Hu, Yingyao; Shum, Matthew

Working Paper

Estimating first-price auctions with an unknown number of bidders: A misclassication approach

Working Paper, No. 541

Provided in Cooperation with:
Department of Economics, The Johns Hopkins University

Suggested Citation: Hu, Yingyao; Shum, Matthew (2007) : Estimating first-price auctions with an unknown number of bidders: A misclassication approach, Working Paper, No. 541, The Johns Hopkins University, Department of Economics, Baltimore, MD

This Version is available at:
https://hdl.handle.net/10419/49868

Estimating First-Price Auctions with an Unknown Number of Bidders: A Misclassification Approach *

Yingyao Hu and Matthew Shum
Dept. of Economics, Johns Hopkins University

December 19, 2007

Abstract

In this paper, we consider nonparametric identification and estimation of first-price auction models when N^*, the number of potential bidders, is unknown to the researcher, but observed by bidders. Exploiting results from the recent econometric literature on models with misclassification error, we develop a nonparametric procedure for recovering the distribution of bids conditional on the unknown N^*. Monte Carlo results illustrate that the procedure works well in practice. We present illustrative evidence from a dataset of procurement auctions, which shows that accounting for the unobservability of N^* can lead to economically meaningful differences in the estimates of bidders' profit margins.

In many auction applications, researchers do not observe N^*, the number of bidders in the auction. (In the parlance of the literature, N^* is the "number of potential bidders", a terminology we adopt in the remainder of the paper.) The most common scenario obtains under binding reserve prices. When reserve prices bind, the number of potential bidders N^*, which is observed by auction participants and influences their bidding behavior, differs from the observed number of bidders A ($\leq N^*$), which is the number of auction participants whose bids exceed the reserve price. Other scenarios which would cause N^* to be unknown to the researcher include bidding or participation costs. In other cases, the number of auction participants may simply not be recorded in the researcher's dataset.

*The authors can be reached at yhu@jhu.edu and mshum@jhu.edu. We thank seminar participants at Brown, Caltech, FTC, UC-Irvine, Iowa, NC State, Toronto, Yale, and SITE (Stanford) for helpful comments. Guofang Huang provided exceptional research assistance.

In this paper, we consider nonparametric identification and estimation of first-price auction models when N^* is observed by bidders, but not by the researcher. Using recent results from the literature on misclassified regressors, we show how the equilibrium distribution of bids, given the unobserved N^*, can be identified and estimated. In the case of first-price auctions, these bid distributions estimated using our procedure can be used as inputs into established nonparametric procedures (Guerre, Perrigne, and Vuong (2000), Li, Perrigne, and Vuong (2002)) to obtain estimates of bidders' valuations.

Accommodating the possibility that the researcher does not know N^* is important for drawing valid policy implications from auction model estimates. Because N^* is the level of competition in an auction, not knowing N^* (or using a mismeasured value for N^*) can lead to wrong implications about the degree of competitiveness in the auction, and also the extent of bidders' markups and profit margins. Indeed, a naïve approach where the number of observed bids is used as a proxy for N^* will tend to overstate competition, because the unknown N^* is always (weakly) larger than the number of observed bids. This bias will be shown in the empirical illustration below.

Not knowing the potential number of bidders N^* has been an issue since the earliest papers in the structural empirical auction literature. In the parametric estimation of auction models, the functional relationship between the bids b and number of potential bidders N^* is explicitly parameterized, so that not knowing N^* need not be a problem. For instance, Laffont, Ossard, and Vuong (1995) used a goodness-of-fit statistic to select the most plausible value of N^* for French eggplant auctions. Paarsch (1997) treated N^* essentially as a random effect and integrates it out over the assumed distribution in his analysis of timber auctions.

In a nonparametric approach to auctions, however, the relationship between the bids b and N^* must be inferred directly from the data, and not knowing N^* (or observing N^* with error) raises difficulties. Within the independent private-values (IPV) framework, and under the additional assumption that the unknown N^* is fixed across all auctions (or fixed across a known subset of the auctions), Guerre, Perrigne, and Vuong (2000) showed how to identify N^* and the equilibrium bid distribution in the range of bids exceeding the reserve price. Hendricks, Pinkse, and Porter (2003) allowed N^* to vary across auctions, and assume that $N^* = L$, where L is a measure of the number of potential bidders which they construct.

The main contribution of this paper is to present a solution for the nonparametric identification and estimation of first-price auction models in which the number of bidders N^*

is observed by bidders, but unknown to the researcher. We develop a nonparametric procedure for recovering the distribution of bids conditional on unknown N^* which requires neither N^* to be fixed across auctions, nor for an (assumed) perfect measure of N^* to be available. Our procedure applies results from the recent econometric literature on models with misclassification error, such as e.g. Mahajan (2006), Hu (2007).

For first-price auctions, allowing the unknown N^* to vary across auctions is not innocuous. Because N^* is observed by the bidders, it affects their equilibrium bidding strategies. Hence, when N^* is not known by the researcher, and varies across auctions, the observed bids are drawn from a mixture distribution, where the "mixing densities" $g(b|N^*)$ and the "mixing weights" $\Pr(A|N^*)$ are both unknown. This motivates the application of econometric methods developed for models with a misclassified regressor, where (likewise) the observed outcomes are drawn from a mixture distribution.

Most closely related to our work is a paper by Song (2004). She solved the problem of the nonparametric estimation of ascending auction models in the IPV framework, when the number of potential bidders N^* is unknown by the researcher (and varies in the sample). She showed that the distribution of valuations can be recovered from observation of any two valuations of which rankings from the top is known.[1] However, her approach cannot be applied to first-price auctions, which are the focus of this paper. The reason for this is that, in IPV first-price auctions (but not in ascending- or second-price auctions), even if the distribution of bidders' valuations do not vary across the unknown N^*, the equilibrium distribution of bids still vary across N^*. Hence, because the researcher does not know N^*, the observed bids are drawn from a mixture distribution, and estimating the model requires deconvolution methods which have been developed in the econometric literature on measurement error.[2]

In a different context, Li, Perrigne, and Vuong (2000) applied deconvolution results from the (continuous) measurement error literature to identify and estimate conditionally independent auction models in which bidders' valuations have common and private (idiosyncratic) components. Krasnokutskaya (2005) also used deconvolution results to estimate auction models with unobserved heterogeneity. To our knowledge, however, our paper is the first

[1] Adams (2007) also considers estimation of ascending auctions when the distribution of potential bidders is unknown.

[2] Song (2006) showed that the top two bids are also enough to identify first-price auctions where the number of active bidders is *not observed* by bidders. Under her assumptions, however, the observed bids are i.i.d. samples from a homogeneous distribution, so that her estimation methodology would not work for the model considered in this paper.

application of (discrete) measurement error results to estimate an auction model where the number of potential bidders is unknown.

The issues considered in this paper are close to those considered in the literature on entry in auctions: eg. Li (2005), Li and Zheng (2006), Athey, Levin, and Seira (2005), Krasnokutskaya and Seim (2005), Haile, Hong, and Shum (2003). While the entry models considered in these papers differ, their one commonality is to model more explicitly bidders' participation decisions in auctions, which can cause the number of observed bidders A to differ from the number of potential bidders N^*. For instance, Haile, Hong, and Shum (2003) consider an endogenous participation model in which the number of potential bidders is observed by the researcher, and equal to the observed number of bidders (i.e., $N^* = A$), so that non-observability of N^* is not a problem. However, A is potentially endogenous, because it may be determined in part by auction-specific unobservables which also affect the bids. By contrast, in this paper we assume that N^* is unobserved, and that $N^* \neq A$, but we do not consider the possible endogeneity of N^*.[3]

In section 2, we describe our auction framework. In section 3, we present the main identification results, and describe our estimation procedure. In section 4, we provide Monte Carlo evidence of our estimation procedure, and discuss some practical implementation issues. In section 5, we present an empirical illustration, using data from procurement auctions in New Jersey. In section 6, we consider extensions of the approach to scenarios where only the winning bid is observed. Section 7 concludes. Proofs of the asymptotic properties of our estimator are presented in the Appendix.

1 Model

In this paper, we consider the case of first-price auctions under the symmetric independent private values (IPV) paradigm, for which identification and estimation are most transparent. For a thorough discussion of identification and estimation of these models when the number of potential bidders N^* is known, see Paarsch and Hong (2006, Chap. 4). For concreteness, we focus on the case where a binding reserve price is the reason why the number of potential bidders N^* differs from the observed number of bidders, and is not known by the researcher.

