Working Paper #11-28

October 2011

Nonlinear Pricing with Product Customization in Mobile Service Industry

Yao Luo
Pennsylvania State University

Nonlinear Pricing with Product Customization in Mobile Service Industry

Yao Luo*

Pennsylvania State University

September 25, 2011

Abstract

This paper proposes to incorporate product customization in the Maskin and Riley (1984) nonlinear pricing model in order to capture major features of mobile service data. In particular, consumers are characterized by a two-dimensional type. One dimension is observed by the provider and integrates product customization, while the other is a standard parameter of adverse selection, which is unobserved by the provider and makes it necessary for the provider to discriminate among consumers with different tastes through nonlinear pricing. We then propose a novel method to aggregate the multiple-dimensional voice consumption into one-dimensional index. We show that the model structure is identified under the following conditions: The marginal utility function is multiplicatively separable in consumers' tastes, and consumers' observed and unobserved heterogeneity are independent. Empirical results show that both dimensions of heterogeneity are important. Due to asymmetric information, 50% of the "second-best" social welfare is left "on the table" in order to screen heterogeneous consumers. Moreover, if costly product customization does not affect subscribers' utility, 20% of subscribers would not be served.

Key Words: Nonlinear Pricing, Product Customization, Mobile Service

JEL Codes: L11, L12, L25, L96

*E-mail: *yluo@psu.edu*. I am grateful for the advice and support of Isabelle Perrigne and Quang Vuong. I have also benefited from comments and discussions with Liang Chen, Frank Erickson, Paul Grieco, Mark Roberts and Felix Tintelnot. I thank the NET Institute (www.NETinst.org) for financial support.

A commodity is a good or a service completely specified physically, temporally, and spatially.
– Debreu (1959)

1 Introduction

Nowadays, modern technology is starting another era of customized mass production. The telecommunication industry is one of the first entering this era. While a traditional landline phone only allows users to talk to each other through a fixed telephone line, mobile service now enables seamless phone calls even when users are moving around wide areas. Through cellular sites and mobile-services switching centers, providers can get phone call services delivered whenever and wherever they are needed. This has dramatically increased the variety of services which can be provided. In addition, mobile service providers typically offer complex nonlinear tariffs. For example, when a consumer chooses a plan with a higher monthly fee, he usually gets a lower minute rate, more free services, a lower rate for minutes beyond the free quota and no peak-time call charge.

In this paper, we develop a bidimensional screening model to explain the observed voice consumptions and payments relying on the Maskin and Riley (1984) model. A monopoly mobile service provider has a technological infrastructure that allows consumers to customize their own services and use them in various quantities. In the market we study, the provider provides a full range of different kinds of voice service that can be temporally and spatially categorized. A particular combination represents a variety of mobile service. We assume that a customer's need for a certain variety of service is determined exogenously by his lifestyle and is observed by the monopolist. This assumption is consistent with the product customization literature (see, e.g., Bernhardt, Liu, and Serfes (2007)) and also real-life practice.[1] Hereafter, we call the variety variable location for convenience. It is any characteristic of the consumer that is observed by both the provider and the consumer but not observed by the analyst. We introduce a second dimension of consumer heterogeneity, i.e., the consumer' taste for mobile voice service, which is unobserved by the provider. Therefore, to maximize profit, the provider needs to discriminate among consumers with different tastes through nonlinear pricing. To summarize, we assume that consumers are heterogeneous along two dimensions. Namely, they are located on a segment, so that they are identified by their positions, and they also differ in their tastes. The former is observed by the provider while the latter is not.

This paper contributes to the existing literature in several dimensions. First, it offers a

[1] For example, while a phone is turned on, its geographical location can be easily determined by calculating the differences in time for a signal to travel from the mobile phone to each of several cellular towers nearby.

methodology to incorporate product customization into a nonlinear pricing model, thereby allowing for an (endogenous) continuous price schedule despite the discrete number of basic plans. Up to now, several papers have empirically analyzed nonlinear pricing using discrete choice models while considering exogenous price schedules. See, e.g., Leslie (2004) for broadway theatre tickets, McManus (2007) for speciality coffee, and Cohen (2008) for paper towel, Mostly related to our paper, Miravete (2002), Miravete and Röller (2004), Economides, Seim, and Viard (2008) and Seim and Viard (2010) study nonlinear pricing in local telephone service and mobile service industry. We propose to use a continuous framework when the large number of varieties stemming from product customization makes discrete choice models intractable.[2] On one hand, the choice set approximates a continuum of choices when the number of varieties is large enough. On the other hand, product customization can be understood as bundling of a large number of services. By the Law of Large Numbers, multiple taste parameters (typically implied by multiple services) even out in large bundles, which means that one only needs to consider the systematic difference (for example, income) across consumers.

Second, we propose a method to aggregate a multiple-dimensional variable into an one-dimensional index and we show how the parameters in the aggregation function are identified. To the best of our knowledge, we are the first to construct such an index without data on another variable (which is a function of the index) or making parametric assumptions on the unobserved variables. In the former case with additional data, the weights can be estimated by projecting the variable on the various characteristics or consumptions. For example, to aggregate the various food categories, Aguiar and Hurst (2005) derive the weights by projecting the (estimated) permanent income on the quantities consumed of various types of foods. To aggregate characteristics of houses into an one-dimensional quality index, Murphy (2007) assumes that the price of a house is a function of its quality index and derive weights by projecting the price on the house characteristics. In the latter case with parametric assumptions on the unobservable variables, the weights can be estimated by maximizing the log likelihood of the data as a function of all the parameters. For example, many empirical studies of production relationships are based on aggregate indexes of capital and labor inputs. Sankar (1970) utilizes a CES production function with a normal distributed error term and estimates the parameters using maximum likelihood. In our case, the multiple-dimensional voice consumptions are aggregated according to a Cobb-Douglas function. The difficulty is how to estimate the weights in the Cobb-Douglas function while estimating the distribution of the unobservable variables nonparametrically without additional data. Obviously the

[2]For example, the multinomial probit/logit model becomes computationally infeasible with too many alternatives. In addition, the nested multinomial logit model depends on strong assumptions.

previous two approaches do not apply to our case. Instead, we exploit the independence between the two unobserved variables to show that the weights are identified.

Third, our paper contributes to the growing literature on the mobile service industry. See Gruber (2005) for an excellent survey. It is worthnoting that the previous literature relies heavily on parametric specifications of the model. For example, using data from an experiment by the South Central Bell, Miravete (2002) constructs a fully parametric principal-agent model which provides closed-form solutions. Using data from a mobile service provider in an Asian country, Kim, Telang, Vogt, and Krishnan (2010) assume a quadratic utility function and normally distributed error terms. In this paper, we investigate the nonparametric identification of the utility function and type distribution from the observed consumptions and payments. This allows us to draw policy conclusions that are robust to functional misspecification. Moreover, in a preference-based structural model, consumers are assumed to be heterogeneous. In this kind of empirical analysis, individual-level data are preferred. However, they have been less available to researchers. For example, the efficiency of the estimates in Miravete and Röller (2004) is confined by the lack of information on individual consumptions. In this paper, we have individual-level data on more than 20,000 subscribers of a mobile service provider in China.

Fourth, our paper builds on the literature on contract models with incomplete information. Some papers focus on one-dimensional context. See, e.g., Perrigne and Vuong (2011a). Recently, several empirical studies document the implications of multidimensional heterogeneity in insurance market. See, e.g., Finkelstein and McGarry (2006) and Cohen and Einav (2007). In view of this, Aryal, Perrigne, and Vuong (2009) study the identification of insurance models with multidimensional screening. Another strand of literature studies the implicatoins of sequentially revealed bidimensional consumer type. See, e.g., Miravete (2002) and Miravete (2005) in the context of telecommunication. In this paper, we remark that mobile service subscribers are characterized by both their (unobserved) taste and (observed) location. Thus, we employ a bidimensional screening model which boils down to a series of one-dimensional screening model. We study its identification in the spirit of Guerre, Perrigne, and Vuong (2000) and Perrigne and Vuong (2011a).

Lastly, our paper contributes to the growing management literature on product customization. Product customization relates to the ability of providing individually designed products and services to every customer through a high process flexibility and integration as defined by Da Silveira, Borenstein, and Fogliatto (2001). Since the seminal paper of Thisse and Vives (1988), the Hotelling's spatial competition model has been used extensively to study product customization. See, e.g., Dewan, Jing, and Seidmann (2000), Chen and Iyer (2002) and Dewan, Jing, and Seidmann (2003). However, they mainly focus on manufacturing operations. There are still few studies dealing with product customization in service operations. Hereafter, we consider perfect product customization in the mobile service industry using a nonlinear pricing model.

The rest of the paper is organized as follows. Section 2 describes the data with a particular attention to aspects that are incorporated in the model. Section 3 presents the model. Identification and estimation are discussed in Section 4. Section 5 presents the estimation results and some counterfactuals. Section 6 concludes with future lines of research.

2 Mobile Phone Data

We collected data on voice consumptions and payments of subscribers to a mobile service provider in a major metropolitan area of China for the billing period of May 2009. Among the three mobile service providers allowed to operate in this area, the firm from which we obtained the data has 72% of the mobile subscribers. It provides mobile service under three brands, each of which has a specific target market: One for students, one for rural residents and one for business people and others. We focus on the latter, which holds an even higher market share. Thus it is reasonable to assume that the firm acts as a monopolist in this market segment.

The firm proposes a new menu of plans every year. A menu usually consists of a basic plan list and many add-on plan lists. In May 2009, eight basic plans were offered. Consumers can change their plans every month at no additional cost. Whenever an existing customer wants to change his plan or a new customer wants to join one, he can only choose from the currently proposed menu. Since it is optional for the existing customers to change to a new plan, there is still a certain proportion of them using old ones. We focus on the customers who subscribed to the eight basic plans offered in May 2009, which gives a sample with more than 20,000 observations.

Each basic plan specifies a monthly fee, a free quota of minutes, a price schedule beyond the free quota and so on. When a subscriber chooses a plan, he pays a monthly fee and is allowed to use a free quota of voice minutes. The free quota is limited to certain kinds of phone calls. Once he uses up the free quota, he incurs a constant price for overtime minutes. The basic plans distinguish outgoing and incoming calls according to the position of the subscribers when the calls are made as shown in Table 1. For example, A is the outgoing call minutes used by consumers when they are in the city.

We observe consumers' voice service consumptions of A, D, $(B+C)$ and $(E+F)$ measured in the number of minutes and their total bills in May 2009.[3] Table 2 contains some summary

[3]There is very little consumption of international calls, which account for less than 0.2% of the total

statistics. The average bill is 182.49 RMB ($26.72 US dollars). While accounting for 8% of the subscribers, this market segment contributes to 20% of the revenue. Subscribers mainly stay in the home city. Hence they consume 5 times more local call minutes than roaming minutes. Slightly less incoming call minutes are consumed than outgoing call minutes. We also calculate the number of days since their subscription dates. An average of 28 billing periods shows the loyalty of subscribers.

