran, F., and Calders, T. (2011). Handling conditional discrimination. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, pages 992–1001.

APPENDICES

A. Detail of the simulation in Chapter 1

The dimension of the product characteristics on utility function X_{jt} is set to be 2, where the first characteristic is constant and the second is randomly generated independently across products and periods. The cost side characteristics Y_{jt} includes X_{jt} and one additional characteristic also drawn independently. The characteristics are drawn from a normal distribution of mean 0 and standard deviation $.1$. The unobserved error terms ξ_{jt} and λ_{jt} are also drawn from a normal distribution of mean zero and standard deviation 1, independently across products and markets. The true values of parameters other than price coefficient α are $\beta = (2., 1.)$ and $\gamma = (3., 0., 1.)$. Those values are chosen to ensure that the marginal cost does not fall below zero, and the resulting share of outside option is not too close to zero for the invertibility of Δ. Given the generated characteristics and the errors, the prices are simulated by solving the profit maximization problem by sequential least square quadratic programming. The results are robust to variety of parameter setting and distributional assumption.

B. Detail of the estimation procedure in Chapter 1

B.1 Hyper parameter setting

The discounting factor β is set to be .9 for dynamic models both on demand and supply side. The draw of consumer types is generated from Halton sequence. The number of consumer segments is set to be 7. The initial market size M_{ij1} is defined by the sum of the sales over the considered period in the subcategory that j belongs to, divided by the number of consumer segments.

B.2 Converting supply side constraints to FOC

The equilibrium constraints on supply side includes the retailer's profit maximization. In the estimation, we substitute it by first order condition. Let us define the derivative of the demand function with respect to price, $\frac{\partial D^m_{ijt}}{\partial p_{jt}}$, to be another set of endogenous variable of MPEC that represents the derivative of the demand function from a consumer i of a product j at period t evaluated at the realized price. Also define $\frac{\partial D^m_{jt}}{\partial p_{jt}}$ be a derivative of the overall demand function, again at the observed price. In static pricing model, the MPEC constraints are converted to:

$$\frac{\partial D^m_{ijt}}{\partial p_{jt}} = M^m_{ijt} s^m_{ijt}(1 - s^m_{ijt})$$

$$\frac{\partial D^m_{jt}}{\partial p_{jt}} = \sum_i \frac{\partial D_{ijt}(p_{jt})}{\partial p_{jt}}$$

$$D_{jt}(p_{jt}) + \frac{\partial D^m_{jt}}{\partial p_{jt}}(p_{jt} - MC_{jt}) = 0$$

$$MC_{jt} = X^s_{jt}\gamma + \lambda_{jt}$$

$$\forall (j,t).$$

In dynamic pricing model, FOC include a derivative of the value function. In addition to the ones above, we define two sets of additional endogenous variables: the realized value

function of product j at period t, v_{jt}, and the derivative of value function at next period with respect to current price evaluated at the observed price, $\frac{\partial V_{jt+1}}{\partial p_{jt}}$. Then FOC and the Bellman equations translate to MPEC constraints:

$$D_{jt}(p_{jt}) + \frac{\partial D^m_{jt}}{\partial p_{jt}}(p_{jt} - MC_{jt}) + \beta \frac{\partial V_{jt}}{\partial p_{jt}} = 0$$
$$v_{jt} = D_{jt}(p_{jt} - MC_{jt}) + \beta v_{jt+1}$$
$$\forall (j,t).$$

As we do not parametrically estimate the value function, the difficulty arises to calculate the derivative of the value function. The state variable at $t+1$ that are influenced by p_{jt} are the market size of consumer segments M_{ijt+1}. Thus, define the derivative of the value with respect to market size, $\frac{\partial V_{jt+1}}{\partial M_{ijt+1}}$, as another set of endogenous variable. Then,

$$\begin{aligned} \frac{\partial V_{jt+1}}{\partial p_{jt}} &= \sum_i \frac{\partial V_{jt}}{\partial M_{ijt+1}} \frac{\partial M_{ijt+1}}{\partial p_j t} p_{jt} \\ &= \sum_i \frac{\partial V_{jt}}{\partial M_{ijt+1}} \left(-\frac{\partial D_{ijt}(p_{jt})}{\partial p_{jt}} \right). \end{aligned}$$

We still have to approximate $\frac{\partial v_{jt}}{\partial M_{ijt+1}}$. One methodology is to utilize the estimated values of v_{jt}. The realized value v_{jt} should be equal to the value function evaluated at the realized state Ω_{jt}. Therefore, by comparing v_{jt} and M_{ijt}, we are able to infer how value function changes with respect to M_{ijt}. In the estimation, we do so by linear approximation such as

$$\frac{\partial V_{jt}}{\partial M_{ijt+1}} = \frac{1}{2} \left(\frac{v_{jt+1} - v_{jt}}{M_{ijt+1} - M_{ijt}} + \frac{v_{jt} - v_{jt-1}}{M_{ijt} - M_{ijt-1}} \right).$$

C. Related work to Chapter 2

This section reviews the previous work on fairness-aware machine learning algorithms.Table C.1 compares our algorithm with the existing ones.