[3] In principle, we recover the distribution of bids (and hence the distribution of valuations) separately for each value of N^*, which accommodates endogeneity in a general sense. However, because we do not model the entry process explicitly (as in the papers cited above), we do not deal with endogeneity in a direct manner.

There are N^* bidders in the auction, with each bidder drawing a private valuation from the distribution $F(x)$ which has support $[\underline{x}, \bar{x}]$. N^* can vary freely across the auctions, and while it is observed by the bidders, it is not known by the researcher. There is a reserve price r, assumed to be fixed across all auctions, where $r > \underline{x}$.[4] The equilibrium bidding function for bidder i with valuation x_i is

$$b(x_i; N^*) \begin{cases} = x_i - \frac{\int_r^{x_i} F(s)^{N^*-1} ds}{F(x_i)^{N^*-1}} & \text{for } x_i \geq r \\ 0 & \text{for } x_i < r. \end{cases} \tag{1}$$

Hence, the number of bidders observed by the researcher is $A \equiv \sum_{i=1}^{N^*} \mathbf{1}(x_i > r)$, the number of bidders whose valuations exceed the reserve price.

For this case, the equilibrium bids are i.i.d. and, using the change-of-variables formula, the density of interest $g(b|N^*, b > r)$ is equal to

$$g(b|N^*, b > r) = \frac{1}{b'(\xi(b; N^*); N^*))} \frac{f(\xi(b; N^*))}{1 - F(r)}, \text{ for } b > r \tag{2}$$

where $\xi(b; N^*)$ denotes the inverse of the equilibrium bid function $b(\cdot; N^*)$ evaluated at b. In equilibrium, each observed bid from an N^*-bidder auction is an i.i.d. draw from the distribution given in Eq. (2), which does not depend on A, the observed number of bidders.

We propose a two-step estimation procedure. In the first step, the goal is to recover the density $g(b|N^*; b > r)$ of the equilibrium bids, for the truncated support $(r, +\infty)$. (For convenience, in what follows, we suppress the conditioning truncation event $b > r$.) To identify and estimate $g(b|N^*)$, we use the results from Hu (2007).

In second step, we use the methodology of Guerre, Perrigne, and Vuong (2000) to recover the valuations x from the joint density $g(b|N^*)$. For each b in the marginal support of $g(b|N^*)$, the corresponding valuation x is obtained by

$$\xi(b, N^*) = b + \frac{1}{N^* - 1} \left[\frac{G(b|N^*)}{g(b|N^*)} + \frac{F(r)}{1 - F(r)} \cdot \frac{1}{g(b|N^*)} \right]. \tag{3}$$

For most of this paper, we focus on the first step of this procedure, because the second step is a straightforward application of standard techniques.

[4] Our estimation methodology can potentially also be used to handle the case where $N^*$is fixed across all auctions, but r varies freely across auctions.

2 Nonparametric identification

In this section, we apply the results from Hu (2007) to show the identification of the first-price auction model with unknown N^*. The procedure requires two auxiliary variables:

1. a proxy N, which is a mismeasured version of N^*
2. an instrument Z, which could be a second corrupted measurement of N^*.

The auxiliary variables (N, Z) must satisfy three conditions. The first two conditions are given here:

Condition 1 $g(b|N^*, N, Z) = g(b|N^*)$.

This assumption implies that N or Z affects the equilibrium density of bids only through the unknown number of potential bidders N^*. In the econometric literature, this is known as the "nondifferential" measurement error assumption.

In what follows, we only consider values of b such that $g(b|N^*) > 0$, for $N^* = 2, \ldots, K$. This requires, implicitly, knowledge of the support of $g(b|N^*)$, which is typically unknown to the researcher. Below, when we discuss estimation, we present a two-step procedure for $g(b|N^*)$ which circumvents this problem.

Condition 2 $g(N|N^*, Z) = g(N|N^*)$.

This assumption implies that the instrument Z affects the mismeasured N only through the number of potential bidders. Roughly, because N is a noisy measure of N^*, this condition requires that the noise is independent of the instrument Z, conditional on N^*.

Examples of N and Z Here we consider several examples of variables which could fulfill the roles of the auxiliary variables N and Z.

1. One advantage to focusing on the IPV model is that A, the observed number of bidders, can be used in the role of N. Particularly, for a given N^*, the sampling density of any equilibrium bid exceeding the reserve price — as given in Eq. (2) above — does not depend

on A, so that Condition 1 is satisfied.[5] A good candidate for the instrument Z could be a noisy estimate of N^*:

$$Z = h(N^*, \eta).$$

In order to satisfy conditions 1 and 2, we would require $b \perp \eta | N^*$, and also $A \perp \eta | N^*$. Because we are focused on the symmetric IPV model in this paper, we will consider this example in the remainder of this section, and also in our Monte Carlo experiments and in the empirical illustration.

2. More generally, N and Z could be two noisy measures of N^*:

$$\begin{aligned} N &= f(N^*, \upsilon) \\ Z &= h(N^*, \eta). \end{aligned} \tag{4}$$

In order to satisfy conditions 1 and 2, we would require $b \perp (\upsilon, \eta) | N^*$, as well as $\eta \perp \upsilon | N^*$.[6]

3. Another possibility is that N is a noisy measure of N^*, as in example 2, but Z is an exogenous variable which directly determines participation:

$$\begin{aligned} N &= f(N^*, \upsilon) \\ N^* &= k(Z, \nu). \end{aligned} \tag{5}$$

In order to satisfy conditions 1 and 2, we would require $b \perp (\upsilon, Z) | N^*$, as well as $\upsilon \perp Z | N^*$. This implies that Z is excluded from the bidding strategy, and affects bids only through its effect on N^*.

Furthermore, in this example, in order for the second step of the estimation procedure (in which we recover bidders' valuations) to be valid, we also need to assume that $b \perp \nu | N^*$. Importantly, this rules out the case that the participation shock ν is a source of unobserved auction-specific heterogeneity.[7] Note that ν will generally be (unconditionally) correlated with the bids b, which our assumptions allow for. ■

We observe a random sample of $\left\{\vec{b}_t, N_t, Z_t\right\}$, where $\vec{b}_t$ denotes the vector of observed bids $\{b_{1t}, b_{2t}, \ldots, b_{A_t t}\}$. (Note that we only observe A_t bids for each auction t.) We assume the

[5] This is no longer true in affiliated value models.

[6] This allows ν and η to be correlated through N^* and, indeed, allows both ν and η to depend on N^*.

[7] In the case when N^* is observed, correlation between bids and the participation shock ν can be accommodated, given additional restriction on the $k(\cdots)$ function. See Guerre, Perrigne, and Vuong (2005) and Haile, Hong, and Shum (2003) for details. However, when N^* is unobserved, as is the case here, it is not clear how to generalize these results.

variables N, Z, and N^* share the same support $\mathcal{N} = \{2..., K\}$. Here K can be interpreted as the maximum number of bidders, which is fixed across all auctions.[8]

By the law of total probability, the relationship between the observed distribution $g(b, N, Z)$ and the latent densities is as follows:

$$g(b, N, Z) = \sum_{N^*=2}^{K} g(b|N^*, N, Z)g(N|N^*, Z)g(N^*, Z). \tag{6}$$

Under conditions 1 and 2, Eq. (6) becomes

$$g(b, N, Z) = \sum_{N^*=2}^{K} g(b|N^*)g(N|N^*)g(N^*, Z). \tag{7}$$

We define the matrices

$$\begin{aligned} G_{b,N,Z} &\equiv [g(b, N = i, Z = j)]_{i,j}, \\ G_{N|N^*} &\equiv [g(N = i|N^* = k)]_{i,k}, \\ G_{N^*,Z} &\equiv [g(N^* = k, Z = j)]_{k,j}, \\ G_{N,Z} &\equiv [g(N = i, Z = j)]_{i,j}, \end{aligned}$$

and

$$G_{b|N^*} \equiv \begin{pmatrix} g(b|N^* = 2) & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & g(b|N^* = K) \end{pmatrix}. \tag{8}$$

All of these are $(K-1)$-dimensional square matrices. With this notation, Eq. (7) can be written as

$$G_{b,N,Z} = G_{N|N^*}G_{b|N^*}G_{N^*,Z}. \tag{9}$$

Condition 2 implies that

$$g(N, Z) = \sum_{N^*=2}^{K} g(N|N^*)g(N^*, Z), \tag{10}$$

which, using the matrix notation above, is equivalent to

$$G_{N,Z} = G_{N|N^*}G_{N^*,Z}. \tag{11}$$

[8] Our identification results still hold if Z has more possible values than N and N^*.