The firm implements nonlinear pricing across and within basic plans. First, as one moves from a cheaper plan to a more expensive one, the ratio of the monthly fee to free minutes decreases with the overtime rates decreasing, more add-ons are offered free and more types of incoming calls becoming free. Second, besides the basic plan list, there are many add-on plan lists. Some add-on plan lists give quantity discounts for certain kinds of calls. Third, there are continuous promotion events. Usually, deeper discounts are offered to the customers who consume more. Fourth, there is a bonus credit accumulation program. The accumulation rate is higher for the customers who consume more. Bonus credits can be redeemed for account balance, add-on plans or gifts. It is worth noting that almost all discounts are given to different levels of payments rather than combinations of calls. On one hand, this suggests that there is a mapping from consumers' willingness to pay to payments. On the other hand, it suggests that there is no implicit ranking of "quality" among different kinds of calls. If there were, discounts should be offered to those who consume more calls with high "quality", which is not the case in our data.

The firm can customize its services perfectly and melds the customized services with corresponding prices. First, it has the best nationwide coverage and provides a full range of phone call services.[4] Thus, the firm can deliver services whenever and whereever they are needed. Customers can customize their own services according to their needs. Second, with the help of its computer system, the mobile service provider implements the price schedule based on where, when and how subscribers consume. For each phone call, the phone numbers, location and networks of the two end nodes, starting time, ending time and duration are recorded. Rates change with all the features of the call, such as the location of the initiator and receiver, the length of the call, the time of the day, the day of the week and so on. However, the analyst does not observe all the temporal and spacial features of a phone call except the total duration for several categories. On the other hand, for consumers consuming the same quantities, the difference in their payments reflects their different locations. To see how heterogeneous consumes are in their locations, we sample consumers whose payments

consumption. To simplify the analysis, we add it to A. Following Kim, Telang, Vogt, and Krishnan (2010), we drop the observations whose total minutes are less than 10. We also drop the observations whose payments are less than 2.01 units or more than 1000 units, which account for less than 0.4% of the final sample.

[4]For example, roaming is not available to subscribers of certain plans of other brands.

fall into chosen bins. Table 3 presents the summary statistics of total minutes and payment. While the standard deviation of payment is controlled to be small in each bin, consumers' consumption of total minutes has a significantly higher variation.

Several features of the mobile service data will be incorporated in our model. First, we will aggregate the observed four-dimensional voice consumption into an one-dimensional index. We argue that consumers' decision is one-dimensional. From the consumers' point of view, if their decision is along all dimensions, no "free lunch" should be left on the table and cheaper services should be consumed more. However, this is not the case in our data. A large proportion of subscribers did not use their free minutes. If a consumer pays the monthly fee, his first usage of $B + C + F$ is free. We compare subscribers' free minutes and actual usage. Since we can not back out E from data on $(A, D, B + C, E + F)$, we assume $\frac{B}{E} = \frac{B+C}{E+F}$[5] and calculate $\frac{B+C+F}{FreeMinutes}$ for each consumer. More than 75% of observations are less than 1, 62% are less than 0.75 and 44% are less than 0.50. Moreover, cheaper calls are not consumed more. While incoming calls are always cheaper than outgoing calls, Table 2 shows that the average consumption of D is 473.44 minutes, which is even lower than the average consumption of A, 518.87 minutes. The consumption of $E + F$ is also lower than $B + C$. Hence, the decision to consume at certain time and place is determined exogenously by the consumers' lifestyle. From the monopolist's point of view, if the decision of consumers is along all dimensions, he should offer discounts along those dimensions to maximize his profit. For example, in the basic plan, the monopolist could have specified a monthly fee and a free quota for each kind of call. Instead, it specifies only a total monthly fee and a total free quota. Moreover, almost all discounts are given to different levels of payments, which is one dimensional. In sum, consumers only decide on one dimension based on their tastes.

Second, we use a continuous framework. Specifically, we consider that the price schedule offered by the monopolist is continuous. In our data, voice consumptions are continuous while the number of basic plans is less than 10. According to our conversations with several employees in this firm, there are literally thousands of codes which represent different discounts offered to consumers. Consumers using the same basic plan can be further discriminated by the nonlinear pricing specification of add-on plans, continuous promotions and bonus credit accumulation program. Thus, the firm can and does offer a large number of plans to approximate a continuous price schedule, which leaves no room for pooling.

Third, we assume that the firm's total cost function is separable across consumers with a fixed term.[6]. Mobile service is a good example of information goods. On one hand, providing

[5]This is assuming that the proportion of incoming and outgoing calls are the same when the consumer is roaming in or outside of his home province.

[6]Sundararajan (2004) intepretes it as transaction cost. In contrast, Perrigne and Vuong (2011a) assume a cost function for the total production in the context of yellow pages. We note that their *total* cost function

mobile telephone service to a customer, the variable cost of producing an additional unit of service is small. On the other hand, the most important cost is building mobile phone base stations. These stations have limited capacities which are used as long as a subscriber's cellular phone is open. Hence, the cost depends mainly on how many subscribers it serves, rather than how many minutes subscribers use.

Fourth, we assume that consumers know their types well for two reasons. On one hand, they can change their plans every month at no additional cost. On the other hand, most of them have been subscribed to this firm for a long time. Table 2 presents the average number of days since they first subscribed to the firm. On average, they have been subscribed for more than 840 days, which correspond to more than 28 months. After this long period of "learning", it is reasonable to assume that the consumers know their types perfectly well.

3 Model

Our model builds on Maskin and Riley (1984). A monopoly sells mobile voice service that may be customized into different varieties (in other words, delivered to different locations) and used by customers in varying quantities. A consumer is characterized by a pair (θ, ϵ). The term θ is his taste for voice service which is known only to him. The term ϵ is his location. It actually represents all the consumer heterogeneity which is known both to the monopolist and the consumer.[7] Assume that (θ, ϵ) is distributed as $\Phi(\cdot, \cdot)$ with a continuous density $\phi(\cdot, \cdot)$ on support $[\underline{\theta}, \overline{\theta}] \times [\underline{\epsilon}, \overline{\epsilon}]$, $0 \leq \underline{\theta} < \overline{\theta} < \infty$ and $0 \leq \underline{\epsilon} < \overline{\epsilon} < \infty$.

The utility function of a consumer of type (θ, ϵ) takes the form

$$\int_0^q v(z; \theta, \epsilon) dz - \tau(q; \epsilon),$$

where q is the quantity of voice services purchased and $\tau(q; \epsilon)$ is the total payment for q units of voice service when his location is ϵ.

The monopolist chooses optimally the function $q(\cdot, \cdot)$ and $\tau(\cdot; \cdot)$ to maximize its profit. The function $q(\cdot, \cdot)$ is defined on $\left[\underline{\theta}, \overline{\theta}\right] \times [\underline{\epsilon}, \overline{\epsilon}]$. We assume for the moment that for given ϵ, $q(\cdot, \epsilon)$ is a strictly increasing function on $[\underline{\theta}, \overline{\theta}]$. Later, we will show that with additional assumptions the resulting optimal $q(\cdot, \epsilon)$ is a strictly increasing function. For each $\epsilon \in [\underline{\epsilon}, \overline{\epsilon}]$, the payment $\tau(\cdot; \epsilon)$ is defined on $\left[0, q\left(\overline{\theta}, \epsilon\right)\right]$. The monopolist's profit can be written as

and our *separable* cost function are nonnested.

[7]In a version of Hotelling's spatial competition model, Bernhardt, Liu, and Serfes (2007) study the investment decisions on customization technology. In their model, consumers are characterized by two attributes: Spatial locations and brand name preference. The former is interpreted as the customizable dimension while the latter is non-customizable. To make comparision with Bernhardt, Liu, and Serfes (2007), the two attribute dimensions in our model are both customizable.

$$\int_{\underline{\theta}}^{\overline{\theta}} \int_{\underline{\epsilon}}^{\overline{\epsilon}} [\tau(q(\theta,\epsilon);\epsilon) - c(q(\theta,\epsilon);\epsilon)]\phi(\theta,\epsilon) d\theta d\epsilon,$$

where $c(q(\theta,\epsilon);\epsilon)$ is the the cost for producing $q(\theta,\epsilon)$ to serve a consumer at location ϵ.

The revelation principle ensures that the monopolist can restrict its attention to a direct mechanism. Moreover, since there is no asymmetric information on ϵ, the monopolist's profit is maximized if and only if it is maximized for each subpopulation with the same ϵ. Hence the monopolist's profit maximization problem boils down to a series of profit maximization problems: for any $\epsilon \in [\underline{\epsilon}, \overline{\epsilon}]$,

$$\max_{q(\cdot,\epsilon), T(\cdot;\epsilon)} \int_{\underline{\theta}}^{\overline{\theta}} [\tau(q(\theta,\epsilon);\epsilon) - c(q(\theta,\epsilon);\epsilon)]\phi(\theta|\epsilon) d\theta$$

over all $(q(\cdot,\epsilon), \tau(\cdot;\epsilon))$ that satisfies incentive compatibility (IC) and individual rationality (IR) constraints: for any $\theta \in [\underline{\theta}, \overline{\theta}]$,

$$(IC): \qquad \theta = \arg\max_{x \in [\underline{\theta}, \overline{\theta}]} \int_0^{q(x,\epsilon)} v(z;\theta,\epsilon) dz - \tau(q(x,\epsilon);\epsilon),$$

$$(IR): \qquad 0 \leq \int_0^{q(\theta,\epsilon)} v(z;\theta,\epsilon) dz - \tau(q(\theta,\epsilon);\epsilon).$$

The IC constraint says that "telling the truth" is optimal for every subscriber. Some remarks on the IR constraint are in order here. We follow Miravete (2002) and assume that the outside option gives the same utility to every subscriber. From a theoretical perspective, we avoid countervailing incentives as studied by Lewis and Sappington (1989) and Maggi and Rodriguez-Clare (1995). From an empirical perspective, this reflects that the outside option provides low utility because consumers have exogenous variety needs. To a mobile service subscriber, his outside option is using a fixed-line phone or phone booth to make calls and not getting incoming calls when away from a fixed-line phone. It is extremely inconvenient. To summarize, when the outside option provides low utility to everyone, even if there is variation, it is a valid first-order approximation to assume that it is the same to everyone.

The following assumptions are made on $v(q;\theta,\epsilon)$.

Assumption A1: *The marginal utility function $v(\cdot;\cdot,\cdot)$ is continuously differentiable on $[0,+\infty) \times [\underline{\theta}, \overline{\theta}] \times [\underline{\epsilon}, \overline{\epsilon}]$, and $\forall q \geq 0$, $\forall \theta \in [\underline{\theta}, \overline{\theta}]$ and $\forall \epsilon \in [\underline{\epsilon}, \overline{\epsilon}]$*

(i) $v(q;\theta,\epsilon) > 0,$

(ii) $v_1(q;\theta,\epsilon) < 0,$

(iii) $v_2(q;\theta,\epsilon) > 0$,

(iv) $v_{22}(q;\theta,\epsilon) \leq 0$,

(v) $\frac{\partial}{\partial\theta}\left\{\frac{-v_1(q;\theta,\epsilon)}{v(q;\theta,\epsilon)}\right\} \leq 0$,

(vi) $\frac{c_{11}(q;\epsilon)}{c_1(q;\epsilon)} > \frac{v_1(q;\theta,\epsilon)}{v(q;\theta,\epsilon)}$,

(vii) $\frac{\partial}{\partial\theta}\left\{\frac{1-\Phi(\theta|\epsilon)}{\phi(\theta|\epsilon)}\right\} \leq 0$.

