A Claryfying Note on Methodological Issues Arising in the Analysis of Online Markets"

Elena Krasnokutskaya* Johns Hopkins University

Kyungchul Song† University of British Columbia

Xun Tang‡ Rice University

April 25, 2017

Here we answer the questions about the methodology in "The Role of Quality in Internet Service Markets" that we have encountered numerous times when presenting this paper. The explanations we provide are more detailed than what we were able to include in the paper.

1 Question: Why not fixed effects for permanent sellers? Why do you have to use the classification algorithm?

Let us begin by saying that we do model the permanent sellers' qualities as fixed effects. However, these fixed effects are *not at the seller's level but at the group level*. Since the groups are unobserved in the data, we use the classification procedure to link individual sellers to their quality groups (and thus, their fixed effects). To illustrate the usefulness of our approach, we will organize our reply to your question in two steps. First, we imagine attempting to estimate the model that includes fixed effects at the individual permanent seller level. As the discussion will illustrate, this approach faces a number of challenges. In the second step, we explain how the classification algorithm (together with our insight on linking the supports of permanent and transitory bidders) simultaneously overcomes all of these challenges.

Estimating seller-level fixed effects. Let us begin by simplifying the environment and assuming that only permanent sellers are present in the market. Even in this case, at least two difficulties arise: the market shares of individual sellers can be very small and very few auctions feature the same set of individual bidders.

To appreciate the first difficulty, note that in the market we study, the sellers choose to participate in a specific auction, and at the time of this decision they do not observe either the identities or the prices of other sellers contacting the same buyer. As a result, even if the probability of a winning for a given seller conditional on the choice set could be estimated, such probability quite often would be very small. This is problematic both for the MLE and GMM

*Email: ekrasno1@jhu.edu.
†Email: kysong@mail.ubc.ca.
‡Email: xun.tang@rice.edu.

methods because their accuracy suffers when the choice probabilities are close to zero.[1] In the differentiated products setting similar issues are often resolved by aggregating the products to the level where market shares become of reasonable size (for example, by treating as one product all the models from the same car family within a producer, etc.). There is no obvious way to similarly aggregate individual sellers in the market we study – so that the members of the group are roughly interchangeable from the buyer's point of view – because they are not related to each other in any way.

The second difficulty is rooted in the fact that in the data such as ours, there is a very large number of potential sellers and only small subset of these sellers choose to participate in any given auction. Thus, the data contain a very large number of distinct choice sets. In contrast, the number of projects in the data is comparable to the number of possible choice sets. This gives rise to a problem in the context of GMM estimation because the probability of choice conditional on choice set – the basis for GMM estimation in the analysis of differentiated products markets – cannot be precisely estimated (once again, because only an extremely small number of buyers face precisely the same set of sellers in our data). Note that, in particular, this rules out a direct application of the GMM methodology utilizing the moments proposed in Berry, Levinsohn, and Pakes (1995) (referred to as BLP below).

These problems may possibly be alleviated by aggregating over the choice sets which include a given seller:

$$\Pr(i \text{ wins}|B_l = b, i \in A_l) = \sum_{\{a: i \in a\}} \Pr(i \text{ wins}|B_l = b, A_l = a) \Pr(A_l = a|i \in A_l), \tag{1}$$

where B_l denotes the bid vector in auction l and A_l the set of participants (i.e., the buyer's choice set) in auction l.

However, it is not obvious that such moments could be used to identify seller fixed effects. Specifically, the invertibility argument underlying BLP methodology does not apply to these moments because the probability of observing a given choice set, $\Pr(A_l = a|i \in A_l)$, depends in the model on the qualities of potential sellers. This is not to say that this system is not invertible, it may well be.[2] The only point that we would like to highlight is that pursuing this strategy would involve modifying the standard arguments in a way that makes it applicable to a market such as ours. Also, it would be necessary to establish that the identification of the distribution of random coefficients (buyers' tastes) would be possible under such full aggregation.[3] Further, even if this mechanism would work in theory it is not at all certain that it would perform well in practice given that the weight probabilities used in aggregation above ($\Pr(A_l = a|j \in A_l)$) are very small. It might be preferable to consider an entirely different basis for aggregation which

[1] It is likely that for a given seller the probability of winning under some choice sets would be of a reasonable magnitude. One might imagine that we could choose a choice set judiciously and in this way overcome the problem we describe here. Note, however, that such choice sets would have to be found for each of the 300-500 permanent sellers in our setting. In addition, for each choice set that we deem acceptable, every seller in this set would have to have a sufficiently high probability of winning. Choosing such choice sets can be a challenging problem. This discussion, of course, assumes that the probability of winning conditional on the choice set can be estimated from the data which is not true in our market as we discuss next.