These algorithms can be classified into two categories: Algorithms of the first category process datapoints before or after putting them into classifier or regressor. Such an

Table C.1.: List of fair estimators and their capabilities.

algorithms	categorical sensitive attrs	numeric sensitive attrs	explanatory attrs	classification	regression
Kamiran et al. (Kamiran and Calders, 2010)	✗	✗	✗	✓	✗
Zliobaite et al. (Zliobaite et al., 2011)	✗	✗	✓	✓	✗
Kamishima et al. (Kamishima et al., 2012b)	✗	✗	✗	✓	✗
Calders et al. (Calders et al., 2013)	✗	✗	✓	✓	✓
Zemel et al. (Zemel et al., 2013)	✓	✗	✗	✓	✓
Fish et al. (Fish et al., 2015)	✗	✗	✗	✓	✗
Feldman et al. (Feldman et al., 2015)	✓	✗	✗	✓	✓
Zafar et al. (Zafar et al., 2017b)	✓	✓	✗	✓	✗
This paper	✓	✓	✓	✓	✓

Note: "Categorical sensitive attrs" indicates that an algorithm can deal with more than binary sensitive variables. "Numeric sensitive attrs" indicates that an algorithm can deal with continuous sensitive variables. "explanatory attrs" indicates that an algorithm utilizes some variables that justify the treatment (e.g. the effect of working hours on wages) (Zliobaite et al., 2011). The checkmark indicates the capability of the algorithm in the corresponding aspect.

algorithm typically transforms training datasets so as to remove any dependency between the sensitive attribute and target attribute. The advantage of these algorithms is generality: they can be combined with a larger class of off-the-shelf algorithms for classification and regression. Moreover, the transformed data can be considered as a "fair representation" (Feldman et al., 2015) that is free from discrimination. The biggest downside of these algorithms that they treat a classifier as a black-box, and as a result, they need to change the datapoints drastically, which tends to reduce accuracy. Regarding the algorithms of this category, Kamiran et al. (Kamiran and Calders, 2010) proposed a data-debiasing scheme by using a ranking algorithm. They were inspired by the idea that the datapoints close to the class borderline are prone to discrimination, and they resample datapoints so as to satisfy fairness constraints. Zliobaite et al. (Zliobaite et al., 2011) argued that some part of discrimination is explainable by some attributes. They also proposed resampling and relabelling methods that help in training fair classifiers. Zemel et al. (Zemel et al., 2013) proposed a method to learn a discrete intermediate fair representation. Feldman et al. (Feldman et al., 2015) considered a quantile-based transformation of each attribute. Hardt et al. (Hardt et al., 2016) studied the condition of equalized odds, and provided a post-processing method that fulfills the condition.

Algorithms of the second category directly classify or regress datapoints. Such algorithms tend to perform well in practice since they do not need to conduct explicit data transformation that loses some information. The downside of these algorithms is that one needs to modify an existing classifier for each task. Regarding the algorithms of this approach, Ristanoski et al. (Ristanoski et al., 2013) proposed a version of support vector machine (SVM), called SVMDisc, that involves a discrimination loss term. Fish et al. (Fish et al., 2015) shifted the decision boundary of the classical AdaBoost algorithm so that fairness is preserved. Goh et al. (Goh et al., 2016) considered a constrained optimization that satisfies various constraints including the one of fairness. Kamishima et al. (Kamishima et al., 2012b) proposed prejudice index and proposed a regularizer to reduce prejudice. Zafar et al. (Zafar et al., 2017b) considered a constrained optimization

for classification tasks that maximizes accuracy (resp. fairness) subject to fairness (resp. accuracy) constraint.

Our two-stage approach lies somewhere between the data preprocessing approach and direct approach. The first stage of 2SDR transforms datasets to make the classifier or regressor in the second stage fair. Unlike most data preprocessing algorithms, the transformation of the first stage in 2SDR conducts the minimum amount of transformation that is primarily intended for linear algorithms, and thus, it does not degrade the original information by much. Moreover, any class of linear algorithm can be used in the second stage, and as a result our algorithm can handle more diverse range of tasks and conditions than the existing algorithms can.

Note that other tasks have been considered in the literature of fairness-aware machine learning. To name a few, Kamishima et al. (Kamishima et al., 2012a; Kamishima et al., 2016) considered methods for removing discrimination in recommendation tasks. Joseph et al. (Joseph et al., 2016) considered fairness in the context of online content selection. Bolukbasi et al. (Bolukbasi et al., 2016) considered fairness in dense word representation learnt from text corpora.

Table D.1.: List of regression or classification datasets.

datasets	Regression or Classification	D	N
Adult	Classification	49	45,222
Communities & Crime (C&C)	Regression	101	1,994
Compas	Classification	12	5,855
German	Classification	47	1,000
LSAC	Classification	24	20,798

Note: D is the number of binary or numeric attributes (after expanding unordered categorical attributes into dummies (i.e., set of binary dummy attributes)), and N is the number of datapoints.

D. Summary of the datasets in the main analysis of Chapter 2

Table D.1 summarizes the datasets used in the main analysis.

D.1 Other Datasets in Chapter 2

Furthermore, we conducted additional experiments on two other datasets (Table D.2). The ProPublica Compas dataset (Angwin et al., 2016) is a collection of criminal offenders screened in Broward County, Florida during 2013-2014, where x is a demographic and criminal record of offenders and y is whether or not a person recidivated within two years after the screening. We set sex as the sensitive attribute s. The Law School Admissions

Table D.2.: Results for the Compas and LSAC datasets.