Assumption A1-(i) says that the marginal utility is always positive, A1-(ii) says that the marginal utility is decreasing in the quantity purchased and A1-(iii) says the subscribers with a higher taste enjoy a higher utility across every q. Assumption A1-(iv) says that the increase in demand price is diminishing as the taste increases, while A1-(v) says that the utility function has a nonincreasing absolute risk aversion. Assumption A1-(vi) says that the cost function is "not too concave" in q. Assumption A1-(vii) says that the conditional distribution of θ has a nonincreasing inverse hazard rate.

Proposition 1: *Under Assumptions A1, the functions* $(q(\cdot,\epsilon), \tau(\cdot;\epsilon))$ *that solve the monopolist's optimization problem satisfy: there exists* $\theta_0(\epsilon) \in [\underline{\theta}, \overline{\theta}]$ *such that consumers with* $\theta < \theta_0(\epsilon)$ *are not served by the provider, and whenever* $q(\cdot,\epsilon) > 0$,

$$\tau_1\left(q\left(\theta,\epsilon\right);\epsilon\right) = v\left(q\left(\theta,\epsilon\right);\theta,\epsilon\right), \tag{1}$$

$$v(q(\theta,\epsilon);\theta,\epsilon) = c_1(q(\theta,\epsilon);\epsilon) + v_2(q(\theta,\epsilon);\theta,\epsilon)\frac{1-\Phi(\theta|\epsilon)}{\phi(\theta|\epsilon)}. \tag{2}$$

Equation (1) says that the marginal utility equals the marginal price at the designated consumption of each subscriber. Equation (2) says that the marginal utility equals the marginal cost plus a distortion term due to incomplete information.

4 Identification and Estimation

In this section, we first specify the econometric model. We then study the identification of our model. Finally, we propose a multistep estimation procedure in view of our identification results.

4.1 The Econometric Model

We first restrict the marginal utility function in view of the first order condition (1). Consider the infeasible case in which we observe θ and ϵ. For any arbitrary values of θ and ϵ, $q(\theta, \epsilon)$ in uniquely determined. Therefore, it is not possible to independently vary (q, θ, ϵ) and trace out $v_0(\cdot; \cdot, \cdot)$ on its 3 dimensional domain. In view of the identification results in Perrigne and Vuong (2011a,b), the following assumption is made on $v(z; \theta, \epsilon)$.

Assumption B1: *The consumer's marginal utility function is of the form*

$$v(z; \theta, \epsilon) = \theta\epsilon v_0(z\epsilon), \forall\theta \in [\underline{\theta}, \overline{\theta}], \forall\epsilon \in [\underline{\epsilon}, \overline{\epsilon}]$$

where $v_0(\cdot)$ *satisfies* $\forall q \in [0, +\infty)$, $v_0(q) > 0$ *and* $v_0'(q) < 0$.

Now the consumer's utility function can be written as

$$\int_0^q \theta\epsilon v_0\left(z\epsilon\right) dz - \tau(q; \epsilon) = \theta \int_0^{\epsilon q} v_0\left(z\right) dz - \tau(q; \epsilon).$$

We interpret $v_0(\cdot)$ as the base marginal utility function and ϵ as a consumption multiplier. Recall that ϵ captures the effects of the consumer heterogeneity which is known to both consumers and the monopolist but unknown to the analyst. Consumers with higher ϵ locate further away and appreciate more convenience from consuming the same amount of minutes. For example, consumers who travel more would consume more roaming minutes, which are of greater convenience than the same amount of local minutes. If we think of q as the base consumption, the consumption multiplier ϵ measures how much this quantity of consumption increases in response to a change in his base consumption.

We now consider the cost function. A similar rationale for Assumption B1 applies here. We need to further restrict the cost function. The following assumption is made on $c(q; \epsilon)$.

Assumption B2: *The cost function is of the form*

$$c(q; \epsilon) = \begin{cases} K + c_0(q\epsilon), & \textit{if } q > 0 \\ 0, & \textit{if } q = 0 \end{cases}$$

where $K \geq 0$ *and* $c_0(0) = 0$.

The term K captures the cost that is triggered by any positive usage. For example, infrastructure costs stem from keeping mobile phones connected and administration costs stem from delivering statements. The term $c_0(q\epsilon)$ expresses the cost that is related to the amount of services delivered. For example, costs stem from monitoring, recording and reporting usages. Product customization is costly. That is, delivering service to consumers locating further away is more costly. For example, the cost of delivering long distance calls is higher because more cellular sites and mobile-services switching centers are involved and it requires more financial settlements between the two providers involved. Under Assumption B1 and B2, (2) can be written as

$$\theta v_0 \left(\epsilon q \left(\theta, \epsilon\right)\right) = c_0'(\epsilon q(\theta, \epsilon)) + \frac{1 - \Phi \left(\theta|\epsilon\right)}{\phi \left(\theta|\epsilon\right)} v_0 \left(\epsilon q \left(\theta, \epsilon\right)\right). \tag{3}$$

We further restrict the distribution of consumers' type.

Assumption B3: $\theta \perp \epsilon$.

Assumption B3 is strong. However, it greatly facilitates the solution of our model. Similar assumptions are widely used in the literature. For example, Bernhardt, Liu, and Serfes (2007) assume the consumer's unobserved brand name preference is independent of his observed location. Heckman, Matzkin, and Nesheim (2010) also assume that the consumer's unobservable heterogeneity is independent of his observed characteristics. Moreover, nonparametric identification of nonlinear nonseparable structural models is often achieved under the assumption of independence between the model's latent variables and exogenous variables. Likewise, independence assumption is the key of our first identification result.

Let $f(\cdot)$ and $F(\cdot)$ denote the density and distribution functions of θ, and let $g(\cdot)$ and $G(\cdot)$ denote the density and distribution functions of ϵ. Under Assumption B3, (3) implies that $\epsilon q\left(\theta, \epsilon\right)$ would not depend on ϵ. Thus, there exists a function $Q(\cdot)$ such that $Q(\theta) \equiv \epsilon q(\theta, \epsilon)$ and

$$\theta v_0(Q(\theta)) = c_0'(Q(\theta)) + \frac{1 - F(\theta)}{f(\theta)} v_0(Q(\theta)).$$

Equation (1) then can be rewritten as

$$\frac{\tau_1(q(\theta, \epsilon); \epsilon)}{\epsilon} = \theta v_0(Q(\theta)). \tag{4}$$

We again remark that the right-hand side only depends on θ. Hence, there exists a function $T(\cdot)$ such that $\tau(q; \epsilon) = T(\epsilon q)$. It can be easily verified that $\frac{\tau_1(q(\theta,\epsilon);\epsilon)}{\epsilon} = T'(Q(\theta))$.

The necessary condition (4) becomes

$$T'(Q(\theta)) = \theta v_0(Q(\theta)).$$

The following proposition formalizes these results.

Corollary 1: *Under Assumptions B1, B2 and B3, the optimal price schdule offered by the monopolist should satisfy the following conditions:*

(i) There exists $\theta_0 \in [\underline{\theta}, \overline{\theta}]$ *such that consumers with* $\theta < \theta_0$ *are not served by the provider,*

(ii) There exists a pair of functions $(Q(\cdot), T(\cdot))$ *such that:* $\forall \theta \in [\theta_0, \overline{\theta}]$

$$T'(Q(\theta)) = \theta v_0(Q(\theta)), \tag{5}$$

$$\theta v_0(Q(\theta)) = c_0'(Q(\theta)) + \frac{1 - F(\theta)}{f(\theta)} v_0(Q(\theta)), \tag{6}$$

(iii) $\forall \theta \in [\theta_0, \overline{\theta}]$ *and* $\forall \epsilon \in [\underline{\epsilon}, \overline{\epsilon}]$, $q(\theta, \epsilon) = \frac{Q(\theta)}{\epsilon}$,

(iv) $\forall \theta \in [\theta_0, \overline{\theta}]$ *and* $\forall \epsilon \in [\underline{\epsilon}, \overline{\epsilon}]$, $\tau(q(\theta, \epsilon); \epsilon) = T(\epsilon q(\theta, \epsilon))$.

We can see from (6) that the cutoff taste is defined by

$$\theta_0 = \min \Big\{ \theta \in [\underline{\theta}, \overline{\theta}] : T(Q(\theta)) \geq K + c_0(Q(\theta)) \Big\}, \tag{7}$$

where $T(\cdot)$ and $Q(\cdot)$ are defined by (5) and (6) and boundary condition

$$T(Q(\theta_0)) = K + c_0(Q(\theta_0)). \tag{8}$$

It is easy to see that $\forall K > 0$, $Q(\theta_0) > 0$.

The variables T and q are the observed payment and consumption. Equations (5) and (6) characterize the optimal schedule and tariff $Q(\cdot)$ and $T(\cdot)$. Since $Q(\theta) \equiv \epsilon q(\theta, \epsilon)$ and $q_1(\theta, \epsilon) > 0$, we know that $Q(\cdot)$ is strictly increasing in θ. Moreover, Tirole (1988) indicates that for a given ϵ, we have $\tau_1(\cdot; \epsilon) > 0$ and $\tau_{11}(\cdot; \epsilon) < 0$. It is easy to verify that $T(\cdot)$ is strictly increasing and concave in Q. Therefore, there is a unique strictly increasing mapping between the unobserved taste θ and the observed bill T, which is the key of our second identification result. Its concavity is directly related to the magnitude of the asymmetry information relative to the distribution of consumers' taste. This is consistent with the observation that almost all discounts are based on the level of payment in practice.

Before studying the identification of the model primitives, we aggregate the multiple-dimensional consumptions into a one-dimensional consumption index. We aggregate consumptions of minutes by $q = h(q^A, q^D, q^{BC}, q^{EF})$ and parameterize it in the following way:

Assumption B4: $h(\cdot, \cdot, \cdot, \cdot)$ *is of the form*

$$h(q^A, q^D, q^{BC}, q^{EF}) = (q^A)^{\alpha^A} (q^D)^{\alpha^D} (q^{BC})^{\alpha^{BC}} (q^{EF})^{\alpha^{EF}},$$

where $\alpha^A, \alpha^D, \alpha^{BC}, \alpha^{EF} \geq 0$.

Though we could allow a more general function form, we assume a Cobb-Douglas specification for simplicity. The Cobb-Douglas specification is widely used in many empirical studies. For example, Murphy (2007) uses a Cobb-Douglas specification to aggregate characteristics of houses into an one-dimensional quality index. Moreover, first introduced by Solow (1957), the Cobb-Douglas specification is also extensively used for aggregating production function. In Consumption-Based CAPM models, the Cobb-Douglas specification is used to construct a consumption index for the representative agent. See, e.g., Dunn and Singleton (1986).