Equations (9) and (11) summarize the unknowns in the model, and the information in the data. The matrices on the left-hand sides of these equations are quantities which can be recovered from the data, whereas the matrices on the right-hand side are the unknown quantities of interest. As a counting exercise, we see that the matrices $G_{b,N,Z}$ and $G_{N,Z}$ contain $2(K-1)^2-(K-1)$ known elements, while the unknown matrices $G_{N|N^*}$, $G_{N^*,Z}$ and $G_{b|N^*}$ contain at most a total of also $2(K-1)^2-(K-1)$ unknown elements. Hence, in principle, there is enough information in the data to identify the unknown matrices. The key part of the proof below is to characterize the solution and give conditions for uniqueness. Moreover, the proof is constructive in that it immediately suggests a way for estimation.

The third condition which the auxiliary variables N and Z must satisfy is a rank condition:

Condition 3 *$Rank\left(G_{N,Z}\right)=K-1$.*

Note that this condition is directly testable from the sample. It essentially ensures that the instrument Z affects the distribution of the proxy variable N (resembling the standard instrumental relevance assumption in usual IV models).

Because Eq. (11) implies that

$$Rank\left(G_{N,Z}\right)\leq\min\left\{Rank\left(G_{N|N^*}\right),Rank\left(G_{N^*,Z}\right)\right\}, \tag{12}$$

it follows from Condition 3 that $Rank\left(G_{N|N^*}\right)=K-1$ and $Rank\left(G_{N^*,Z}\right)=K-1$. In other words, the matrices $G_{N,Z}$, $G_{N|N^*}$, and $G_{N^*,Z}$ are all invertible. Therefore, postmultiplying both sides of Eq. (9) by $G_{N,Z}^{-1}=G_{N^*,Z}^{-1}G_{N|N^*}^{-1}$, we obtain the key equation

$$G_{b,N,Z}G_{N,Z}^{-1}=G_{N|N^*}G_{b|N^*}G_{N|N^*}^{-1}. \tag{13}$$

The matrix on the left-hand side can be formed from the data. For the expression on the right-hand side, note that because $G_{b|N^*}$ is diagonal (cf. Eq. (8)), the RHS matrix represents an eigenvalue-eigenvector decomposition of the LHS matrix, with $G_{b|N^*}$ being the diagonal matrix of eigenvalues, and $G_{N|N^*}$ being the corresponding matrix of eigenvectors. This is the key representation which will identify and facilitate estimation of the unknown matrices $G_{N|N^*}$ and $_{b|N^*}$.

In order to make the eigenvalue-eigenvector decomposition in Eq. (13) unique, we assume:

Condition 4 *For any $i,j\in\mathcal{N}$, the set $\{(b):g(b|N^*=i)\neq g(b|N^*=j)\}$ has nonzero Lebesgue measure whenever $i\neq j$.*

This assumption (which is actually implied by equilibrium bidding) guarantees that the eigenvalues in $G_{b|N^*}$ are distinctive for some bid b, which ensures that the eigenvalue decomposition in Eq. (13) exists and is unique, for some bid b. This assumption guarantees that all the linearly independent eigenvectors are identified from the decomposition in Eq. (13). Suppose that for some value $\widetilde{b}$, $g(\widetilde{b}|N^* = i) = g(\widetilde{b}|N^* = j)$, which implies that the two eigenvalues corresponding to $N^* = i$ and $N^* = j$ are the same. In this case, the two corresponding eigenvectors cannot be uniquely identified, because any linear combination of the two eigenvectors is still an eigenvector. Assumption 4 guarantees that there exists another value $\overline{b}$ such that $g(\overline{b}|N^* = i) \neq g(\overline{b}|N^* = j)$. Eq. (13) holds for every b, implying that $g(\widetilde{b}|N^* = i)$ and $g(\overline{b}|N^* = i)$ correspond to the same eigenvector, as do $g(\widetilde{b}|N^* = j)$ and $g(\overline{b}|N^* = j)$. Therefore, although we cannot use $\widetilde{b}$ to uniquely identify the two eigenvectors corresponding to $N^* = i$ and $N^* = j$, we can use the value $\overline{b}$ to identify them.

Given Condition 4, Eq. (13) shows that an eigenvalue decomposition of the observed $G_{b,N,Z}G_{N,Z}^{-1}$ matrix identifies $G_{b|N^*}$ and $G_{N|N^*}$ up to a normalization and ordering of the columns of the eigenvector matrix $G_{N|N^*}$.

There is a clear appropriate choice for the normalization constant of the eigenvectors; because each column of $G_{N|N^*}$ should add up to one, we can multiply each element $G_{N|N^*}(i,j)$ by the reciprocal of the column sum $\sum_i G_{N|N^*}(i,j)$, as long as $G_{N|N^*}(i,j)$ is non-negative.

The appropriate ordering of the columns of $G_{N|N^*}$ is less clear, and in order to complete the identification, we need an additional assumption which pins down the ordering of these columns. One such assumption is:

Condition 5 $N \leq N^*$.

The condition $N \leq N^*$ is natural, and automatically satisfied, when $N = A$, the observed number of bidders. This condition implies that for any $i, j \in \mathcal{N}$

$$g\left(N = j|N^* = i\right) = 0 \text{ for } j > i. \tag{14}$$

In other words, $G_{N|N^*}$ is an upper-triangular matrix. Since the triangular matrix $G_{N|N^*}$ must be invertible (by Eq. (12), its diagonal entries are all nonzero, i.e.,

$$g\left(N = i|N^* = i\right) > 0 \text{ for all } i \in \mathcal{N}. \tag{15}$$

In other words, Condition 5 implies that, once we have the columns of $G_{N|N^*}$ obtained as the eigenvectors from the matrix decomposition (13), the right ordering can be obtained by re-arranging these columns so that they form an upper-triangular matrix.

Hence, under Conditions 1-5, $G_{b|N^*}$, $G_{N|N^*}$ and also $G_{N^*,Z}$ are identified (the former pointwise in b).

3 Nonparametric Estimation: two-step procedure

In this section, we give details on the estimation of $(b|N^*)$ given observations of (b, N, Z), for the symmetric independent private values model. In the key equation (13), the matrix $G_{N|N^*}$ is identical for all b.[9] This suggests a convenient two-step procedure for estimating the unknown matrices $G_{N|N^*}$ and $G(b|N^*)$.

Step One In Step 1, we estimate the eigenvector matrix $G_{N|N^*}$. To maximize the convergence rate in estimating $G_{N|N^*}$, we average across values of the bid b. Specifically, from Eq. (7), we have

$$E(b|N,Z)g(N,Z) = \sum_{N^*=2}^{K} E(b|N^*)g(N|N^*)g(N^*,Z). \tag{16}$$

Define the matrices

$$G_{Eb,N,Z} \equiv \left[E\left(b|N=i,Z=j\right)g(N=i,Z=j)\right]_{i,j}, \tag{17}$$

and

$$G_{Eb|N^*} \equiv \begin{pmatrix} E\left[b|N^*=2\right] & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & E\left[b|N^*=K\right] \end{pmatrix}.$$

Then

$$G_{Eb,N,Z} = G_{N|N^*}G_{Eb|N^*}G_{N^*,Z}$$

[9]This also implies that there is a large degree of overidentification in this model, and suggests the possibility of achieving identification with weaker assumptions. In particular, it may be possible to relax the non-differentiability condition 1 so that we require $g(b|N^*,N,Z) = g(b|N^*)$ only at one particular value of b. We are exploring the usefulness of such possibilities in ongoing work.

and, as before, postmultiplying both sides of this equation by $G_{N,Z}^{-1} = G_{N^*,Z}^{-1} G_{N|N^*}^{-1}$, we obtain an integrated version of the key equation:

$$G_{Eb,N,Z} G_{N,Z}^{-1} = G_{N|N^*} G_{Eb|N^*} G_{N|N^*}^{-1}. \tag{18}$$

This implies

$$G_{N|N^*} = \psi\left(G_{Eb,N,Z} G_{N,Z}^{-1}\right),$$

where $\psi(\cdot)$ denotes the mapping from a square matrix to its eigenvector matrix following the identification procedure in the previous section.[10] As mentioned in Hu (2007), the function $\psi(\cdot)$ is a nonstochastic analytic function. Therefore,we may estimate $G_{N|N^*}$ as follows:

$$\widehat{G}_{N|N^*} := \psi\left(\widehat{G}_{Eb,N,Z} \widehat{G}_{N,Z}^{-1}\right), \tag{19}$$

where $\widehat{G}_{Eb,N,Z}$ and $\widehat{G}_{N,Z}$ may be constructed directly from the sample. In our empirical example, we estimate $\widehat{G}_{Eb,N,Z}$ using a sample average:

$$\widehat{G}_{Eb,N,Z} = \left[\frac{1}{T}\sum_t \frac{1}{N_j}\sum_{i=1}^{N_j} b_{it}\mathbf{1}(N_t = N_j, Z_t = Z_k)\right]_{j,k}. \tag{20}$$