Algorithm (Dataset)	P%-rule	Accuracy
OLS (Compas)	0.59	0.73
2SDR (Compas)	0.92	0.73
OLS (Compas-R)	0.19	0.65
2SDR (Compas-R)	0.93	0.65
OLS (LSAC)	0.21	0.75
2SDR (LSAC)	0.86	0.73

Note: We balanced training data by resampling in LSAC dataset to cope with class inbalance problem. Compas-R is a version of the Compas dataset where predictive attributes are dropped: In this version, we dropped the attributes of the original dataset whose correlation with y was stronger than 0.3. This significantly reduces the prediction accuracy and fairness of the OLS estimator which tries to utilize the available information as much as possible. Unlike OLS, the fairness of 2SLS does not decrease even if these attributes are dropped.

Council (LSAC) dataset [1] is a survey among students attending law schools in the U.S. in 1991, where y indicates whether each student passed the first bar examination. We set whether or not the race of the student is black as the sensitive attribute. Similar to the German dataset, we used $2/3$ (resp. $1/3$) of the datapoints as training (resp. testing) datasets, and results are averaged over 100 runs. The results, shown in Table D.2 implies that 2SDR complies with the 80%-rule with an insignificant deterioration in classification performance on these datasets.

[1] http://www2.law.ucla.edu/sander/Systemic/Data.htm

Table E.1.: Performance of 2SDR on the Adult dataset where s is (sex, age).

Algorithm	P%-rule	CC	Accuracy
OLS	0.30	0.22	0.84
2SDR	0.65	0.10	0.82

Note: We show p%-rule (resp. correlation coefficient, CC) with respect to sex (resp. age). Both fairness criteria are improved by using 2SDR.

E. Other settings

This section shows results with several other settings.

Multiple s: Here, we report the result for multiple sensitive attributes. Table E.1 lists the results for the Adult dataset, where s is sex (binary) and age (numeric). One can see that (i) 2SDR reduces discrimination for both of sensitive attributes with a very small deterioration on the classification performance, and (ii) the power of removing discrimination is weaker than in the case of applying 2SDR to a single s.

Effect of ordinal transformation: The method proposed by Feldman et al. (Feldman et al., 2015) conducts a quantile-based transformation. We have also combined the transformation with 2SDR. Let $x_{i,(k)}$ be the k-th attribute in x_i. A quantile-based transformation maps each attribute $x_{i,(k)}$ into its quantile rank among its sensitive attributes s_i:

$$\mathrm{Rank}_{i,k} = \frac{\sum_{j \in \{1,2,\ldots,n\}: s_i = s_j} \mathrm{I}[x_{i,(k)} > x_{i,(k)}]}{|\{j \in \{1,2,\ldots,n\} : s_i = s_j\}|}. \tag{E.1}$$

Feldman et al. (Feldman et al., 2015) showed that the dependence between $x_{i,(k)}$ and s can be removed by using such a quantile-based transformation (c.f. Figure 1 in Feldman et al. (Feldman et al., 2015)). Table E.2 and E.3 list the results of applying the transformation

Table E.2.: Classification results for the Adult dataset, with or without the ordinal transformation of Eq. (E.1).

	Without ordinal trans.		With ordinal trans.	
Algorithm	P%-rule	Accuracy	P%-rule	Accuracy
OLS	0.30	0.84	0.29	0.83
2SDR	0.83	0.82	0.82	0.82
OLS (cont. only)	0.22	0.81	0.06	0.80
2SDR (cont. only)	0.88	0.79	0.87	0.78

Note: "With ordinal trans." (resp. "Without ordinal trans.") indicates an ordinal transformation is conducted (resp. is not conducted) for each attribute. "cont. only" indicates that non-numeric attributes in x are discarded beforehand.

Table E.3.: Regression results for the C&C dataset.

	Without ordinal trans.		With ordinal trans.	
Algorithm	MD	RMSE	MD	RMSE
OLS	0.22	0.14	0.23	0.16
2SDR	0.02	0.18	0.02	0.19

of Eq. (E.1) for each non-binary attribute. Applying an ordinal transformation slightly decreased accuracy (or increased RMSE in regression), as it discards the modal information on the original attribute.

Table E.4.: Classification result of 2SDR combined with logistic regression.

Algorithm	dataset	P%-rule	Accuracy
2SDR	Adult	0.72	0.83
2SDR	German	0.80	0.73

Generalized linear models: We also tried logistic regression in the second stage classifier. Logistic regression is a binary classification model that assumes the following relation between the attributes x_i and target y_i:

$$\mathbb{P}[y_i = 1|x_i] = \frac{1}{1 + e^{-x_i\,\beta}}, \tag{E.2}$$

where β is the model parameter to be learnt. Table E.4 shows the results of classification when we replaced the second-stage classifier with the logistic regression. Compared with a linear model (Ridge classifier), this yielded a lower p%-rule in the Adult dataset. This fact is consistent with Theorem 2.4.2. It states that u is asymptotically uncorrelated to s: However, a non-linear map such as the sigmoid function in Eq. (E.2) can cause bias between the mapped u and s. We should also note that the more involved non-linear second stage classifiers, such as naive Bayes classifiers, support vector machines, and gradient boosting machines, resulted in a significantly lower p%-rule than logistic regression because of their strong non-linearity.