4.2 Identification

We define the game structure and the observables. The model primitives are $[v_0(\cdot), F(\cdot), G(\cdot), \alpha, K, c_0(\cdot)]$, which are the base marginal utility function, the taste distribution, the distribution of location, the weights used to aggregate the multidimensional voice consumption and cost function. The data provide information on the consumers' consumption choices and their payment. We denote the observable distribution as $\mathcal{G}(\cdot, \cdot, \cdot, \cdot, \cdot)$, where the first four arguments are the four-dimensional voice consumption and the last argument is the payment.

We proceed in several steps. First, we study the identification of the weights to aggregate different kinds of minute consumptions and the tariff function $T(\cdot)$. Second, we use a strategy similar to Perrigne and Vuong (2011a) to identify the type distribution and marginal utility function of consumers.

4.2.1 Identification of α and $T(\cdot)$

Identification of α implies the identification of the consumption index q given information on $(q^A, q^D, q^{BC}, q^{EF})$. In the previous literature, there are mainly three methods to identify the weights $(\alpha^A, \alpha^D, \alpha^{BC}, \alpha^{EF})$. The first is to parameterize the distribution of the unobservables. This is not applicable here because we want to identify the distribution of the unobservables nonparametrically. The second is to find an observable/estimable variable which is a function of the consumption index. This is not applicable here either since we do not have additional data. Third, theory may provide some information about the weights. For example, if consumers were optimizing their choices of voice consumptions, the Cobb-Douglas specification leads to have the ratio of different consumptions equal to the inverse of the ratio of their marginal prices. We have shown that this is not the case in Section 2.

Therefore, it is not possible to identify the weights by equating the ratio of consumptions with the inverse of the ratio of their prices.

Instead, we explore the restrictions we put on consumers' taste and location to identify the weights. Assumption B4 implies that

$$T = T((q^A)^{\alpha^A}(q^D)^{\alpha^D}(q^{BC})^{\alpha^{BC}}(q^{EF})^{\alpha^{EF}} \times \epsilon),$$

where $T(\cdot)$ is strictly increasing and concave.

Considering the inverse function T^{-1} gives

$$T^{-1}(T) = (q^A)^{\alpha^A}(q^D)^{\alpha^D}(q^{BC})^{\alpha^{BC}}(q^{EF})^{\alpha^{EF}} \times \epsilon.$$

In addition, the natural logarithm leads to

$$\alpha^A \log(q^A) + \alpha^D \log(q^D) + \alpha^{BC} \log(q^{BC}) + \alpha^{EF} \log(q^{EF}) = \log(T^{-1}(T)) - \log(\epsilon),$$

which can be written as

$$\alpha' Y = \Lambda(X) + e, \tag{9}$$

where $Y \equiv (\log(q^A), \log(q^D), \log(q^{BC}), \log(q^{EF}))'$, $\Lambda(\cdot) \equiv \log(T^{-1}(\cdot))$, $X \equiv T$ and $e \equiv -\log(\epsilon)$. We denote $Y^l \equiv \log(q^l)$, where $l \in \{A, D, BC, EF\}$.

We motivate the identification of α and T by comparing our econometric model with two most related ones: The transformation model and the single index model. We remark that (i) ϵ is independent of θ, and (ii) T only depends on θ. Thus $\log(\epsilon)$ is independent of $\log(T^{-1}(T))$. On one hand, our model does not specify the relationship between q^l and ϵ. That is, $\log(\epsilon)$ is not necessarily independent of q^l. Thus we cannot transform (9) into a transformation model.[8] On the other hand, if the αs were known, we can calculate the scalar dependent variable on the left hand side and our model becomes a degenerated single index model.[9]

Note that (9) continues to hold if α, Λ, and e are replaced by $c\alpha$, $c\Lambda$, and ce for any $c > 0$. It also holds if Λ and e are replaced by $\Lambda + c$ and $e - c$ for any $c \in \mathbb{R}$. Therefore, location and scale normalizations are needed for identification. Moreover, the coefficients αs are not identified if there is perfect multicollinearity among the elements in Y. The following

[8]The transformation model is $\Lambda(Y) = X'\beta + e$, where Y is a scalar dependent variable, $\Lambda(\cdot)$ is a strictly increasing function, X is a vector of explanatory variables, β is the vector of corresponding coefficients, and e is an unobserved error term independent of X.

[9]The semiparametric single index regression model is $Y = \Lambda(X'\beta) + e$, where Y is a scalar dependent variable, $\Lambda(\cdot)$ is an unknown link function, X is a vector of explanatory variables, β is the vector of corresponding coefficients, and e is an unobserved error term independent of X.

assumption formalizes these observations.

Assumption C1:

(i) $\alpha^A + \alpha^D + \alpha^{BC} + \alpha^{EF} = 1$,

(ii) $E[\log(\epsilon)|\theta] = 0$,

(iii) There exists a triple (t_1, t_2, t_3) *such that*

$$\begin{vmatrix} E[e\frac{\partial Y^A}{\partial t_1}] & E[e\frac{\partial Y^D}{\partial t_1}] & E[e\frac{\partial Y^{BC}}{\partial t_1}] & E[e\frac{\partial Y^{EF}}{\partial t_1}] \\ E[e\frac{\partial Y^A}{\partial t_2}] & E[e\frac{\partial Y^D}{\partial t_2}] & E[e\frac{\partial Y^{BC}}{\partial t_2}] & E[e\frac{\partial Y^{EF}}{\partial t_2}] \\ E[e\frac{\partial Y^A}{\partial t_3}] & E[e\frac{\partial Y^D}{\partial t_3}] & E[e\frac{\partial Y^{BC}}{\partial t_3}] & E[e\frac{\partial Y^{EF}}{\partial t_3}] \\ 1 & 1 & 1 & 1 \end{vmatrix} \neq 0. \tag{10}$$

Assumption C1-(i) says that the aggregation function has constant return to scale. Assumption C1-(ii) says that $\log(\epsilon)$ is mean-independent of θ, which is needed for the identification of $T(\cdot)$. Assumption C1-(iii) restricts how ϵ and q^l are correlated. Violation of it would lead to nonidentification. For instance, we assume that $Y^A = k_1 Y^D + k_2 Y^{BC} + k_3 Y^{EF}$, then the left-hand side of equation (10) equals 0 because the vector of column 1 can be written as a linear combination of vectors of the other columns. In this case, α is not identified. In fact, for any structure with $\alpha = (\alpha^A, \alpha^D, \alpha^{BC}, \alpha^{EF})$, we can define another structure with the same primitives except for the weights, (say) $\tilde{\alpha} = (0, \alpha^A k_1 + \alpha^D, \alpha^A k_2 + \alpha^{BC}, \alpha^A k_3 + \alpha^{EF})$. The two structures leads to the same observable $\mathcal{G}(\cdot,\cdot,\cdot,\cdot,\cdot)$.

Proposition 2: *Under Assumption C1,* α *and* $T(\cdot)$ *are identified.*

To conclude this subsection, we compare our aggregation method with the quality-adjusted quantity method used in Perrigne and Vuong (2011a). In Perrigne and Vuong (2011a), in addition to quantitative characteristics, some qualitative characteristics of the plans are observed by the analyst. Since the publisher does not use different qualities to discriminate among firms, they construct a quality-adjusted quantity index from one of the price schedules. In our model, locations of consumers are not observed by the analyst. However, for consumers with the same observed consumptions, their different payments reflects their different locations. Identification of the aggregation parameters comes from the independence between payment and the "demeaned" consumption index.

4.2.2 Identification of K, $c_0(\cdot)$, $v_0(\cdot)$, $F(\cdot)$ and $G(\cdot)$

In this subsection, we first note that further restriction on the cost function is necessary for identification. With (5) and (6), it is impossible to uncover the three functions $Q(\cdot)$, $v_0(\cdot)$ and $c_0(\cdot)$. As in Perrigne and Vuong (2011a), the marginal cost function is identified at $Q(\overline{\theta})$. While no further restriction on the cost function is needed for the identification of the other model primitives in their paper, it is not the case here. Due to its separability, our cost function is involved in (5) and (6) everywhere on the support of Q. In view of this, we restrict the cost function to be linear.

Assumption D1: $c_0(\cdot) = \gamma \times \cdot$, *where* $\gamma > 0$.

Under Assumption D1, we can show the identification of K and γ once α and $T(\cdot)$ are identified.

Lemma 1: *The parameters K and γ are identified. In particular,*

$$\gamma = T'(T^{-1}(\bar{t})),$$
$$K = t_0 - \gamma T^{-1}(t_0),$$

where $t_0 \equiv T(Q(\theta_0))$ and $\bar{t} \equiv T(Q(\overline{\theta}))$.

Now, we turn to the identification of v_0, $F(\cdot)$ and $G(\cdot)$. We use a strategy similar to Perrigne and Vuong (2011a). They exploit the known increasing one-to-one mapping between the consumer's type and consumption. Since consumers' consumption choices are observed, they will be able to recover their types. Hence, one can identify the distribution of consumer types from the distribution of quantity consumed. The identification of the utility function arises from the FOCs that the optimal price schedule should satisfy. In our case, the first order conditions (5) and (6) define a unique strictly increasing mapping from θ to Q. However, Q is not observed by the analyst. Instead we exploit the unique strictly increasing mapping between θ and T since the latter is observed. Specifically, we rewrite the necessary conditions (5) and (6) as follows.

Lemma 2: $\forall \theta_m \in [\theta_0, \overline{\theta}]$, *the necessary conditions can be rewritten as*

$$v_0(Q) = \frac{T'(Q)}{\theta_m}[1 - H(T(Q))]^{1-\frac{\gamma}{T'(q)}} \exp\left\{-\gamma \int_{Q(\theta_0)}^{Q} \frac{T''(x)}{T'(x)^2} \log[1 - H(T(x))]dx\right\}, \quad (11)$$

$$\theta(Q) = \theta_m[1 - H(T(Q))]^{\frac{\gamma}{T'(Q)}-1} \exp\left\{\gamma \int_{Q(\theta_0)}^{Q} \frac{T''(x)}{T'(x)^2} \log[1 - H(T(x))]dx\right\}, \quad (12)$$

where $H(\cdot)$ is the distribution of consumers' payments.

Notice that everything on the right-hand side of (11) and (12) are identified except θ_m. Lemma 1 suggests that any normalization of $\theta_m \in [\theta_0, \overline{\theta}]$ leads to the identification of the consumers' marginal utility function and taste distribution.

Assumption D2: $\theta_0 = 1$.

Under Assumption D2, $v_0(\cdot)$ can be interpreted as the marginal utility function for the type $(1, 1)$. By Proposition 2, for any $t \in [t_0, \overline{t}]$, Q is identified, and the marginal utility and taste can be obtained using (11) and (12). Thus, the location of a consumer with payment t and consumptions $(q^A, q^D, q^{BC}, q^{EF})$ can be identified as,

$$\epsilon = \frac{Q}{(q^A)^{\alpha^A}(q^D)^{\alpha^D}(q^{BC})^{\alpha^{BC}}(q^{EF})^{\alpha^{EF}}}. \tag{13}$$

The following proposition summarizes these results.