Step Two In Step 2, we estimate $g(b|N^*)$. With $G_{N|N^*}$ estimated by $\widehat{G}_{N|N^*}$ in step 1, we may proceed to estimate $g(b|N^*)$, pointwise in b. From Eq. (13), we have for any b

$$G_{N|N^*}^{-1}\left(G_{b,N,Z} G_{N,Z}^{-1}\right) G_{N|N^*} = G_{b|N^*}. \tag{21}$$

Define $e_{N^*} = (0, ...0, 1, 0, ..., 0)^T$, where 1 is at the N^*-th position in the vector. We have

$$g(b|N^*) = e_{N^*}^T\left[G_{N|N^*}^{-1}\left(G_{b,N,Z} G_{N,Z}^{-1}\right) G_{N|N^*}\right] e_{N^*}. \tag{22}$$

which holds for all $b \in (-\infty, \infty)$. Hence, we may estimate $g(b|N^*)$ as follows:

$$\widehat{g}(b|N^*) := e_{N^*}^T\left[\widehat{G}_{N|N^*}^{-1}\left(\widehat{G}_{b,N,Z} \widehat{G}_{N,Z}^{-1}\right) \widehat{G}_{N|N^*}\right] e_{N^*}, \tag{23}$$

[10] In order for $G_{N|N^*}$ to be recovered from this eigenvector decomposition, Condition 4 from the previous section must be strengthened so that the conditional means $E[b|N^*]$, which are the eigenvalues from this decomposition, are distinct for every N^*.

where $\widehat{G}_{N|N^*}$ is estimated in step 1 and $\widehat{G}_{b,N,Z}$ may be constructed directly from the sample. In our empirical work, we use a kernel estimate for $\widehat{g}_{b,N,Z}(b, N_j, Z_k)$:

$$\widehat{g}_{b,N,Z}(b, N_j, Z_k) = \left[\frac{1}{Th} \sum_t \frac{1}{N_t} \sum_{i=1}^{N_t} K\left(\frac{b - b_{it}}{h} \right) \mathbf{1}(N_t = N_j, Z_t = Z_k) \right]. \tag{24}$$

The bid b may have a different unknown support for different N^*. That is,

$$g(b|N^*) = \begin{cases} > 0 & \text{for } b \in [r, u_{N^*}] \\ = 0 & \text{otherwise} \end{cases},$$

where u_{N^*}, the upper bound of the support of $g(b|N^*)$, may not be known by the researcher. In practice, we may estimate the upper bound u_{N^*} as follows:

$$\hat{u}_{N^*} = \sup \{ b : \hat{g}(b|N^*) > 0 \}.$$

We analyze the asymptotic properties of our estimator in detail in the appendix. Here we provide a brief summary. Given the discreteness of N, Z, and the use of a sample average to construct $\widehat{G}_{Eb,N,Z}$ (via. Eq. (20)), the estimates of $\widehat{G}_{N|N^*}$ (obtained using Eq. (19)) and $\widehat{G}_{N,Z}$ should converge at a $\sqrt{T}$-rate (where T denotes the total number of auctions).

Hence, pointwise in b, the convergence properties of $\widehat{g}(b|N^*)$ to $g(b|N^*)$, where $\widehat{g}(b|N^*)$ is estimated using Eq. (23), will be determined by the convergence properties of the kernel estimate of $g(b, N, Z)$ in Eq. (24), which converges at a rate slower than $\sqrt{T}$. In the Appendix, we show that, pointwise in b, $(Th)^{1/2} [\widehat{g}(b|N^*) - g(b|N^*)]$ converges to a normal distribution. We also present a uniform convergence rate for $\widehat{g}(b|N^*)$.

The matrix $G_{N|N^*}$, which is a by-product of the estimation procedure, can be useful for specification testing, when $N = A$, the observed number of bidders. Under the assumption that the difference between the observed number of bidders A and the number of potential bidders N^* arises from a binding reserve price, and that the reserve price r is fixed across all the auctions with the same N^* in the dataset, it is well-known (cf. Paarsch (1997)) that

$$A|N^* \ \sim \ \text{Binomial}(N^*, 1 - F_v(r)) \tag{25}$$

where $F_v(r)$ denotes the CDF of bidders' valuations, evaluated at the reserve price. This suggests that the recovered matrix $G_{A|N^*}$ can be useful in two respects. First, using Eq. (25), the truncation probability $F_v(r)$ could be estimated. This is useful when we use the first-order condition (3) to recover bidders' valuations. Alternatively, we could also test whether the columns of $G_{A|N^*}$, which correspond to the probabilities $\Pr(A|N^*)$ for a fixed N^*, are consistent with the binomial distribution in Eq. (25).

4 Monte Carlo Evidence

In this section, we present some Monte Carlo evidence for the IPV model. Using simulated bids, we estimate the bid densities and bidder valuations using the procedure presented in section 3.

We consider first price auctions where bidders' valuations $x_i \sim U[0,1]$, independently across bidders i. With a reserve price $r > 0$, the equilibrium bidding strategy with N^* bidders is:

$$b^*(x; N^*) = \begin{cases} \left(\frac{N^*-1}{N^*}\right) x + \frac{1}{N^*} \left(\frac{r}{x}\right)^{N^*-1} r & \text{if } x \geq r \\ 0 & \text{if } x < r. \end{cases} \tag{26}$$

For each auction t, we need to generate the equilibrium bids b_{jt}, for $j = 1, \ldots N_t^*$, as well as (N_t^*, N_t, Z_t). In this exercise, N_t is taken to be the number of observed bidders A_t, and Z_t is a corrupted measure of N_t^*.

For each auction t, the number of potential bidders N_t^* is generated uniformly on $\{2, 3, \ldots, K\}$, where K is the maximum number of bidders. Subsequently, we generate Z_t as a corrupted measure of N_t^*:

$$Z_t = \begin{cases} N_t^* & \text{with probability } q \\ \text{unif.}\{2, 3, \ldots, K\} & \text{with probability } 1 - q. \end{cases} \tag{27}$$

For each auction t, and each bidder $j = 1, \ldots, N_t^*$, we draw valuations $x_j \sim U[0,1]$, and construct the corresponding equilibrium bids using Eq. (26). Finally, the number of observed bidders is determined as the number of bidders whose valuations exceed the reserve price:

$$A_t = \sum_{j \in \mathcal{N}_t^*} \mathbf{1}(x_j \geq r) \tag{28}$$

The estimation procedure in section 3 above requires the matrix $G_{A|N^*}$ to be square, but in generating the variables here, the support of A is $\{1, 2, .., K\}$ while the support of N^* is $\{2, .., K\}$. To accommodate this, we define

$$N = \begin{cases} A & \text{if } A \geq 3 \\ 2 & \text{if } A \leq 2 \end{cases}.$$

Therefore, N has the same support as N^*. This redefinition does not affect any of the identification arguments given above.

4.1 Results

We present results from $S = 200$ replications of a simulation experiment. First, we consider the case where K (the maximum number of bidders) is equal to 4. The performance of our estimation procedure is illustrated in Figure 1. The estimator perform well for all values of $N^* = 2, 3, 4$, and for a modest-sized dataset of $T = 302$ auctions. Across the Monte Carlo replications, the estimated density functions track the actual densities quite closely. In these graphs, we also plot $g(b|A = n)$, the bid density conditioned on the observed number of bidders, for $n = 2, 3, 4$, which we consider a "naïve" estimator for $g(b|N^* = n)$. For $N^* = 2, 3$, our estimator outperforms the naïve estimator, especially for the case of $N^* = 2$.

In Figure 2, we present estimates of bidders' valuations. In each graph on the left-hand-side of the figure, we graph the bids against three measures of the corresponding valuation: (i) the actual valuation, computed from Eq. (3) using the actual bid densities $g(b|N^*)$, and labeled "True values"; (ii) the estimated valuations using our estimates of $g(b|N^*)$, labeled "Estimated value"[11]; and (iii) naïve estimates of the values, computed using $g(b|A)$, the observed bid densities conditional on the observed number of bidders.[12]

The graphs show that there are sizeable differences between the value estimates, across all values of the bids. For all values of N^* (=2,3,4), we see that our estimator tracks the true values quite closely. In contrast, the naïve approach underestimates the valuations. This is to be expected – because $N^* \geq A$, the set of auctions with a given value of A actually have a true level of competition larger than A. Hence, the naïve approach overstates the true level of competition, which leads to underestimation of bidders' markdowns $(v - b)/v$. The markdowns implied by our valuation estimates are shown in the right-hand-side graphs in Figure 2.