F. Proof of Lemmas and Propositions in Chapter 3

F.1 Proof of Lemma 3.3.1 (on Page 82)

Lemma 3.3.1 *For a given β, the perceived expected utility is higher for the ill-informed consumer than for the well-informed one. Hence, the market share from the well-informed consumers is smaller than or equal to that from ill-informed ones. That is, $\tilde{U}_1 > \tilde{U}_2$ and $\beta_2^* \geq \beta_1^*$.*

Proof We prove the first part initially:

$$\begin{aligned}
\tilde{U}_1 &= (1-r_1)u(x;\tilde{q}_1,\beta) + r_1 u(x; g^{-1}(\tilde{q}_1;s),\beta) \\
&= (1-r_1)u(x;g(q;s),\beta) + r_1 u(x;q,\beta) \\
&> (1-r_1)u(x;q,\beta) + r_1 u(x;q,\beta) \\
&= u(x;q,\beta) \\
&= (1-r_2)u(x;q,\beta) + r_2 u(x;q,\beta) \\
&> (1-r_2)u(x;q,\beta) + r_2 u(x; g^{-1}(q;s),\beta) \\
&= (1-r_2)u(x;\tilde{q}_2,\beta) + r_2 u(x; g^{-1}(\tilde{q}_2;s),\beta) \\
&= \tilde{U}_2.
\end{aligned}$$

Both inequalities hold because the utility is increasing in quality and $g^{-1}(q;s) < q < g(q;s)$. It proves the first part of the lemma. Next,

$$\max_x \tilde{U}_1 = \tilde{U}_1(x_1^*) > \tilde{U}_1(x_2^*) > \tilde{U}_2(x_2^*) = \max_x \tilde{U}_2. \tag{F.1}$$

The first inequality comes from the unique optima and the second is from the first part of the lemma.

Given Equation F.1, it is obvious that $\beta_2^* > \beta_1^*$ since

$$U_{no}(\beta_1^*) = \tilde{U}_1(\beta_1^*, q) > \tilde{U}_2(\beta_1^*, q),$$

which means that a consumer whose β is at β_1^* does not purchase the product if she belongs to $t = 2$.

■

F.2 Proof of Lemma 3.3.2 (on Page 83)

Lemma 3.3.2 *$m(q)$ is a continuous, differentiable, and strictly increasing function of q. $n_t(s, r_t)$ is a continuous and differentiable function of s and r_t. $\frac{\partial n_1(s,r_1)}{\partial s} > 0$; $\frac{\partial n_2(s,r_2)}{\partial s} < 0$; and $\frac{\partial n_t(s,r_t)}{\partial r_t} \leq 0$ for $t = \{1, 2\}$ but the inequality is strict only if $s > 0$. The utility from not purchasing the security software, $U_{no}(\beta) = 0$ for any β.*

Proof Recall that from Equation 3.3, we have

$$\beta m(q) n_t(s, r_t) = \max_x \tilde{U}_t = \tilde{U}_t^* \tag{F.2}$$

$$\tilde{U}_t = (1 - r_t)u(x; \tilde{q}_t, \beta) + r_t u(x; g^{-1}(\tilde{q}_t; s), \beta). \tag{F.3}$$

It is clear that the function on the right hand side is a continuous differentiable function of q, s and r_t. Therefore, we can apply the envelope theorem to imply that $m(q)$ and $n_t(s, r_t)$ are continuous and differentiable.

To compute $\frac{\partial n_t(s,r_t)}{\partial s}$, we invoke the envelope theorem, Following that,

$$\frac{\partial \tilde{U}_1^*}{\partial s} = (1 - r_1)\frac{\partial u(x_1^*; g(q; s), \beta)}{\partial q}\frac{\partial g(q; s)}{\partial s}.$$

Note that, by assumption, $\frac{\partial u(x_1^*; g(q;s), \beta)}{\partial q} > 0$ and $\frac{\partial g(q;s)}{\partial s} > 0$. Consequently, because of Equation F.2, $\frac{\partial n_1(s,r_1)}{\partial s} > 0$. The proof is similar to establishing that $\frac{\partial n_2(s,r_2)}{\partial s} < 0$ because $\frac{\partial g^{-1}(q;s)}{\partial s} < 0$.

If $s = 0$, for $t = \{1, 2\}$, we can also apply the envelope theorem:

$$\frac{\partial \tilde{U}_t^*}{\partial r_t} = -u(x_t^*; \tilde{q}_t, \beta) + u(x_t^*; g^{-1}(\tilde{q}_t; s), \beta).$$

Because $u(x, q, \beta)$ is a strictly increasing function of q and $\tilde{q}_t > g^{-1}(\tilde{q}_t; s)$, $\frac{\partial \tilde{U}_t^*}{\partial r_t} < 0$. Hence, $\frac{\partial n_t(s,r_t)}{\partial r_t} < 0$.

Next, we prove that $U_{no} = \max_x u(x; 0, \beta) = 0$. Since $m(0) = 0$ by assumption,

$$\begin{aligned}
\beta m(0) n_2(s, r_2) &= 0 = \max_x \; r_2 u(x; g^{-1}(0; s), \beta) + (1 - r_2) u(x; 0, \beta) \;\leq \max_x u(x; 0, \beta) \\
\beta m(0) n_1(s, r_1) &= 0 = \max_x \left(r_1 u(x; 0, \beta) + (1 - r_1) u(x; g(0; s), \beta) \right) \geq \max_x u(x; 0, \beta)
\end{aligned}$$

$0 \leq \max_x u(x; 0, \beta) \leq 0$. Thus, $\max_x u(x; 0, \beta) = 0$. ■

F.3 Proof of Proposition 3.4.1 (on Page 85)