Proposition 3: *Under Assumptions D1 and D2, the marginal utility function $v_0(\cdot)$ and the consumers' taste distribution $F(\cdot)$ are identified on $[Q_0, \overline{Q}]$ and $[\theta_0, \overline{\theta}]$, respectively. In addition, the location distribution $G(\cdot)$ is identified on $[\underline{\epsilon}, \overline{\epsilon}]$.*

4.3 Estimation

We observe the subscribers' consumptions of A, D, $(B + C)$ and $(E + F)$ and their total bill in May 2009. We denote them as $\left\{(q_i^A, q_i^D, q_i^{BC}, q_i^{EF}), t_i)\right\}_{i=1}^N$. The term t_i is the total bill paid by subscriber i. Our semiparametric identification result of Section 4.2 leads naturally to a semiparametric procedure for estimation. We propose a multistep estimation procedure.

STEP 1: From the observed consumptions and corresponding payments, we estimate α and $T(\cdot)$. This allows us to compute $\left\{\widehat{Q}_i\right\}_{i=1}^N$ and the pseudo locations $\{\widehat{\epsilon}_i\}_{i=1}^N$ using (13).

STEP 2: In view of Lemma 2, we estimate γ and K by replacing t_0 by $t_{\min} \equiv \min_{i=1,\ldots,N} t_i$ and $\overline{t}$ by $t_{\max} \equiv \max_{i=1,\ldots,N} t_i$. This allows us to obtain an estimate for the marginal payoff $v_0(\cdot)$ and to construct a sample of pseudo tastes using (11) and (12). We then estimate the taste and location densities by using a kernel estimator.

We provide below detailed information on every step.[10]

[10]The asymptotic properties of these estimators will be collected in the supplemental web appendix.

4.3.1 Estimation of α and $T(\cdot)$

To estimate α, we partition the range of payment into $\mathcal{B}$ bins and define dummy variables $D_b(t) = 1$ if $t \in b$ and 0 otherwise, where $b \in \{1, \ldots, \mathcal{B}\}$. For any value of $b \in \{1, \ldots, \mathcal{B}\}$, we have

$$\alpha' \mathrm{E}\Big\{D_b(T)\Big[(Y - \mathrm{E}(Y|T))(Y - \mathrm{E}(Y|T))' - \sum_{b=1}^{\mathcal{B}} \mathrm{E}\left\{D_b(T)[(Y - \mathrm{E}(Y|T))(Y - \mathrm{E}(Y|T))']\right\}\Big]\Big\}\alpha = 0.$$

Thus, the estimate $\widehat{\alpha}$ is obtained by minimizing a least squares function:

$$\min_{\alpha} \sum_{b=1}^{\mathcal{B}} \left\{\alpha' \left[\frac{1}{N_b} \sum_{i:t_i \in b} (y_i - \widehat{\mathrm{E}}(y_i|t_i))(y_i - \widehat{\mathrm{E}}(y_i|t_i))' - \frac{1}{N} \sum_{i=1}^{N} (y_i - \widehat{\mathrm{E}}(y_i|t_i))(y_i - \widehat{\mathrm{E}}(y_i|t_i))'\right] \alpha\right\}^2,$$

where N_b is the number of observations in bin b, $y_i \equiv (\log(q_i^A), \log(q_i^D), \log(q_i^{BC}), \log(q_i^{EF}))'$ and $\widehat{\mathrm{E}}(y_i|t_i)$ is the conditional expection of the consumption bundle at t_i, which can be estimated using the usual nonparametric regression estimator

$$\widehat{\mathrm{E}}(y_i^l|t_i) = \frac{\sum_{k=1}^{N} y_k^l K(\frac{t_i - t_k}{h_t})}{\sum_{k=1}^{N} K(\frac{t_i - t_k}{h_t})},$$

for any $l \in \{A, D, BC, EF\}$, where $K(\cdot)$ is a symmetric kernel function with compact support and h_t is some bandwidth.

We then use the constrained smoothing estimator with regression splines proposed by Dole (1999) to approximate $T^{-1}(\cdot)$, which is strictly increasing and convex. The approximation spline proposed for strictly increasing and convex functions is

$$\psi(x; \beta, \delta) = \beta_0 + \beta_1 x + \sum_{j=0}^{n} \delta_j s_j(x),$$

where $\beta \equiv (\beta_0, \beta_1) \geq 0$, $\delta \equiv (\delta_0, \ldots, \delta_n) \geq 0$, n is the number of interior knots and $s_j(x)$ are cubic spline functions.[11]

For given $(\widehat{\alpha}, \beta, \delta)$, we can construct a pseudo sample of $e \equiv -\log(\epsilon)$

$$\widehat{e}_i(\widehat{\alpha}, \beta, \delta) = \Big(\widehat{\alpha}_A \log(q_i^A) + \widehat{\alpha}_D \log(q_i^D) + \widehat{\alpha}_{BC} \log(q_i^{BC}) + \widehat{\alpha}_{EF} \log(q_i^{EF})\Big) - \log\Big(\psi(t_i; \beta, \delta)\Big),$$

where $i \in \{1, 2, \ldots, N\}$ and N is the number of observations.

[11]See the appendix for further implementation details of the b-splines.

Our estimate of (β, δ) solves:

$$\min_{(\beta,\delta)} SSE(\widehat{\alpha}, \beta, \delta) \equiv \sum_{i=1}^{N} \left[\widehat{e}_i(\widehat{\alpha}, \beta, \delta)\right]^2, \tag{14}$$

where $\beta \geq 0$, $\delta \geq 0$. We denote the corresponding estimate as $(\widehat{\beta}, \widehat{\delta})$.

With the estimate of (β, δ), the tariff function is estimated by $\widehat{T}(\cdot) = \psi^{-1}(\cdot; \widehat{\beta}, \widehat{\delta})$. We can also calculate an estimate of Q for each payment using

$$\widehat{Q}(t) = \psi(b; \widehat{\beta}, \widehat{\delta}) = \widehat{\beta}_0 + \widehat{\beta}_1 t + \sum_{j=0}^{n} \widehat{\delta}_j s_j(t),$$

where $t \in [t_0, \overline{t}]$, and the $s_j(\cdot)$s are known cubic spline functions.

Second, with the estimate of α, we can also estimate the consumption index $q_i \equiv q(\theta_i, \epsilon_i)$ of consumer i using

$$\widehat{q}_i = (q_i^A)^{\widehat{\alpha}^A} (q_i^D)^{\widehat{\alpha}^D} (q_i^{BC})^{\widehat{\alpha}^{BC}} (q_i^{EF})^{\widehat{\alpha}^{EF}},$$

where $i \in \{1, 2, \ldots, N\}$.

Third, with the estimate of Q_i and q_i, we can estimate the location of consumer i using

$$\widehat{\epsilon}_i = \frac{\widehat{Q}_i}{(q_i^A)^{\widehat{\alpha}^A} (q_i^D)^{\widehat{\alpha}^D} (q_i^{BC})^{\widehat{\alpha}^{BC}} (q_i^{EF})^{\widehat{\alpha}^{EF}}},$$

where $\widehat{Q}_i = \psi(t_i; \widehat{\beta}, \widehat{\delta})$. Hence, a pseudo sample of locations can be constructed, $\{\widehat{\epsilon}_i\}_{i=1}^N$.

4.3.2 Estimation of $v_0(\cdot)$, $F(\cdot)$ and $G(\cdot)$

We estimate γ by $\widehat{\gamma} = \frac{1}{\psi'(t_{\max}; \widehat{\beta}, \widehat{\delta})}$, and K by $\widehat{K} = t_{\min} - \gamma\psi(t_{\min}; \widehat{\beta}, \widehat{\delta})$. We estimate $Q_{\max}$ by $Q_{\max} = \psi(t_{\max}; \widehat{\beta}, \widehat{\delta})$. We estimate $H(\cdot)$ as the empirical distribution of payment,

$$\widehat{H}(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(t_i \leq t),$$

where $\mathbb{1}(\cdot)$ is an indicator function and $t \in [0, t_{\max}]$.

The estimate for $v_0(\cdot)$ is given by

$$\widehat{v}_0(Q) = \begin{cases} \frac{\widehat{T}'(Q)}{\widehat{\xi}(Q)}, & \text{if } Q \in (0, Q_{\max}) \\ \lim_{x \uparrow Q_{\max}} \widehat{v}_0(x), & \text{if } Q \in (Q_{\max}, Q(\overline{\theta})) \end{cases}$$

and the estimate of $\theta(\cdot)$ by

$$\widehat{\theta}(Q) = \widehat{\xi}(Q),$$

where

$$\widehat{\xi}(Q) = [1 - \widehat{H}(\psi^{-1}(Q; \widehat{\beta}, \widehat{\delta}))]^{\frac{\widehat{\gamma}}{\widehat{T}'(q)} - 1} \exp\left\{\widehat{\gamma} \int_{Q(\theta_0)}^{Q} \frac{\widehat{T}''(x)}{\widehat{T}'(x)^2} \log[1 - \widehat{H}(\psi^{-1}(x; \widehat{\beta}, \widehat{\delta}))] dx\right\}.$$

In the above equations, $\psi^{-1}(\cdot; \widehat{\beta}, \widehat{\delta})$ is the inverse function of $\psi(\cdot; \widehat{\beta}, \widehat{\delta})$. $\widehat{T}'(\cdot), \widehat{T}''(\cdot)$ are the first and second derivative of $\psi^{-1}(\cdot; \widehat{\beta}, \widehat{\delta})$. Hence, for any $Q \in [Q(\theta_0), Q(\overline{\theta})]$, we can estimate the marginal utility and taste. A pseudo sample of taste can be constructed as $\left\{\widehat{\theta}_i\right\}_{i=1}^{N}$.

Finally, with the pseudo sample of taste and location, $\left\{(\widehat{\theta}_i, \widehat{\epsilon}_i)\right\}_{i=1}^{N}$, we estimate the truncated density of taste and the density of location by using kernel estimators

$$\widehat{f}^*(\theta) = \frac{1}{N h_\theta} \sum_{i=1}^{N} K\Big(\frac{\theta - \widehat{\theta}_i}{h_\theta}\Big),$$
$$\widehat{g}(\epsilon) = \frac{1}{N h_\epsilon} \sum_{i=1}^{N} K\Big(\frac{\epsilon - \widehat{\epsilon}_i}{h_\epsilon}\Big),$$

for $(\theta, \epsilon) \in [\theta_0, \overline{\theta}] \times [\underline{\epsilon}, \overline{\epsilon}]$, where $K(\cdot)$ is a symmetric kernel function with compact support, h_θ and h_ϵ are some bandwidths.

5 Empirical Results

In this section, we present the estimation results and examine several counterfactual experiments. The minimum of consumption and payment are both close to 0. Our estimate of K and $Q(\theta_0)$ are $\widehat{K} = 0$ and $\widehat{Q}(\theta_0) = 0$, respectively. Therefore, no consumer is excluded. This stems from the fact that we only consider voice consumption in this paper. The population can be understood as those who are already subscribed to this firm.