In a second set of experiments, we consider the case where the maximum number of bidders is $K = 6$. In these experiments, we increased the number of auctions to be $T = 1000$. Graphs summarizing these simulations are presented in Figure 3. Clearly, our estimator continues to perform well. In both the $N^* = 4$ as well as the $N^* = 6$ case, we see that the differences between our estimator and the naïve estimator diminish. This may not be surprising, because as N^* increases, the bidding strategies are less distinguishable for

[11] In computing these valuations, the truncation probability $F(r)$ in Eq. (3) is obtained from the first-step estimates of the misclassification probability matrix $G_{N|N^*}$ as $\hat{F}(r) = 1 - \left[\hat{G}(N^*|N^*)\right]^{1/N^*}$.

[12] In computing the values for the naïve approach, we use the first-order condition $\xi(b; A) = b + \frac{G(b|A)}{(A-1)\cdot g(b|A)}$, which ignores the possibility of a binding reserve price.

Figure 1: Monte Carlo Evidence: $K = 4$

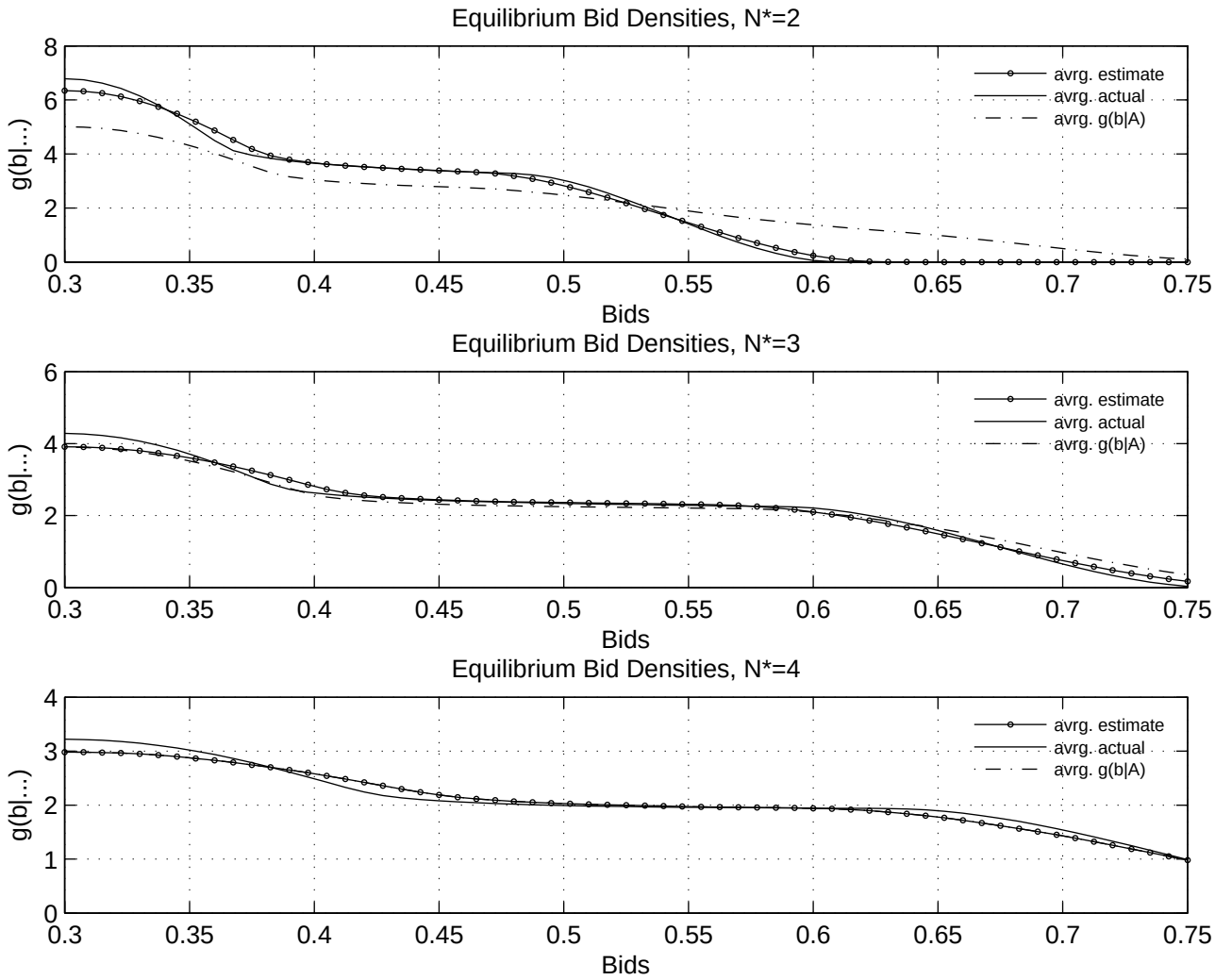
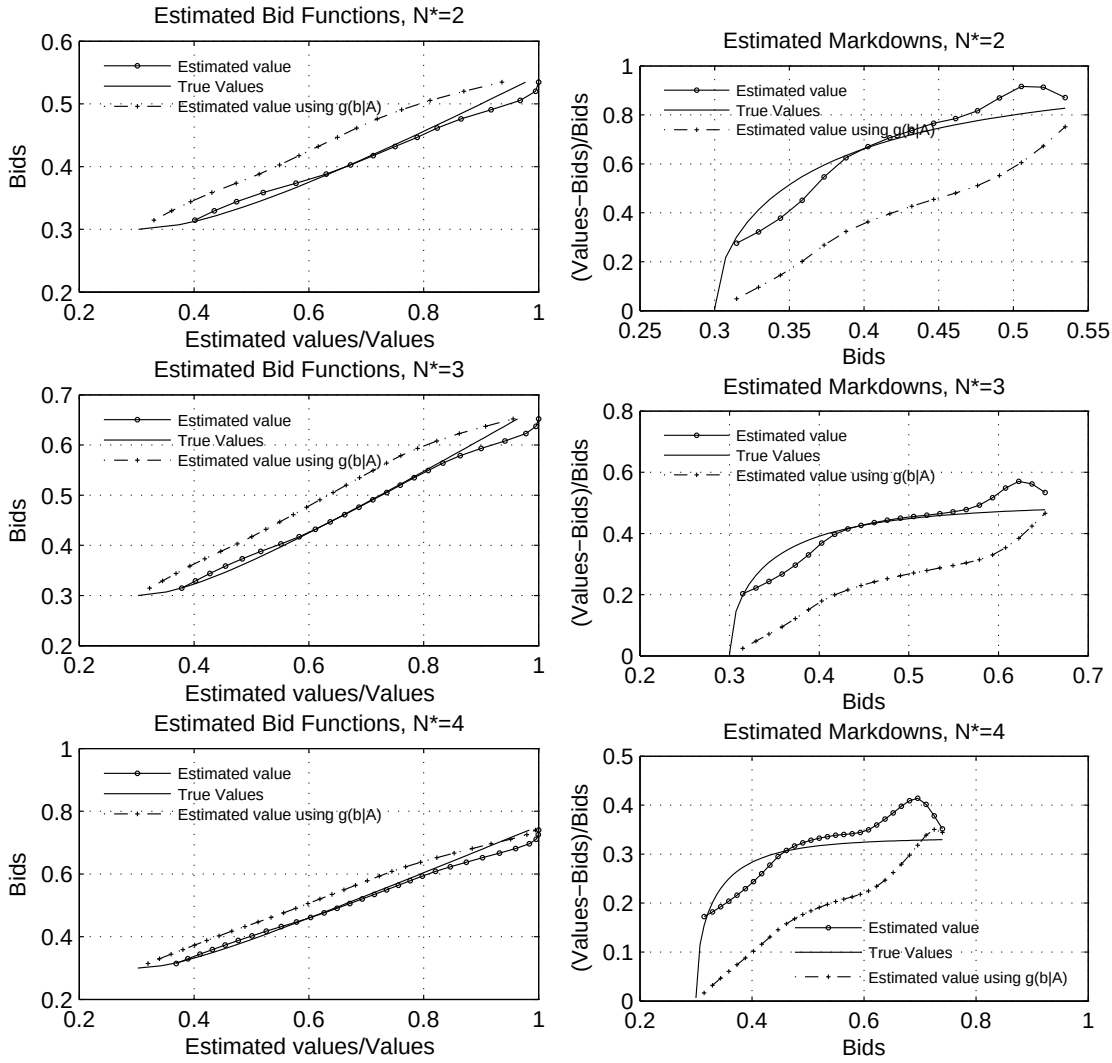


Figure 2: Estimates of bid functions and implied markdowns, $K = 4$ experiments



different values for N^* and, in the limit, as $N^* \to \infty$, the equilibrium bid density will approach the distribution of the valuations x. Hence, the error in using $g(b|A = n)$ as the estimator for $g(b|N^* = n)$ for larger n will be less severe.