[2] For example, it is possible that the inversion could be made to work if we use empirical probabilities of observing different choice sets in the expression on the right-hand side.

[3] Specifically, we would loose variation in choice sets which is typically used for identification of the distribution of buyers' tastes. It is possible that the variation in prices could be exploited instead but it is not clear whether it would be sufficient to identify multi-dimensional buyers' tastes (represented by α_l and β_l).

would maximize the performance of the estimator given the available data structure.

However, the foregoing discussion is purely hypothetical if one is interested in the analysis of online markets or markets for services more generally, because these markets tend to have a large share of transitory sellers. The studies of differentiated products tend to drop unsuccessful products from consideration. This is feasible in their context, because all unsuccessful products combined account for a very small market share and constitute a tiny part of any buyer's choice set. However, in our market, often more than half of the choices in any buyers' choice set consists of transitory sellers and the probability that any given buyer will choose a transitory seller is quite large (0.38).

Moreover, the online platform we study creates many opportunities for buyers to learn about the sellers, either transitory or permanent, through direct communication. This allows buyers to take quality of each seller into account when making the choice. In contrast, the researcher does not observe transitory sellers qualities. Thus, the unobserved heterogeneity of transitory sellers, which leads to endogeneity of their prices, has to be taken into account, and yet it cannot be captured by fixed effects since we have a very small number of observations per transitory seller within the context of a non-linear model. Transitory sellers' qualities have to be modeled as random effects correlated with prices. In short, transitory sellers introduce mixture components into the estimation.

The estimation routine based on the system of equation in (1), which contains an equation for each of the permanent sellers, cannot be implemented using the set of auctions where only permanent sellers participate because such auctions constitute less than one tenth of a percent in our data. Instead, permanent sellers' individual fixed effects would have to be estimated jointly with the buyers' tastes and the mixture components associated with transitory sellers.

To begin with, it is not at all obvious that mixture components and therefore everything else could be identified through such procedure since it is well known that identification of such models is very challenging under the best of circumstances. Setting concerns about identification aside, the methodologies for the estimation of mixture models available to us include (1) solving for the relationship between the prices and the transitory sellers' qualities from the model and using this relationship to integrate out the transitory sellers' qualities within the estimation routine; (2) assuming ad hoc distribution of transitory sellers' qualities conditional on their bids and estimating this distribution jointly with other primitives of the model.[4] The first approach is computationally infeasible, because due to the non-standard nature of the auction mechanism, it takes long time to solve for the bidding functions even for a single set of participants, and because, with the seller heterogeneity defined at the level of seller identity, one needs to solve the model for a very large number of choice sets. The second approach has been previously shown to produce highly unreliable estimates.[5] Thus, a novel approach to estimating the mixture model will have to be developed to make progress. We describe how our methodology addresses these challenges next.

Our approach. Our approach is based on the theoretical result which establishes that permanent sellers' group membership can be recovered (identified) from the pairwise comparisons of sellers' performance. These pairwise tests can be implemented nonparametrically without

[4]A researcher may also consider an approach proposed by Kasahara and Shimotsu (2009) in the context of a dynamic discrete choice model. However, the model considered by these authors does not readily map into our environment so the applicability of this method, if possible, is far from obvious.

[5]See Heckman and Singer (1984) for details.

specific knowledge of the distribution of buyers' tastes or the distribution of transitory sellers' qualities conditional on price. The proof is constructive and relies on the data structure typical for online markets (this procedure requires that sufficiently many auctions where one of the two potential bidders is present but the other is not should be available – such requirement is satisfied by our data). The fact, that the informational structure underlying our theoretical identification strategy is a close match to the data, makes it more likely that the estimator based on our constructive identification proof would have good empirical properties. Indeed, in a companion paper Krasnokutskaya, Song, and Tang (2016) we conduct an extensive Monte Carlo evaluation and establish this to be the case.