Proposition 3.4.1 *The optimal vendor profit and optimal quality are non-zero, finite, unique, and continuous in all the parameters s, r_t, and σ_t. The equilibrium price is:*

$$p^* \in \frac{m(q^*)}{2W(s, r_1, r_2, \sigma)},$$

and it is unique iff $W(s, r_1, r_2, \sigma)$ is a singleton. If $W(s, r_1, r_2, \sigma) = W_1$, only the ill-informed consumers are served; if $W(s, r_1, r_2, \sigma) = W_2$, both well- and ill-informed consumers are served. The implicit function that finds the unique optimal quality is

$$\frac{m\,(q^*)}{4} \sum_{t=1}^{t} \sigma_t \quad W_t^{-1} - c\,(q^*) = 0 \text{ if } W_t \in W(s, r_1, r_2, \sigma).$$

Proof

$$\pi = \begin{cases} p \quad (1 - \sigma)(1 - \frac{p}{m(q) n_2(s, r_2)}) + \sigma(1 - \frac{p}{m(q) n_1(s, r_1)}) \quad - c(q) & \text{if } 0 < p \leq m(q) n_2(s, r_2) \\ p\sigma \quad 1 - \frac{p}{m(q) n_1(s, r_1)} \quad - c(q) & \text{if } m(q) n_2(s, r_2) < p \leq m(q) n_1(s, r_1) \\ -c(q) & \text{if } p > m(q) n_1(s, r_2). \end{cases}$$

The profit function is clearly a continuous function. We then demonstrate that the equilibrium price and quality are bounded. Recall that the feasible ranges are $q \in [0, \infty)$ and $p \in [0, \infty)$. If $p > m(q) n_1(s, r_1)$, $\pi \leq 0$ with the inequality being strict for $q > 0$. Now, consider $q = \epsilon > \; 0$ and $p = m(q) n_2(s, r_2)$. Then,

$$\begin{aligned}
\pi &= \sigma n_2(s, r_2) m(\;) \quad 1 - \frac{n_2(s, r_2)}{n_1(s, r_1)} \Big) - c(\;) \\
\frac{\partial \pi}{\partial \epsilon} &= \sigma n_2(s, r_2) m\,(\;) \quad 1 - \frac{n_2(s, r_2)}{n_1(s, r_1)} \Big) - c\,(\;).
\end{aligned}$$

Recall that $\lim_{\epsilon \to 0} c(\epsilon) = 0$. From Lemma 3.3.2, $\lim_{\epsilon \to 0} m(\epsilon) > 0$. Therefore $\lim_{\epsilon \to 0} \frac{\partial \pi}{\partial \epsilon} > 0$. It implies that there exists an $\epsilon > 0$ such that $\pi > 0$. Hence, $p^* \leq m(q)n_1(s, r_1)$.

For a given q, $\pi \leq m(q)n_1(s, r_1) - c(q)$ because $p^* \leq m(q)n_1(s, r_1)$ and $D(p, q) \leq 1$. Note that, by assumptions that $m(q)$ is concave and $c(q)$ is convex in q, $m(q)n_1(s, r_1) - c(q)$ is strictly concave in q. Also, from the assumptions, $m(0)n_1(s, r_1) - c(0) = 0$, $m'(0)n_1(s, r_1) - c'(0) > 0$, and $\lim_{q \to \infty} m(q)n_1(s, r_1) - c(q) = \infty$. It implies that $\bar{q} > 0$, $m(\bar{q})n_1(s, r_1) - c(\bar{q}) = 0$ and $m'(q)n_1(s, r_1) - c'(q) < 0$ for $q \geq \bar{q}$. Therefore, as a follow up to the first statement, $\pi < 0$ for $q > \bar{q}$. So, $q^* \in [0, \bar{q}]$.

From Weirstrauss' theorem, an optimal solution exists since the objective function is continuous and the feasible region is closed and continuous. By Fermat's theorem, the optimal solution can lie only on boundaries, non-differentiable points, or stationary points (obtained from the first order conditions). Boundaries $p = 0$ and $p = m(q)n_1(s, r_1)$ can be ruled out since $\pi \leq 0$ and we can argue along the lines of the previous paragraphs that those prices are infeasible as optimal solutions to our problem. Similarly, the profit generated by the stationary points are at least as much as when price is at the discontinuity point $(m(q)n_2(s, r_2))$. So, the optimal price must belong to the set $\{p_1 = \frac{m(q)}{2W_1}, p_2 = \frac{m(q)}{2W_2}\}$. At $p = p_1$, the resulting profit is $\pi_1 = \frac{m(q)\sigma}{4W_1} - c(q)$; and at $p = p_2$, the profit is $\pi_2 = \frac{m(q)}{4W_2} - c(q)$.

Now we compare π_1 and π_2. Suppose $\{\sum_{t=1}^{1} \sigma_t \quad W_1^{-1} = \sigma n_1(s, r_1) > \{\frac{\sigma}{n_1(s,r_1)} + \frac{1-\sigma}{n_2(s,r_2)}\}^{-1} = \{\sum_{t=1}^{2} \sigma_t \quad W_2^{-1}$. Then, $\pi_1 > \pi_2$. In this case, $p_1 > m(q)n_2(s, r_2)$ is satisfied:

$$
\begin{aligned}
&\sigma n_1(s, r_1) > \left\{\frac{\sigma}{n_1(s, r_1)} + \frac{1-\sigma}{n_2(s, r_2)}\right\}^{-1} \\
&\Leftrightarrow \frac{1}{\sigma n_1(s, r_1)} < \frac{1-\sigma}{n_2(s, r_2)} + \frac{\sigma}{n_1(s, r_1)} \\
&\Leftrightarrow 1 < (1-\sigma)\sigma \frac{n_1(s, r_1)}{n_2(s, r_2)} + \sigma^2 \\
&\Leftrightarrow (1-\sigma^2)n_2(s, r_2) < (1-\sigma)\sigma n_1(s, r_1) \\
&\Leftrightarrow 1 + \frac{1}{\sigma} n_2(s, r_2) < n_1(s, r_1) \\
&\Rightarrow 2n_2(s, r_2) < n_1(s, r_1) \ (\because \sigma < 1) \\
&\Leftrightarrow m(q)n_2(s, r_2) < \frac{m(q)n_1(s, r_1)}{2} = p_1
\end{aligned}
$$

Thus the stationary point $p = p_1$ exists and it is the equilibrium price.