The valuations implied by our estimates of the bid densities, for the $K = 6$ case, are presented in Figure 4. Qualitatively, the results are very similar to the $K = 4$ results presented earlier.

5 Empirical illustration

In this section, we illustrate our methodology using a dataset of low-bid construction procurement auctions held by the New Jersey Department of Transportation (NJDOT) in the years 1989–1997. This dataset was previously analyzed in Hong and Shum (2002), and a full description of it is given there.

Among all the auctions in our dataset, we focus on highway work construction projects, for which the number of auctions is the largest. In Table 1, we present some summary statistics on the auctions used in the analysis. Note that there were six auctions with just one bidder, in which non-infinite bids were submitted. If the observed number of bidders A is equal to N^*, the number of potential bidders observed by bidders when they bid, then the non-infinite bids observed in these one-bidder auctions is difficult to explain from a competitive bidding point of view.[13] However, occurrences of one-bidder auctions is a sign that the observed number of bidders is less than the potential number of bidders, and the methodology developed in this paper allows for this possibility.

For the two auxiliary variables, we used A, the number of observed bidders, in the role of the noisy measure N. In the role of the instrument Z, we constructed a measure of the average number of observed bidders in the five previous auctions of the same project category which took place before a given auction.[14]

[13] Indeed, Li and Zheng (2006, pg. 9) point out that even when bidders are uncertain about the number of competitors they are facing, finite bids cannot be explained when bidders face a non-zero probability that they could be the only bidder.

[14] The validity of using A_{-1} (values of A in previous auctions) to construct the instrument Z can be shown as follows: Let N^* and N^*_{-1} denote the potential number of bidders in the current and the previous auctions. With a binding reserve price, $A_{-1} = \sum_{i=1}^{N^*_{-1}} \mathbf{1}(x_i > r)$. Conditions 1 and 2 require that $g\left(b|N^*, N^*_{-1}, N\right) = g\left(b|N^*\right)$ and $g\left(N|N^*, N^*_{-1}\right) = g\left(N|N^*\right)$. Moreover, for condition 3, we require the matrix $G_{N|N_{-1}}$, elements of which are the conditional probabilities of $N|N_{-1}$, to be invertible. With these conditions, lagged values