The weights in the Cobb-Douglas aggregation function are $\widehat{\alpha} = (0.426, 0, 0.381, 0.193)$. It is not surprising that $\widehat{\alpha}^D = 0$ because incoming calls are free when the subscriber is in his home city. The estimated marginal tariff function $\widehat{T}'(\cdot)$ is displayed in Figure 1-(Up-Left). We obtain $\widehat{\gamma} = 1.026$. The estimated marginal utility function is displayed in Figure 1-(Up-Right). We use a triweight kernel and a rule-of-thumb bandwidth to estimate the taste and location density functions. $\widehat{\theta}_i \in [1, 4.376]$ and $\widehat{\epsilon}_i \in [0.092, 4.242]$.[12] The estimated taste density is displayed in Figure 1-(Bottom-Left). It is a decreasing density. The estimated

[12]We use a reflection method to correct boundary effect. Since the inclusion of some unusually high values skews the estimated location density to the right severely, we also trim the upper 2.5% pseudo locations.

density reveals substantial heterogeneity among subscribers in their taste for mobile voice service. The estimated location density is displayed in Figure 1-(Bottom-Right). It is a unimodal density that is skewed right. We note that the standard deviation of the estimated taste is 0.682 while it is 1.068 for the estimated locations. Thus, subscribers are more heterogeneous in their location than taste. Omitting complete information when studying the interaction between mobile service provider and its subscribers would lead to a significant bias.

With the estimates at hand, we can perform some counterfactuals. With the estimated density functions of consumers' willingness to pay and locations, we can estimate the provider's profit when he offers a linear price schedule. That is, $T(Q) = p \times Q$. The optimal linear price would be 2.758. Figure 2 presents the allocations of Q for subscribers with different taste under the three schemes: first-best, nonlinear pricing, and linear pricing. Both allocations under nonlinear and linear pricing are less than the first-best allocation. The nonlinear pricing allocation converges to the first-best allocation because subscribers with the strongest taste are served efficiently. We then calculate the total production, the provider's profit and social welfare under three schemes, respectively. Table 4 presents the results. Nonlinear pricing achieves nearly 90% of the first-best social welfare. It is especially interesting to compare nonlinear pricing with linear pricing. Note that a necessary condition for nonlinear pricing to increase social welfare is that sufficiently more are produced. The total production increases 82% under nonlinear pricing. However, the social welfare only increases by 23%. This is due to the distribution effect of price discrimination which reduces the welfare gain. Note that a necessary condition for maximizing welfare with given output is to sell at the same price. Under nonlinear pricing, a lower average price and more service are provided to consumers with high taste. Thus, the social welfare only increases by a small amount. Nonlinear prices can increase profits and efficiency when firms cannot tell customers' willingness to pay. Due to asymmetric information, 50% of "second-best" social welfare is left "on the table" in order to induce consumers with high taste to buy large quantities.

Our second counterfactual considers product customization. In our model, product customization is costly and also affects the utility of subscribers. We now study the case in which it does not affect subscribers' utility but is still costly. That is, $v(z;\theta,\epsilon) = \theta v_0(z)$ and $c(q;\epsilon) = \widehat{\gamma} q\epsilon$. Equation (1) and (2) now become

$$\tau_1\left(q\left(\theta,\epsilon\right);\epsilon\right) = \theta v_0\left(q\left(\theta,\epsilon\right)\right),$$
$$\theta v_0\left(q\left(\theta,\epsilon\right)\right) = \gamma\epsilon + v_0\left(q\left(\theta,\epsilon\right)\right)\frac{1-F(\theta)}{f(\theta)}.$$

We note that subscribers with type (θ, ϵ) such that $\frac{\widehat{\gamma}\epsilon}{\theta - \frac{1-\widehat{F}(\theta)}{\widehat{f}(\theta)}} \geq \widehat{v}_0(0)$ will be excluded. The excluded population in the type space is displayed in Figure 3, which accounts for 20% of the subscribers. This has important implication in practice. The provider would rather exclude some subscribers if customizing service is costly but not useful to subscribers. Universal service regulation should be made if the monopolist's optimal exclusion decision greatly harms social welfare.

6 Conclusion

This paper develops a structural model to explain the observed voice consumptions and payments in mobile service industry. Two important features of the data are incorporated in the model. First, discounts are given based on the level of payment. Second, consumers may have different payments even if they consume the same amount of minutes. Our model assumes that consumers are heterogeneous along two dimensions: location and willingness to pay. The former is exogenous and common knowledge to the subscriber and the provider while the latter is known only to the subscriber. A one-to-one mapping between the willingness to pay and payment explains the first feature. Difference in locaions which is unknown to the analyst explains the second feature. Moreover, we propose a new method to aggregate multiple-dimensional variables into one-dimensional index, which does not require additional data on a variable which is a function of the index nor making parametric assumptions on the unobservable variables. Empirical results show that both observed and unobserved heterogeneity are both important. Moreover, if costly product customization does not affect subscribers' utility, 20% of subscribers would be not served by the provider.

Our results rely on the assumption that consumers know their types well, which is supported by our data. However, as noted in Miravete (2002), uncertainty about future consumption at the time of plan choice may be significant. Examples include cases where there is a switching cost or a long term contract. A natural extension of our model is to consider the case in which consumers have uncertainty when choosing plans. A two-stage decision process model can be entertained. Another extension of interest is to consider both voice and short message service (SMS) consumption. Relying on Armstrong (1996), Luo, Perrigne, and Vuong (2011) generalize the methodology in Perrigne and Vuong (2011a) and the techniques developed here to the multiproduct case and apply them to the empirical analysis of voice and SMS in the mobile phone industry.

Appendix A: Proofs

This appendix gives the proofs of the propositions and lemmas stated in Section 4.

Proof of Proposition 1: The proof follows Sundararajan (2004).

□

Proof of Proposition 2: We observe the joint distribution of $(Y^A, Y^D, Y^{BC}, Y^{EF}, T)$. We denote $\alpha \equiv (\alpha^A, \alpha^D, \alpha^{BC}, \alpha^{EF})'$, $\Lambda(\cdot) \equiv \log(T^{-1}(\cdot))$, $e \equiv -\log(\epsilon)$ and $Y^l(T,e) \equiv \log(q^l(T,e))$ where $l \in \{A, D, BC, EF\}$. We want to identify α, $\Lambda(\cdot)$ and the distribution of ϵ, $G(\cdot)$. Our model does not provide any information about $Y^l(\cdot)$ except the following relationship:

$$\sum_{l\in\{A,D,BC,EF\}} \alpha^l Y^l(T,e) = \Lambda(T) + e,$$

where $\Lambda(\cdot)$ is strictly increasing, $\alpha^l \geq 0$ and $\sum_{l\in\{A,D,BC,EF\}} \alpha^l = 1$.

In the first step, we prove that α is identified. Our identification result of α is based on the following observations:

(i) For any given payment $t \in [0, \bar{t}]$, the variability in $\sum_{l\in\{A,D,BC,EF\}} \alpha^l Y^l(t,e)$ equals the variation in e.

(ii) $\forall t \in [0, \bar{t}], E[e^2|T = t] = \sigma_e^2$.

which imply that

$$\text{Var}\ \{\alpha' Y - \Lambda(T)|T = t\} = \sigma_e^2. \tag{15}$$

To prove the identification of α, we need to show that there exists a unique α such that (15) is satisfied. To see this, note that $\mathrm{E}\left[\alpha' Y|T = t\right] = \Lambda(t)$ implies that (15) can be rewriten as

$$\mathrm{E}\left\{\Big[\sum_{l\in\{A,D,BC,EF\}} \alpha^l Y^l(t,e) - \Lambda(t)\Big]^2\right\} = \sigma_e^2.$$

where the expectation is taken on e.

Taking partial derivative with respective to t and using Leibniz's rule gives

$$\int e\Big[\sum_{l\in\{A,D,BC,EF\}} \alpha^l \frac{\partial Y^l}{\partial t} - \Lambda'(t)\Big] g_e(e)de = 0,$$

where $g_e(\cdot)$ is the density function of e. $\mathrm{E}\left[e|T\right] = 0$ implies $\mathrm{E}\left[e\Lambda'(t)\right] = 0$. Thus

$$\int e\Big(\sum_{l\in\{A,D,BC,EF\}} \alpha^l \frac{\partial Y^l}{\partial t}\Big) g_e(e)de = 0,$$

which can be rewritten as

$$\sum_{l\in\{A,D,BC,EF\}} \alpha^l \mathrm{E}\left[e\frac{\partial Y^l}{\partial t}\right] = 0, \tag{16}$$

where the expectation is taken on e.

Supplemented by $\sum_{l \in \{A,D,BC,EF\}} \alpha^l = 1$, (16) evaluated at t_1, t_2 and t_3 provides four linear equations in the four unknowns $(\alpha^A, \alpha^D, \alpha^{BC}, \alpha^{EF})$. By C3, these equations can be solved for an unique solution. Hence, α is identified.

In the second step, we prove that $\Lambda(\cdot)$ is also identified. Let $\mathcal{Y} \equiv \sum_{l \in \{A,D,BC,EF\}} \alpha^l Y^l$ and $\mathrm{E}[\mathcal{Y}|T = t; \Lambda(\cdot), G(\cdot)]$ denote the conditional expectation of $\mathcal{Y}$ given $T = t$, for the distribution generated by $(\Lambda(\cdot), G(\cdot))$. Suppose that $(\widetilde{\Lambda}(\cdot), \widetilde{G}(\cdot))$ and $(\widehat{\Lambda}(\cdot), \widehat{G}(\cdot))$ generate the same joint distribution of $(Y^A, Y^D, Y^{BC}, Y^{EF}, T)$, and $\widetilde{\Lambda}(t^{\#}) \neq \widehat{\Lambda}(t^{\#})$. Because α is identified, they generate the same joint distribution of $(\mathcal{Y}, T)$. Thus, since

$$E[\mathcal{Y}|T = t^{\#}; \widetilde{\Lambda}(\cdot), \widetilde{G}(\cdot)] = \widetilde{\Lambda}(t^{\#})$$
$$E[\mathcal{Y}|T = t^{\#}; \widehat{\Lambda}(\cdot), \widehat{G}(\cdot)] = \widehat{\Lambda}(t^{\#})$$

and both functions are continuous at $t^{\#}$, it follows that

$$F_{\mathcal{Y}|T}(\cdot; \widetilde{\Lambda}(\cdot), \widetilde{G}(\cdot)) \neq F_{\mathcal{Y}|T}(\cdot; \widehat{\Lambda}(\cdot), \widehat{G}(\cdot)).$$

which contradicts the fact that the two structures generate the same joint distrubtion of $(\mathcal{Y}, T)$. Hence, $\Lambda(\cdot)$ is identified. Since $T(\cdot)$ is a known functional of $\Lambda(\cdot)$, it is also identified.

□

Proof of Lemma 1: On one hand, $Q(\overline{\theta})$ is identified because

$$T(Q(\overline{\theta})) = \overline{t},$$

where $T(\cdot)$ is identified by Proposition 2 and $\overline{t}$ is observed.

On the other hand, evaluating (5) and (6) at $\theta = \overline{\theta}$ gives

$$T'(Q(\overline{\theta})) = \gamma.$$

where $T'(\cdot)$ is identified given that $T(\cdot)$ is.

Therefore, γ is identified as $T'(T^{-1}(\overline{t}))$. The identification of K follows from $t_0 = T(Q(\theta_0)) = K + \gamma Q(\theta_0)$.