Similarly, when $\sum_{t=1}^{1} \sigma_t \quad W_1^{-1} > \sum_{t=1}^{2} \sigma_t \quad W_2^{-1}$, the stationary point $p = p_2$ is the equilibrium price. Finally, when $W_1 = W_2$, both $p = p_1$ and $p = p_2$ are valid as stationary points.

Lastly, we show that the continuity of equilibrium outcome with respect to the parameters s, r_t, and σ. It is clear that W is a continuous function of all the parameters. Thus, the equilibrium quality and profit (which are defined by continuous implicit function of W) are also continuous in parameters. The same holds for the price except for the case where $W_1 = W_2$. ■

F.4 Proof of Lemma 3.4.1 (on Page 85)

Lemma 3.4.1 *$W_t \in W(s, r_1, r_2, \sigma)$ is non decreasing in r_t and non increasing in σ. With respect to s, the function decreases when $W(s, r_1, r_2, \sigma) = W_1$ but may increase or decrease when $W(s, r_1, r_2, \sigma) = W_2$.*

Proof First we show the result regarding r_t. Because W is continuous, it is sufficient that we show $\frac{\partial W_1}{\partial r} > 0$ and $\frac{\partial W_2}{\partial r} > 0$.

$$\begin{aligned}
\frac{\partial W_1}{\partial r_1} &= -\frac{1}{\sigma n_1(s, r_1)^2} \frac{\partial n_1(s, r_1)}{\partial r_1} \\
\frac{\partial W_1}{\partial r_2} &= 0 \\
\frac{\partial W_2}{\partial r_1} &= -\frac{\sigma}{n_1(s, r_1)^2} \frac{\partial n_1(s, r_1)}{\partial r_1} \\
\frac{\partial W_2}{\partial r_2} &= -\frac{1-\sigma}{\sigma n_2(s, r_2)} \frac{\partial n_2(s, r_2)}{\partial r_2}
\end{aligned}$$

From Lemma 3.3.2, $\frac{\partial n_t(s, r_t)}{\partial r_t} \leq 0$. Thus all the four equations are greater than zero.

Similarly, with respect to σ:

$$\begin{aligned}
\frac{\partial W_1}{\partial \sigma} &= 0 \\
\frac{\partial W_2}{\partial \sigma} &= -\frac{1}{n_2(s, r_2)} + \frac{1}{n_1(s, r_1)}.
\end{aligned}$$

The first equation is (weakly) less than zero. The second is less than zero because $n_2(s, r_2) \leq n_1(s, r_1)$ from Lemma 3.3.1.

Lastly, we show the result with respect to s. If $W_1 < W_2$, $W = W_1$ and

$$\frac{\partial W_1}{\partial s} = -\frac{1}{n_1(s, r_1)^2} \frac{\partial n_1(s, r_1)}{\partial s} \leq 0$$

from Lemma 3.3.2. If $W_1 > W_2$, $W = W_2$ and

$$\frac{\partial W_2}{\partial s} = -\frac{1-\sigma}{n_2(s, r_2)^2} \frac{\partial n_2(s, r_2)}{\partial s} - \frac{\sigma}{n_1(s, r_1)^2} \frac{\partial n_1(s, r_1)}{\partial s}.$$

From Lemma 3.3.2, $\lim_{\sigma \to 0} \frac{\partial W_2}{\partial s} = -\frac{1}{n_2(s,r_2)^2} \frac{\partial n_2(s,r_2)}{\partial s} \geq 0$ and $\lim_{\sigma \to 1} \frac{\partial W_2}{\partial s} = -\frac{1}{n_1(s,r_1)^2} \frac{\partial n_1(s,r_1)}{\partial s} \leq 0$ ■

F.5 Proof of Theorem 3.4.1 (on Page 86)

Theorem 3.4.1 *When the parameters change, the following hold at the equilibrium.*

1. *If r_t increases (equivalently, decreases): the quality and the profit decrease (increase); if the price is a singleton, it also decreases (increases).*

2. *If σ increases (decreases): the quality and the profit increase (decrease); if the price is a singleton, it also increases (decrease).*

3. *If s increases: the profit, quality, and price* may *decrease.*

Proof From Lemma 3.4.1, we know the sensitivity of $W_t \in W(s, r_1, r_2, \sigma)$ with respect to the exogenous parameters. Hence, we only need to show the equilibrium changes with $W_t \in W(s, r_1, r_2, \sigma)$.