Figure 3: Monte Carlo Evidence: $K = 6$

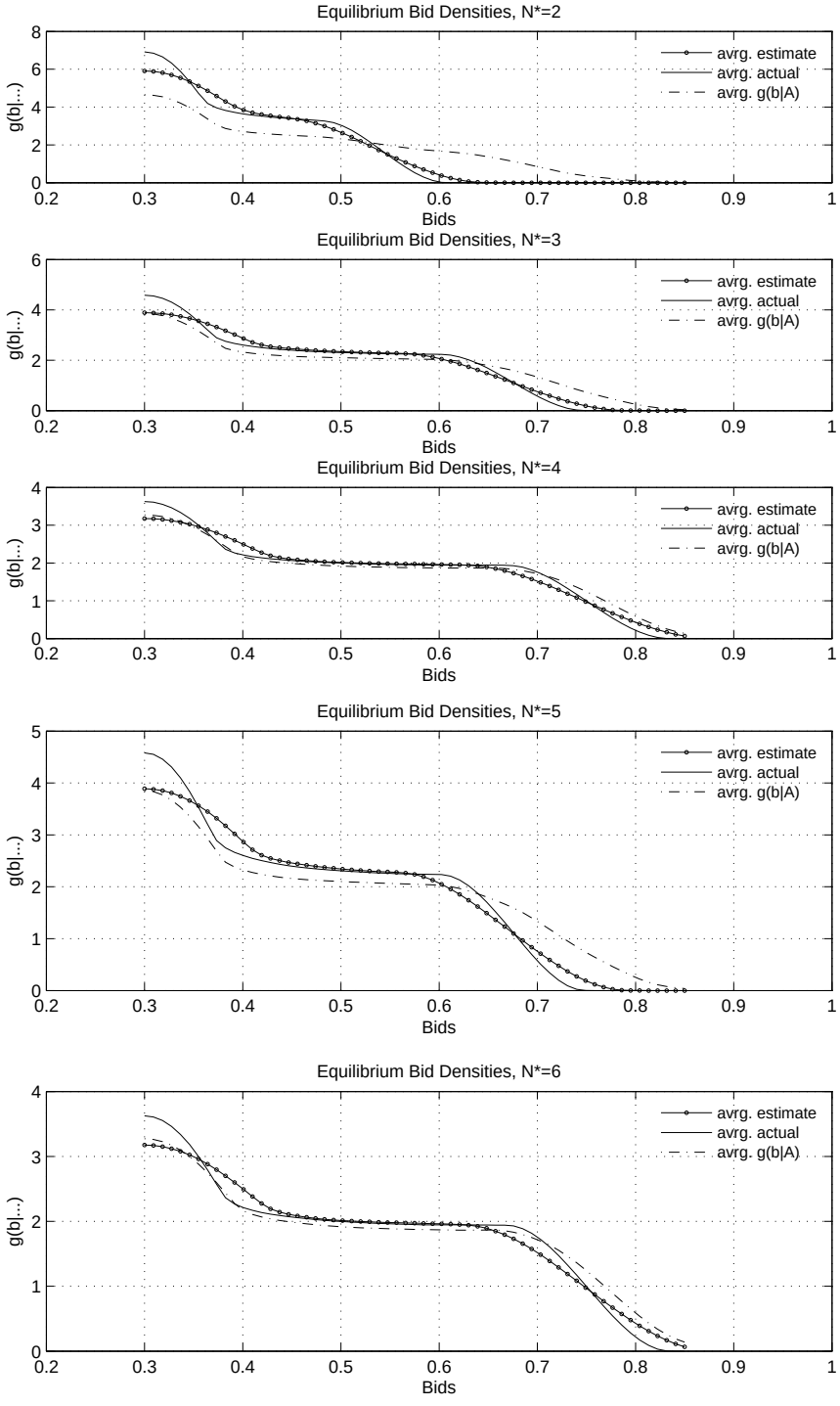


Figure 4: Estimates of bid functions, $K = 6$ experiments

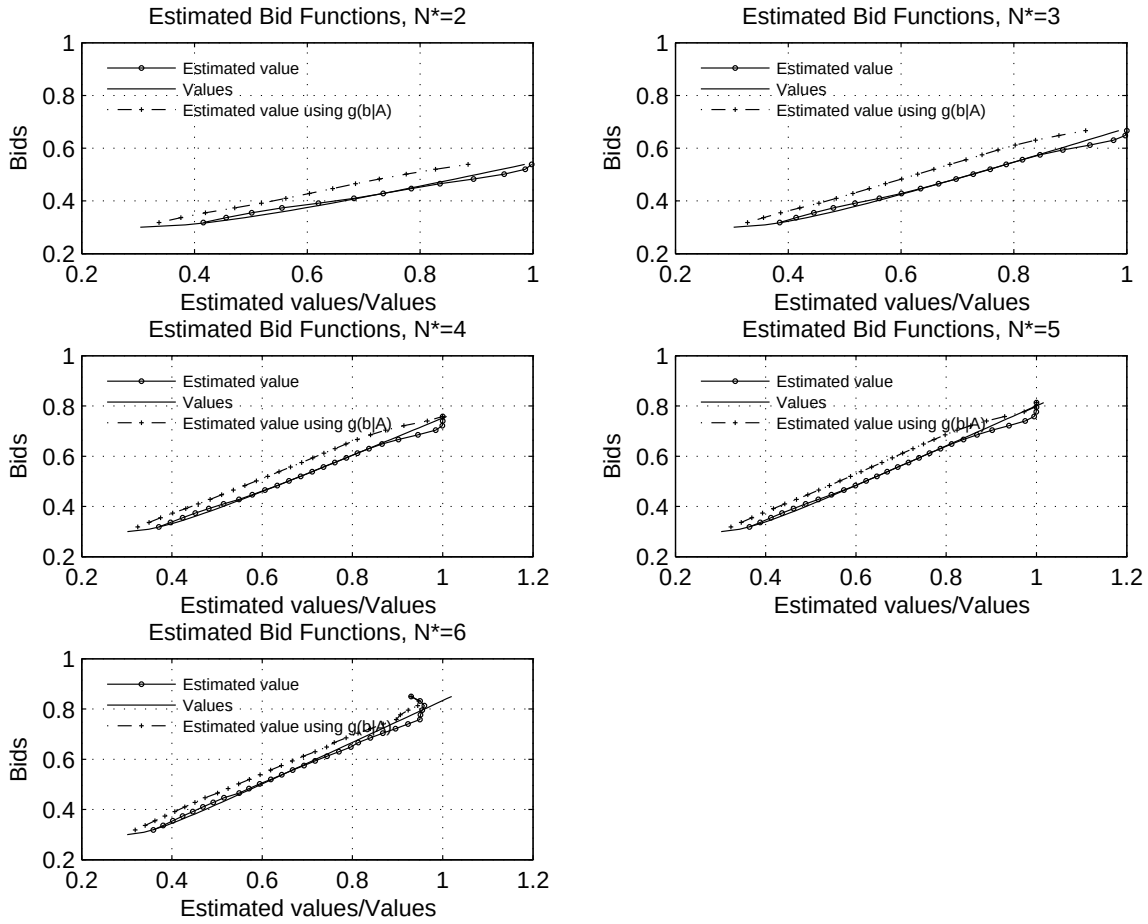


Table 1: Summary statistics of procurement auction data
Highway work auctions

Observed # bidders (A)	# aucs.	Freq.	avg bid[a]
1	6	1.42	0.575
2	12	2.84	5.894
3	31	7.33	1.692
4	46	10.87	1.843
5+	338	77.54	7.920

[a]: in millions of 1989$

In order to satisfy condition 3, which requires that the matrix $G_{N|Z}$ be full rank, we divided the values of A and Z into three categories: $\{(1,2,3),4,5+\}$, and correspondingly consider only three distinct values for $N^* \in \{3,4,5\}$. Furthermore, the ordering assumption that we make is that $A \leq N^*$, which is consistent with the story that bidders decide not to submit a bid due to an (implicit) reserve price.[15]

Because we model these auctions in a simplified setting, we do not attempt a full analysis of these auctions. Rather, this exercise highlights some practical issues in implementing the estimation methodology. There are two important issues. First, the assumption that $A \leq N^*$ implies that the matrix on the right-hand side of the key equation (18) should be upper triangular, and hence that the matrix on the left-hand side, $G_{Eb,N,Z}G_{N|Z}^{-1}$, which is observed from the data, should also be upper-triangular. However, in practice, this matrix may not be upper-triangular. For our empirical results, we impose upper-triangularity on $G_{Eb,N,Z}G_{N|Z}^{-1}$ by setting all lower-triangular elements of the matrix to zero.[16]

Second, even after imposing upper-triangularity, it is still possible that the eigenvectors and eigenvalues could have negative elements, which is inconsistent with the interpretation of them as densities and probabilities.[17] When our estimate of the densities $g(b|N^*)$ took on negative values, our remedy was to set the density equal to zero, but normalize our density estimate so that the resulting density integrated to one.[18]

Results: Highway work auctions Figure 5 contains the graphs of the estimated densities $g(b|N^*)$ for $N^* = 3,4,5$, for the highway work auctions. In each column of this table, we present three estimates of each $g(b|N^*)$: (i) the normalized estimate with the negative portions removed, labeled "trunc est"; (ii) the un-normalized estimate, which includes the negative values for the density, labeled "Orig est"; and (iii) the naïve estimate, given by $g(b|A)$. In each plot, we also include the 5% and 95% pointwise confidence intervals,

A_{-1} can be used in the role of Z.

[15] See Hong and Shum (2002, Appendix B.1) for more discussion of a model with implicit reserve prices, for this dataset.

[16] Indeed, in the Monte Carlo simulations, we sometimes also had to impose this on the simulated data, as the $G_{Eb,N,Z}G_{N|Z}^{-1}$ matrix could be non-upper triangular due to small sample noise. In a previous version of the paper, we also reported estimation results without imposing upper-triangularity, which required an alternative ordering condition (instead of Condition 5) to identify the column order of the eigenvector matrix $G_{N|N^*}$. The results were clearly inferior to those reported here, and so are omitted from this version.

[17] This issue also arose in our Monte Carlo studies, but went away when we increased the sample size.

[18] Here we follow the recommendation of Efromovich (1999, pg. 63).

calculated using bootstrap resampling.[19]

Figure 5 shows that the naïve bid density estimates, using A in place of N^*, overweights small bids, which is reminiscent of the Monte Carlo results. As above, the reason for this seems to be that the number of potential bidders N^* exceeds the observed number of bidders A. In the IPV framework, more competition drives down bids, implying that using A to proxy for the unobserved level of competition N^* may overstate the effects of competition. Because in this empirical application we do not know and control the data-generating process, these economically sensible differences between the naïve estimates (using $g(b|A)$) and our estimates (using $g(b|N^*)$) serve as a confirmatory reality check on the assumptions underlying our estimator.

For these estimates, the estimated $G_{A|N^*}$ matrix was

	$N^*=3$	$N^*=4$	$N^*=5$
$A=(1,2,3)$	1.0000	0.1490	0.2138
$A=4$	0	0.8510	0.4237
$A\geq 5$	0	0	0.3625

Furthermore, for the normalized estimates of the bid densities with the negative portions removed, the implied values for $E[b|N^*]$, the average equilibrium bids conditional on N^*, were 7.984, 7.694, 4.162 for, respectively, $N^*=3,4,5$ (in millions of dollars).

The corresponding valuation estimates, obtained by solving Eq. (3) pointwise in b using our bid density estimates, are graphed in Figure 6. We present the valuations estimated using our approach, as well as a naïve approach using $g(b|A)$ as the estimate for the bid densities. Note that the valuation estimates become negative within a low range of bids, and then at the upper range of bids, the valuations are decreasing in the bids, which violates a necessary condition of equilibrium bidding. These may be due to unreliability in estimating the bid densities $g(b|A)$ and $g(b|N^*)$ close to the bounds of the observed support of bids. Furthermore, in the estimated values for $N^*=3$ and $N^*=4$ in Figure (6), we see that the valuations rise steeply for low bids. This arises from the truncation procedure, which leads to a kink in the bid density at the point when the density changes from zero to a positive value.

[19]The asymptotic variance is derived analytically in Appendix A.2. However, it is tedious to compute in practice, which is why we use the bootstrap to approximate the pointwise variance of the density estimates. Note that in the normalized density estimates, the lower bound of the confidence interval is always zero, uniformly across all values of b.