□

Proof of Lemma 2: In the first step, we consider the infeasible case in which Q is observed. Therefore $T(\cdot)$ is also observed. Denote the distribution of Q as $G_Q(\cdot)$. Following the same lines as in Perrigne and Vuong (2011a), we can express the marginal utility function $v_0(\cdot)$ and $\theta(\cdot)$ as functions of $T(\cdot)$, γ and $G_Q(\cdot)$. For any value $\theta_m \in [\theta_0, \overline{\theta}]$, the necessary conditions can be rewritten as

$$v_0(Q) = \frac{T'(Q)}{\theta_m}[1 - G_Q(Q)]^{1-\frac{\gamma}{T'(q)}} \exp\left\{-\gamma \int_{Q(\theta_0)}^{Q} \frac{T''(x)}{T'(x)^2} \log[1 - G_Q(x)]dx\right\}, \tag{17}$$

$$\theta(Q) = \theta_m[1 - G_Q(Q)]^{\frac{\gamma}{T'(Q)}-1} \exp\left\{\gamma \int_{Q(\theta_0)}^{Q} \frac{T''(x)}{T'(x)^2} \log[1 - G_Q(x)]dx\right\}. \tag{18}$$

In the second step, we consider the case in which Q is not observed. We note that $\forall Q_m \in [Q(\theta_0), Q(\overline{\theta})]$,

$$G_Q(Q_m) = \Pr\{Q \leq Q_m\} = \Pr\{T \leq T(Q_m)\} \equiv H(T(Q_m)).$$

Therefore, replacing $G_Q(\cdot)$ with $H(T(\cdot))$ in (17) and (18) yields (11) and (12).

□

Proof of Proposition 3: When $\Lambda(\cdot)$ is identified, we can also identify $G(\cdot)$ because

$$G(z) = \Pr\{\log(\epsilon) \leq \log(z)\} = \Pr\{e \geq -\log(z)\} = \Pr\{\mathcal{Y} - \Lambda(T) \geq -\log(z)\}.$$

where the joint distribution of $(\mathcal{Y}, T)$ is identified.

Regarding the identification of $v_0(\cdot)$ and $F(\cdot)$, we follow Perrigne and Vuong (2011a).

□

Appendix B: Estimation of α and $T(\cdot)$

This appendix describes how we estimate α and the implementation details of the b-splines to estimate $T(\cdot)$.

To estimate α, we first use a triweight kernel and rule-of-thumb bandwidths to estimate $\widehat{E}[\log(q_i^l)|t_i]$, where $l \in \{A, D, BC, EF\}$. Second, we partition $[0, t_{\max}]$ into 50 bins using empirical quantiles $\left\{\frac{b}{50}\right\}_{b=1}^{50}$ as cut points. We trim several bins on the boundaries because a kernel is used to estimate $\widehat{E}[\log(q_i^l)|t_i]$. The estimate of α minimizes the "adjusted" objective function:

$$\min_{\alpha} \sum_{b=3}^{50-4} \left\{ \alpha' \left[\frac{1}{N_b} \sum_{i:t_i \in b} (y_i - \widehat{E}(y_i|t_i))(y_i - \widehat{E}(y_i|t_i))' - \frac{N}{(N^*)^2} \sum_{i=1}^{N^*} (y_i - \widehat{E}(y_i|t_i))(y_i - \widehat{E}(y_i|t_i))' \right] \alpha \right\}^2,$$

where N^* is the total number of observations in bins included.

The adjustement is necessary because σ_e^2 is estimated differently after trimming. In particular, note that $\sigma_e^2 P_b = \mathrm{E}\,\{D_b(T)[\alpha'(Y - \mathrm{E}\,(Y|T))(Y - \mathrm{E}\,(Y|T))'\alpha]\}$ implies that

$$\begin{aligned}
\sigma_e^2 &= \Big(\sum_{b=3}^{50-4} \mathrm{E}\,\{D_b(T)[\alpha'(Y - \mathrm{E}\,(Y|T))(Y - \mathrm{E}\,(Y|T))'\alpha]\} \Big) / \Big(\sum_{b=3}^{50-4} P_b \Big) \\
&= \sum_{b=1}^{50} \mathrm{E}\,\{D_b(T)[\alpha'(Y - \mathrm{E}\,(Y|T))(Y - \mathrm{E}\,(Y|T))'\alpha]\}.
\end{aligned}$$

To estimate $T(\cdot)$, we approximate its inverse function with splines and find the optimal approximate spline that solves (14). Note that $T^{-1}(\cdot)$ is increasing and convex. We use the constrained smoothing with regression splines proposed in Dole (1999) to approximate it:

$$\psi(x; \beta, \delta) = \beta_0 + \beta_1 x + \sum_{j=0}^{n} \delta_j s_j(x),$$

where $\beta \equiv (\beta_0, \beta_1) \geq 0$, $\delta \equiv (\delta_0, \ldots, \delta_n) \geq 0$ and n is the number of interior knots.

The term $s_0(x)$ is defined as

$$s_0(x) = \begin{cases}
0 & \text{if} \quad x \in (-\infty, k_{(1)}] \\
\left(x - k_{(1)}\right)^3 / 6 \left(k_{(2)} - k_{(1)}\right) & \text{if} \quad x \in \left[k_{(1)}, k_{(2)}\right] \\
x^2/2 - x\left(k_{(1)} + k_{(2)}\right)/2 + \left(k_{(2)} - k_{(1)}\right)^2/6 + k_{(1)}k_{(2)}/2 & \text{if} \quad x \in \left[k_{(2)}, k_{(N-1)}\right] \\
\left(x - k_{(N)}\right)^3 / 6 \left(k_{(N-1)} - k_{(N)}\right) + b_{10}x + b_{00} & \text{if} \quad x \in \left[k_{(N-1)}, k_{(N)}\right] \\
b_{10}x + b_{00} & \text{if} \quad x \in [k_{(N)}, \infty)
\end{cases}$$

where

$$\begin{aligned}
b_{10} &= \left(k_{(N)} + k_{(N-1)} - k_{(2)} - k_{(1)}\right)/2 \\
b_{00} &= \left(\left(k_{(2)} - k_{(1)}\right)^2 - \left(k_{(N-1)} - k_{(N)}\right)^2 + 3k_{(2)}k_{(1)} - 3k_{(N)}k_{(N-1)}\right)/6,
\end{aligned}$$

with $k_{(j)}$ denoting the j'th order statistic of the observed bills ($j = 1, 2, N-1, N$).

For $1 \leq j \leq n$, $s_j(x)$ is defined as

$$s_j(x) = \begin{cases} 0 & \text{if} \quad x \in (-\infty, k_{j-1}] \\ (x - k_{j-1})^3 / 6(k_j - k_{i-1}) & \text{if} \quad x \in [k_{j-1}, k_j] \\ (x - k_{j+1})^3 / 6(k_j - k_{j+1}) + b_1 x + b_0 & \text{if} \quad x \in [k_j, k_{j+1}] \\ b_1 x + b_0 & \text{if} \quad x \in [k_{j+1}, \infty) \end{cases}$$

where

$$\begin{aligned} b_1 &= (k_{j+1} - k_{j-1})/2 \\ b_0 &= \left((k_j - k_{j-1})^2 - (k_j - k_{j+1})^2 + 3(k_{j+1} - k_{j-1}) k_j\right)/6, \end{aligned}$$

when $j - 1 = 0$, we set $k_{j-1} = 0$, and when $j + 1 > n$, we set $k_{j+1} = 1000$.

We remark that $s_0(x)$ is the double integral of the following B-spline

$$B_0(x) = \begin{cases} \left(x - k_{(1)}\right) / \left(k_{(1)} - k_{(2)}\right) & \text{if} \quad x \in \left[k_{(1)}, k_{(2)}\right] \\ 1 & \text{if} \quad x \in \left[k_{(2)}, k_{(N-1)}\right] \\ \left(x - k_{(N)}\right) / \left(k_{(N)} - k_{(N-1)}\right) & \text{if} \quad x \in \left[k_{(N-1)}, k_{(N)}\right] \\ 0 & \quad \textit{otherwise} \end{cases}$$

and $s_j(x)$ is the double integral of the following B-spline

$$B_j(x) = \begin{cases} (x - k_{j-1}) / (k_j - k_{j-1}) & \text{if} \quad x \in [k_{j-1}, k_j] \\ (x - k_{j+1}) / (k_j - k_{j+1}) & \text{if} \quad x \in [k_j, k_{j+1}] \\ 0 & \quad \textit{otherwise} \end{cases}$$