First, we show that q^* is decreasing in $W_t \in W(s, r_1, r_2, \sigma)$. Recall that the equilibrium quality satisfies $\frac{m(q^*)}{4W_t} \sum_{t=1}^{t} \sigma_t - c(q^*) = 0$ for $W_t \in W(s, r_1, r_2, \sigma)$ and is unique. So, $\frac{m(q^*)}{4c(q^*)} = W_t \sum_{t=1}^{t} \sigma_t$ for $W_t \in W(s, r_1, r_2, \sigma)$. Since we assumed $m(q)$ to be concave and $c(q)$ convex, $m(q)$ is decreasing; but $c(q)$ is always increasing in q. So, $\frac{m(q^*)}{c(q^*)}$ decreases as q^* increases. It implies that if the right hand side increases, the equilibrium

quality decreases. When s and r_t change, the treatment is straightforward because the right hand side directly changes with W. We then consider the case of σ_t next.

Consider the case $W_2 \in W(s, r_1, r_2, \sigma)$. The right hand side can be substituted as $\sum_{t=1}^{2} \sigma_t = 1$. Thus, q^* increases when W_2 decreases. If $W_1 = 1/n_i(s, r_i) \in W(s, r_1, r_2, \sigma)$, the right hand side is σW_1. In this case, $W(s, r_1, r_2, \sigma)$ is independent of σ, and q^* increases when $W(s, r_1, r_2, \sigma)$ decreases.

Regarding the price, recall that the price is $p^* = \frac{m(q^*)}{2W(s,r_1,r_2,\sigma)}$. Suppose $W_t \in W(s, r_1, r_2, \sigma)$ increases. From the previous argument, we know that $m(q^*)$ increases. So the numerator increases and denominator decreases. Thus p^* decreases.

The profit can be written as $\frac{m(q^*)}{4W(s,r_1,r_2,\sigma)} \sum_{t=1}^{t} \sigma_t - c(q^*)$ in any case. By applying the envelope theorem, the profit decreases if $W_t \in W(s, r_1, r_2, \sigma)$ increases. ■

F.6 Proof of Theorem 3.4.2 (on Page 90)

Theorem 3.4.2 *Social welfare can increase with the amount of bias.*

Proof Let the utility function be $u(x; \beta, q) = \sqrt{\beta q x} - \lambda x$ and the quality perception be $\tilde{q}_1 = q(s + 1)$. The first term is the benefit consumers obtain from the risky behavior, and the second term is the cost associated with the risky behavior. The benefit is assumed to be concave, and the cost is assumed to be linear. The consumers are assumed to be heterogeneous in the benefit but not in the cost. This specification satisfies the assumptions we made in the section 3.2. To show it also satisfies the separability assumption in 3.2.1, first the expected utility for the ill-informed consumers is

$$EU_1 = r_1 \left(\overline{\beta q x} - \lambda x \right) + (1 - r_1) \left(\overline{\beta q (1 + s) x} - \lambda x \right).$$

From the first order condition, $x_i^* = \frac{\beta q}{4\lambda^2} \left(r_1 + (1 - r_1) \overline{(1 + s)} \right)^2$. The optimal expected utility is:

$$\begin{aligned} EU_1^* &= \beta \cdot \frac{q}{4\lambda} \left(r_1 + (1 - r_1) \overline{(1 + s)} \right)^2 \\ &\equiv \beta m(q) n_1(s, r_1). \end{aligned}$$

Similarly, the expected utility of well-informed consumers can be written as $EU_2^* = \beta m(q) n_2(s, r_2)$ where $n_2(s, r_2) = \left(\frac{r_2}{\sqrt{(1+s)}} + (1 - r_2) \right)^2$.

The social welfare expression is extensive, but $\frac{\partial SW}{\partial s}$ at $s = 0$ can be simplified as

$$\frac{\partial SW}{\partial s}_{s=0} = \frac{\sigma - r_1\sigma - r_2(1-\sigma)}{1024k\lambda^2},$$

which can be positive if r_1 and r_2 are sufficiently small. For example, the parameter we used for figure 3.2(b) $(\sigma, r_1, r_2) = (.6, .2, .2)$ satisfies this condition. Therefore, when s is increased from zero, the social welfare may improve. It implies that the social welfare can increase with the amount of bias. ■

F.7 Proof of Theorem 3.4.3 (on Page 92)

Theorem 3.4.3 *There exist scenarios where social and consumer welfare are higher without security software in the market.*

Proof We use the same specification as mentioned in the proof of Theorem 3.4.2. The social welfare is zero if there is no market for the security software. Thus, we only have to show that there exists a set of parameters with which the social welfare is negative. For example, let $(s, \sigma, r_1, r_2, k, \lambda) = (3, .5, .1, .1, 1, .05)$. The social welfare evaluated at this point is approximately $-.203$. Therefore, there exists a case where the social welfare is smaller with market than without market. ■

VITA

VITA

Hajime Shimao

Santa Fe Institute
1399 Hyde Park Rd
Santa Fe, NM 87501

Phone: (+1) (765) 409-6391
Email: hajime.fr@gmail.com
website: https://sites.google.com/site/hajimeshimao/

Updated: May 23, 2018

Education

Ph.D. Economics, Purdue University August 2018

Dissertation: *Essays on Structural Econometric Modeling and Machine Learning*

M.S. Economics, Purdue University 2014

M.S. Decision Science, Tokyo Institute of Technology 2012

Concentration: Evolutionary Game Theory, Experimental Economics

Thesis Title: *Strict or Graduated Punishment? Effect of Punishment Strictness on the Evolution of Cooperation in Continuous Public Goods Games*

B.A. Psychology, University of Tokyo 2009

Concentration: Decision Theory

Thesis Title: *The Effect of Self-Esteem on the Decision Making in Bargaining Games*