The classification step, which recovers the number of distinct quality levels and assigns each permanent seller to a particular quality group, considerably simplifies the estimation since after this step only a relatively small number of (group-specific) fixed effects has to be estimated. Moreover, using groups rather than identities, we achieve partial aggregation of observations, which (1) creates aggregated products; (2) allows us to introduce an aggregated concept of the choice set. Such aggregation allows us to overcome the problem of small probabilities of choice and (to a degree) the problem associated with the large number of choice sets. In practice, however, we estimate that there are 22 groups, which is still too large a number for the methodology utilizing the moments which condition on the choice set.[6] This leads us to propose a novel set of moments (based on our ability to group sellers in the classification step) which allows us to recover all of the model primitives.

The classification step also provides a foundation for recovering the model primitives in the presence of transitory sellers. Our approach for doing so relies on the second key insight. Specifically, we show that estimation of the mixture model simplifies dramatically if we tie the support of the distribution of the transitory sellers' qualities to that of the distribution of the permanent sellers' qualities. Indeed, conditioning on sellers' fixed effects eliminates the endogeneity of permanent sellers' prices and allows us to use exogenous variation in these prices to identify all of the other primitives of the model, including the frequencies of different quality levels in the population of transitory sellers conditional on price (the levels are recovered as the permanent sellers' fixed effects). We formally establish identification of the model primitives following this line of thought. However, such approach would not be effective in practice if the qualities of permanent sellers were estimated at the seller level. In such case, we would end up with too many mixture components and prohibitively many parameters to estimate (i.e., each mixture component would have two or three parameters and this number would have to be multiplied by the number of permanent sellers). Defining the permanent sellers' heterogeneity at the group level makes this approach feasible.

Thus, the estimation strategy that we develop exhibits a number of desirable properties. It is based on a formal identification proof exploiting the data structure which corresponds closely to the data available for our analysis. The constructive nature of the proof leads directly to the implementation strategy. Moreover, the resulting estimator is quite efficient computationally.

While we do not wish to argue that it is not possible to construct a different estimator sharing the same desirable features, there is no existing of-the-shelf estimation strategy that one can use to study the data collected in the online markets. Possible alternative strategies will have to solve differently the same set of problems that our strategy overcomes.

[6]For example, even with as few as 5 bidders per auction, there are $22^5 = 5,153,632$ distinct choice sets, much larger than the number of auctions in our data.

2 Other Typical Questions

(a) *Where is the problems in the model without transitory sellers? Why BLP cannot be used? There is a finite number of seller types (permanent/transitory with different qualities) even if the discrete qualities, q_k, are unobserved by the researcher.*

(b) *Does the problem arises because of endogeneity of sellers' participation decisions?*

(a) The standard BLP methodology is explicitly characterized by the moments which are based on the probability of choice conditional on the choice set in the context of product-specific fixed effects. As we explain above, such set of moments cannot be used given the structure of our data even if the environment does not have transitory sellers. In our data we see at most five observations characterized by the same set of *permanent* bidders within buyer's choice set. As we mentioned before, in the world without transitory sellers, it might be possible to design an alternative GMM procedure that would be based on the aggregated moments. However, even in that case there are difficulties to overcome and the success is far from certain.

(b) The methodology we propose is likely to be useful even in the context of a simple model with exogenous participation. Indeed, the fact that there is a small number of types is not helpful in estimation unless we can link a seller to his type a-priori. Without the classification step we have to estimate a fixed effect for every seller. Then, if the number of types is small that there will be a small number of distinct clusters among the estimated fixed effects but such regularity would only be revealed post-estimation. Thus, even if participation is exogenous, as long as participation varies and the number of potential sellers is large, the number of possible choice sets is likely to be large, the probability to observe any given choice set would be small. Most importantly, if transitory sellers are present, then even in the context of a model with exogenous participation we have to account for the endogeneity of transitory sellers' prices, i.e. we need to estimate a mixture model. If this is the case, then our methodology offers a viable path forward.

References

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica*, 63, 841–890.

Heckman, J., and B. Singer (1984): "A Method of Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52, 271–320.

Kasahara, H., and K. Shimotsu (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77(1), 135–175.

Krasnokutskaya, E., K. Song, and X. Tang (2016): "Estimating Unobserved Agent Heterogeneity Using Pairwise Comparisons," Working paper, Johns Hopkins University.