Appendix C: Graphs

Figure 1: Estimation Results

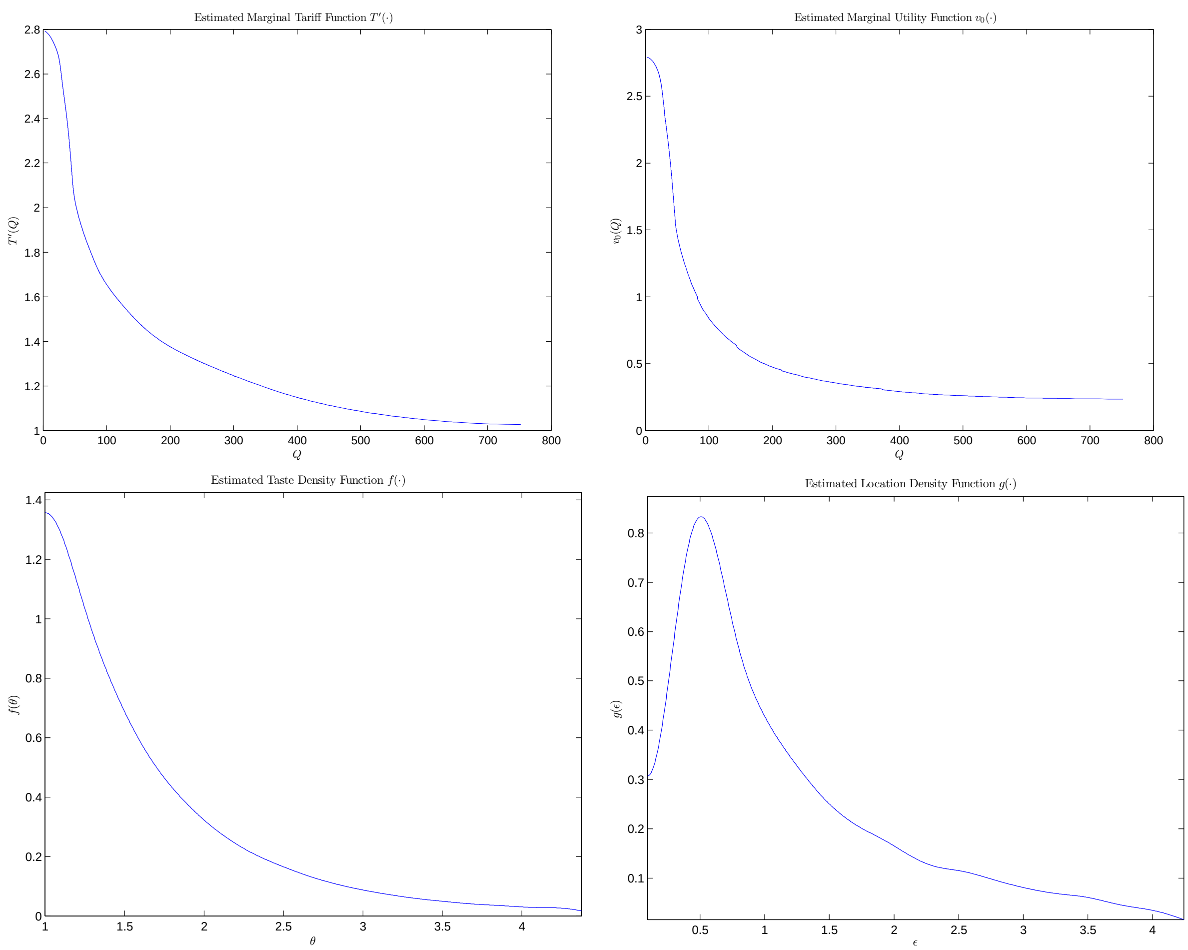


Figure 2: Q under First-best, Nonlinear and Linear Pricing

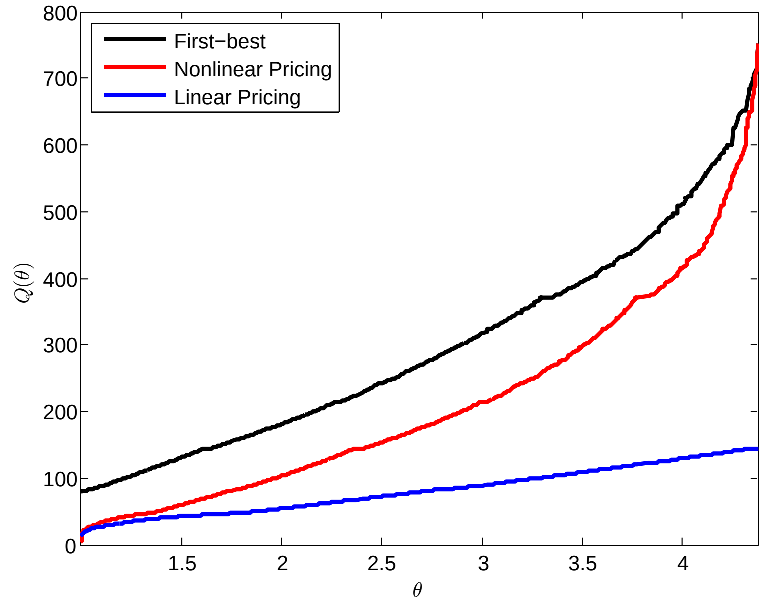
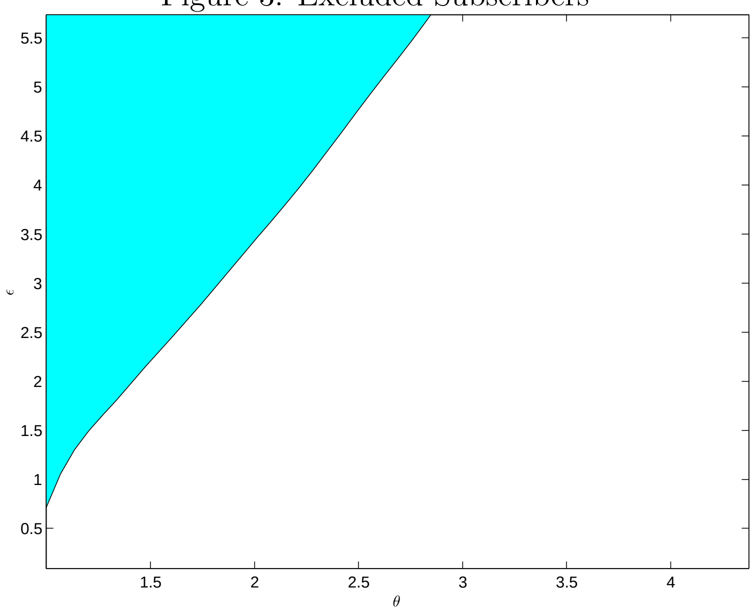


Figure 3: Excluded Subscribers



Appendix D: Tables

Table 1: Different Kinds of Calls

	When the subscriber is		
	at local	in home province	out of home province
Outgoing Calls	A	B	C
Incoming Calls	D	E	F

Table 2: Summary Statistics[a]

Variable	Mean	Std. Dev.	Min	Max
Bill	182.49	125.07	3	997.30
A	518.87	431.88	0	4613
D	473.44	429.41	0	9713
BC	114.31	230.37	0	9923
EF	72.34	162.91	0	3071
Days since Subscription[b]	845.38	825.18	-22	6553

[a] The number of observations $N > 20,000$. We utilize a random sample of 10,000 observations in our estimation.

[b] The number of days between the date when the consumer first subscribed to the provider and April 30, 2009. The minimum is negative because some consumers subscribed after April 30, 2009.

Table 3: Product Customization

$Bill \in$	Variable	# Observations	Mean	Std. Dev.	Min	Max
(175,225]	TotalMin	$0.14 \times N$	924.88	274.44	17	4103
	Bill	$0.14 \times N$	197.29	13.81	175.02	225
(275,325]	TotalMin	$0.06 \times N$	1356.49	399.28	41	3698
	Bill	$0.06 \times N$	297.38	13.55	275.02	325
(375,425]	TotalMin	$0.02 \times N$	1817.58	603.19	24	5583
	Bill	$0.02 \times N$	397.47	13.85	375.04	425
(475,525]	TotalMin	$0.01 \times N$	2135.45	681.21	70	4878
	Bill	$0.01 \times N$	498.81	14.13	475.05	524.9

Table 4: Counterfactual Experiments

	First-best	Nonlinear Pricing	Linear Pricing
Total Production	155.27	84.90	46.66
Firm Profit	214.44	93.77	80.80
Social Welfare	214.44	191.33	155.12

References

Aguiar, M. and E. Hurst (2005): "Consumption Versus Expenditure," *Journal of Political Economy*, 113, 919–948.

Armstrong, M. (1996): "Multiproduct Nonlinear Pricing," *Econometrica*, 64, 51–75.

Aryal, G., I. Perrigne, and Q. Vuong (2009): "Identification of Insurance Models with Multidimensional Screening," Working Paper, Pennsylvania State University.

Bernhardt, D., Q. Liu, and K. Serfes (2007): "Product Customization," *European Economic Review*, 51, 1396–1422.

Chen, Y. and G. Iyer (2002): "Consumer Addressability and Customized Pricing," *Marketing Science*, 21, 197–208.

Cohen, A. (2008): "Package Size and Price Discrimination in the Paper Towel Market," *International Journal of Industrial Organization*, 26, 502–516.

Cohen, A. and L. Einav (2007): "Estimating Risk Preferences From Deductible Choice," *The American Economic Review*, 97, 745–788.

Da Silveira, G., D. Borenstein, and F. Fogliatto (2001): "Mass Customization: Literature Review and Research Directions," *International Journal of Production Economics*, 72, 1–13.

Debreu, G. (1959): *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*, Yale University Press.

Dewan, R., B. Jing, and A. Seidmann (2000): "Adoption of Internet-Based Product Customization and Pricing Strategies," *Journal of Management Information Systems*, 17, 9–28.

——— (2003): "Product Customization and Price Competition On the Internet," *Management Science*, 49, 1055–1070.

Dole, D. (1999): "Cosmo: A Constrained Scatterplot Smoother for Estimating Convex, Monotonic Transformations," *Journal of Business & Economic Statistics*, 17, 444–455.

Dunn, K. and K. Singleton (1986): "Modelling the Term Structure of Interest Rates Under Habit Formation and Durability of Goods," *Journal of Financial Economics*, 17, 27–55.

Economides, N., K. Seim, and V. Viard (2008): "Quantifying the Benefits of Entry Into Local Phone Service," *The RAND Journal of Economics*, 39, 699–730.

Finkelstein, A. and K. McGarry (2006): "Multiple Dimensions of Private Information: Evidence From the Long-Term Care Insurance Market," *American Economic Review*, 96, 938–958.

Gruber, H. (2005): *The Economics of Mobile Telecommunications*, Cambridge University Press.

Guerre, E., I. Perrigne, and Q. Vuong (2000): "Optimal Nonparametric Estimation of First-Price Auctions," *Econometrica*, 68, 525–574.

Heckman, J., R. Matzkin, and L. Nesheim (2010): "Nonparametric Identification and Estimation of Nonadditive Hedonic Models," *Econometrica*, 78, 1569–1591.

Kim, Y., R. Telang, W. Vogt, and R. Krishnan (2010): "An Empirical Analysis of Mobile Voice Service and SMS: A Structural Model," *Management Science*, 56, 234–252.

Leslie, P. (2004): "Price Discrimination in Broadway Theater," *RAND Journal of Economics*, 35, 520–541.

Lewis, T. and D. Sappington (1989): "Countervailing Incentives in Agency Problems," *Journal of Economic Theory*, 49, 294–313.

Luo, Y., I. Perrigne, and Q. Vuong (2011): "Multiproduct Nonlinear Pricing: Mobile Voice Service and SMS," Working Paper, Pennsylvania State University.

Maggi, G. and A. Rodriguez-Clare (1995): "On Countervailing Incentives," *Journal of Economic Theory*, 66, 238–263.

Maskin, E. and J. Riley (1984): "Monopoly with Incomplete Information," *The RAND Journal of Economics*, 171–196.

McManus, B. (2007): "Nonlinear Pricing in an Oligopoly Market: The Case of Specialty Coffee," *The RAND Journal of Economics*, 38, 512–532.

Miravete, E. (2002): "Estimating Demand for Local Telephone Service with Asymmetric Information and Optional Calling Plans," *Review of Economic Studies*, 69, 943–971.

——— (2005): "The Welfare Performance of Sequential Pricing Mechanisms," *International Economic Review*, 46, 1321–1360.

Miravete, E. and L. Röller (2004): "Competitive Nonlinear Pricing in Duopoly Equilibrium: The Early US Cellular Telephone Industry," Working Paper, University of Texas at Austin.

Murphy, A. (2007): "A Dynamic Model of Housing Supply," Working Paper, Washington University in St. Louis.

Perrigne, I. and Q. Vuong (2011a): "Nonlinear Pricing in Yellow Pages," Working Paper, Pennsylvania State University.

——— (2011b): "Nonparametric Identification of Contract Models With Adverse Selection and Moral Hazard," *Econometrica*, 79, 1499–1539.

Sankar, U. (1970): "Elasticities of Substitution and Returns to Scale in Indian Manufacturing Industries," *International Economic Review*, 11, 399–411.

Seim, K. and V. Viard (2010): "The Effect of Entry and Market Structure On Cellular Pricing Tactics," *American Economic Journal: Microeconomics*, forthcoming.

Solow, R. (1957): "Technical Change and the Aggregate Production Function," *The Review of Economics and Statistics*, 39, 312–320.

Sundararajan, A. (2004): "Nonlinear Pricing of Information Goods," *Management Science*, 50, 1660–1673.

Thisse, J. and X. Vives (1988): "On the Strategic Choice of Spatial Price Policy," *The American Economic Review*, 78, 122–137.