Research Interests

Industrial Organization, Econometrics, Machine Learning, Applied Microeconomics, Experimental Economics

Research

Working Papers:

1. ”Cross-Validation Based Model Selection on Generalized Method of Moments with Application to Dynamic Pricing Model” [*Job Market Paper*] (with Junpei Komiyama)

2. ”Estimating Skill-Added: A Revealed Choice Set Approach of College Major and Occupation Choices” (with Xiaoxiao Li, and Sebastian Linde)

Publication and Submitted Papers:

1. ”So You Think You are Safe: Implication of Quality Uncertainty in Security Software.” (with Warut Khernamnuai and Kirthik Kannan) *R&R at* ***Management Science*** (3rd round revision)

2. ”Two-stage Algorithm for Fairness-aware Machine Learning.” (with Junpei Komiyama) *submitted.*

3. ”Reciprocity and Exclusion in Informal Financial Institutions: An Experimental Study of Rotating Savings and Credit Associations.” (with Takehiko Yamato, et. al) *submitted.*

4. ”Strict or Graduated Punishment? Effect of Punishment Strictness on the Evolution of Cooperation in Continuous Public Goods Games.” (with Mayuko Nakamaru) *published at* ***PloS one***,**8**(3).

Work in Progress:

1. ”Improving Instrumental Variables Estimation Using Support Vector Machine.” (with Junpei Komiyama, and Xiaoxiao Li)

2. ”Identifying Complementarity of Goods from Pattern Mining.” (with Junpei Komiyama)

3. ”A Game-theoretic Analysis on Fairness Criteria.” (with Junpei Komiyama)

Refereed conferences

1. Shimao H. and Komiyama J. ”Cross-Validation Based Model Selection on Generalized Method of Moments with Application to Dynamic Pricing Model” American Economic Association: ASSA Annual Meeting, Philadelphia, PA, January 2018 (poster session).

2. Komiyama J., Li X., and Shimao H. ”Improving Instrumental Variables Estimation Using Support Vector Machine” 87th Southern Economics Association Annual Meeting (SEA), Tampa, Florida, November 2017.

3. Shimao H. and Komiyama J. ”Cross-Validation Based Model Selection on Generalized Method of Moments with Application to Dynamic Pricing Model” 87th Southern Economics Association Annual Meeting (SEA), Tampa, Florida, November 2017.

4. Khern-am-nuai, W., Shimao, H., and Kannan, K. "So You Think You Are Safe: Implication of Quality Uncertainty in Security Software." Conference on Information Systems and Technology (CIST), Philadelphia, PA, October 2015.

Invited seminar presentations

Department of Economics, State University of New York, Binghamton	February, 2018
Federal Reserve Board of Governors	February, 2018
Department of Economics, Virginia Tech	January, 2018
Department of Economics, St. Gallen University	December, 2017
Department of Economics, Kent State University	November, 2017
Department of Economics, Villanova University	November, 2016
Purdue Ph.D. Research Symposium	March, 2010
Game Theory Workshop, Kyushu University	December, 2009
Human Behavior and Evolution Society of Japan, Kyushu University	November, 2009
Research Institute for Mathematical Science, Ryukoku University	

Professional Activity

Editorial Staff of ***Letters on Evolutionary Behavioral Science***	Spring 2010 - Summer 2011

Affiliations

American Economic Association (AEA), Southern Economics Association (SEA), Econometric Society

Awards, Grants, and Fellowships

Fellowship from *Japan Student Service Organization*	Summer 2012 - Summer 2015
Graduate Scholarship, Purdue University	Summer 2015 - Summer 2017

Research Experience

Research Assistant to Ralph Siebert	Summer 2015 - Summer 2017

Teaching Experience

Teaching Assistant (Undergrad Level)	
Econ 499 Honors Thesis Course	Fall 2015
Teaching Assistant (Master Level)	
Econ 510 Game Theory	Spring 2016
Mathematical Modeling in Social Science	Spring 2009
Teaching Assistant (Ph.D. Level)	
Econ 621 Applied Industrial Organization	Fall 2016
Econ 631 Industrial Organization	Spring 2016
Econ 673 Time Series Econometrics	Spring 2016
Other Teaching Experience	
Tutor for *Scientific Education Group*	2004-2006

Skills

Programming: Python, C/C++, Matlab, R

Software: Theano, AMPL, Mathematica, Stata, SAS, z-Tree

Languages: English & Japanese (fluent); Chinese & French (beginner)

Econometrics

- Econometrics - OLS, IV, Bootstrap, GMM, Panel; MLE, Binary, EM, Tobit
- Time Series Econometrics
- Advanced Econometrics - Factor Model Analysis
- Theoretical and Applied Bayesian Econometrics
- (Grades in all above Ph.D. Econometrics Courses: A or A+; Overall GPA: 3.8)

Data Experience

- ***Transaction Level Data***: Transaction data on multiple online-retailers (machine learning; pattern mining; structural econometric analysis of consumer behavior);
- ***Market Level Data***: Monthly sales and price data of video game industry (demand estimation with heterogeneous dynamic agent model; analysis of pricing behavior);
- ***Cryptocurrency Data***: Price data of multiple cryptocurrencies in different exchanges (development of automatic data acquisition and trade system; predictive analysis);
- ***Labor Data***: O*NET database 16.0 (**O*NET**); American Community Survey (**ACS**); National Longitudinal Survey of Youth (**NLSY**); Panel Study of Income Dynamics (**PSID**) (structural econometric analysis of occupational choice; estimation of worker skills).