Figure 5: Highway work projects

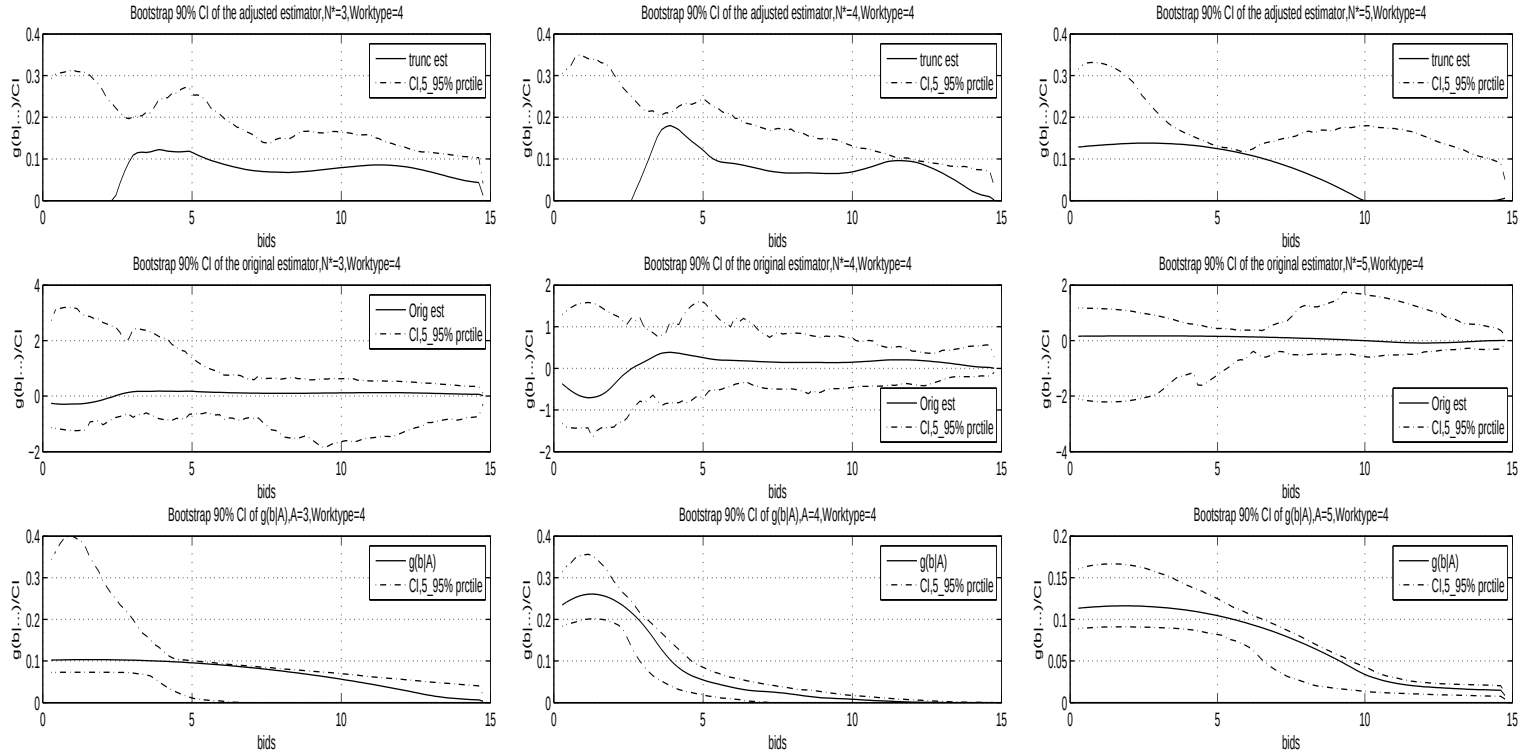
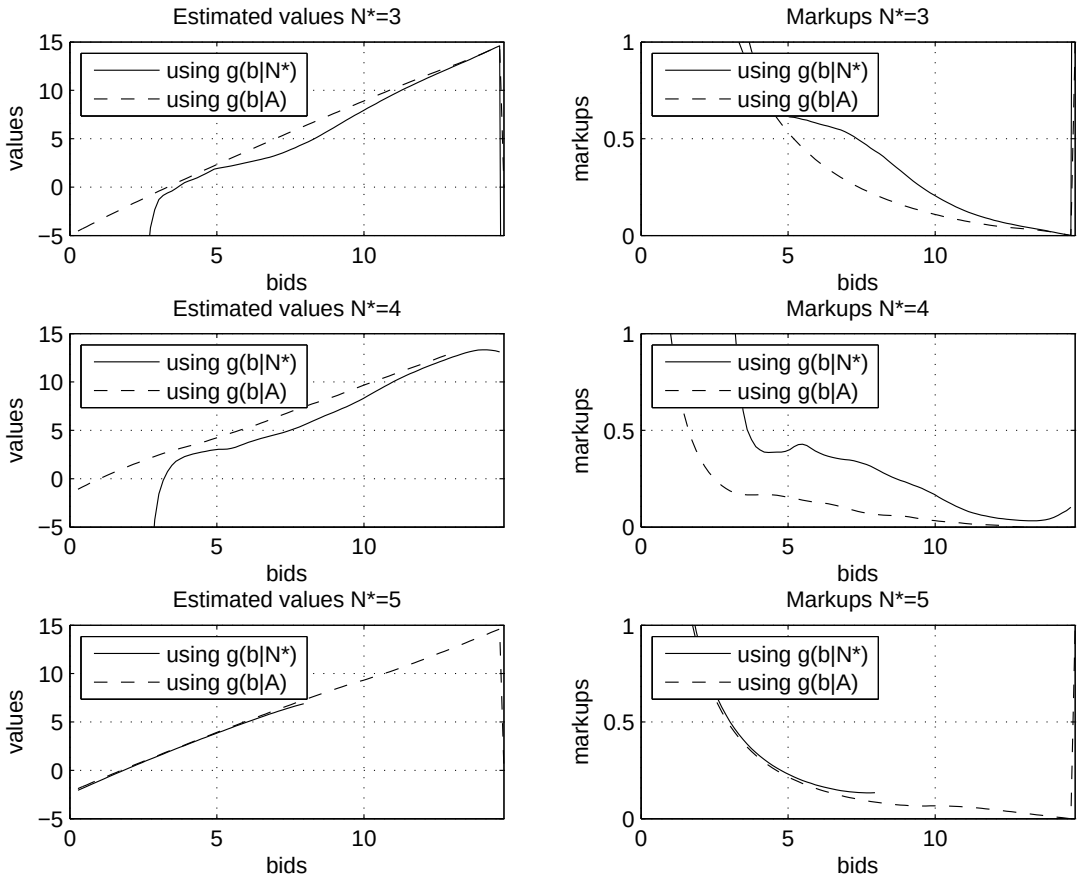


Figure 6: Highway work projects, estimated values



Comparing the estimates of valuations using $g(b|N^*)$, and those obtained using $g(b|A)$, we see that the valuations using $g(b|N^*)$ are smaller than those using $g(b|A)$, for $N^* = 3, 4$ (but virtually indistinguishable for $N^* = 5$). As in the Monte Carlo results, this implies that the markups $(b-c)/b$ are larger using our estimates of $g(b|N^*)$. The differences in implied markups between these two approaches is economically meaningful, as illustrated in the right-hand-side graphs in Figure (6). For example, for $N^* = 4$, at a bid of \$5 million, the corresponding markup using $g(b|A = 4)$ is around 15%, or \$750,000, but using $g(b|N^* = 4)$ is around 40%, or \$2 million. This suggests that failing to account for unobservability of N^* can lead the researcher to understate bidders' profit margins.

One shortcoming of these results is that we do not allow for auction-specific heterogeneity. To address this issue, we collected, for a subset of these auctions, some covariates measuring cost and locational factors associated with the contracts. Using an approach used previously in Haile, Hong, and Shum (2003) and Bajari, Houghton, and Tadelis (2006), among others, we control for the auction-specific heterogeneity by running a first-stage log-linear regression of bids on covariates. Under the assumption that equilibrium bids in an auction are multiplicatively separable into a common auction-specific component (which is a function of the covariates), and an idiosyncratic component which varies across bidders in an auction:

$$b_{it} = \exp(X_t'\beta)\exp(\tilde{b}_{it}), \quad \tilde{b}_{it} \perp X_t,$$

the residuals from this regression can be interpreted as "normalized" bids, which are comparable across auctions, and hence used in our estimation procedure. In Table 2, we present results from the first-stage bid regression. Bids are increasing in the cost components COST INDEX and TRAFFIC, as is expected.

In Figure 7, we present the bid density estimates obtained using the (exponentiated) residuals from the bid regression, and in Figure 8, we present the corresponding estimates of valuations and markups. The results are qualitatively quite similar to the earlier results, which were obtained assuming without controlling for auction-specific heterogeneity. One notable difference, evident in Figure 8, is that the naïve estimator leads to *higher* markups for the $N^* = 4$ case. This is evidence that the distributions of valuations $x|N^*$ are different across N^*, because if the distributions were identical across N^*, the naïve approach would tend to underestimate markups for all values of $N^* < K$ (as shown in the Monte Carlo results present earlier).

Table 2: Results from first-stage log-linear regression

Dependent variable: $\log b_{it}$		
Cost components:		
COST INDEX	1.073	(2.91)**
TRAFFIC	0.135	(10.63)**
Regions:		
GATEWAY	0.178	(3.25)**
SKYLANDS	0.213	(3.90)**
SHORE	-0.307	(5.41)**
DELAWARE	-0.038	(0.68)
SOUTH	0.185	(3.00)**
Constant	-0.304	(1.13)
Observations	3283	
R-squared	0.09	

Robust standard errors in parentheses.
Cost-related covariates: COST is a construction cost index obtained from the trade publication *Engineering News-Record*; TRAFFIC measures weekday traffic volume (in both directions) of the road being repaired. Regional covariates are dummy variables for geographic regions of New Jersey (with the excluded region being the Atlantic City region).

Figure 7: Estimated Bid Densities for Highway Work Projects, Controlling for Auction-Specific Heterogeneity

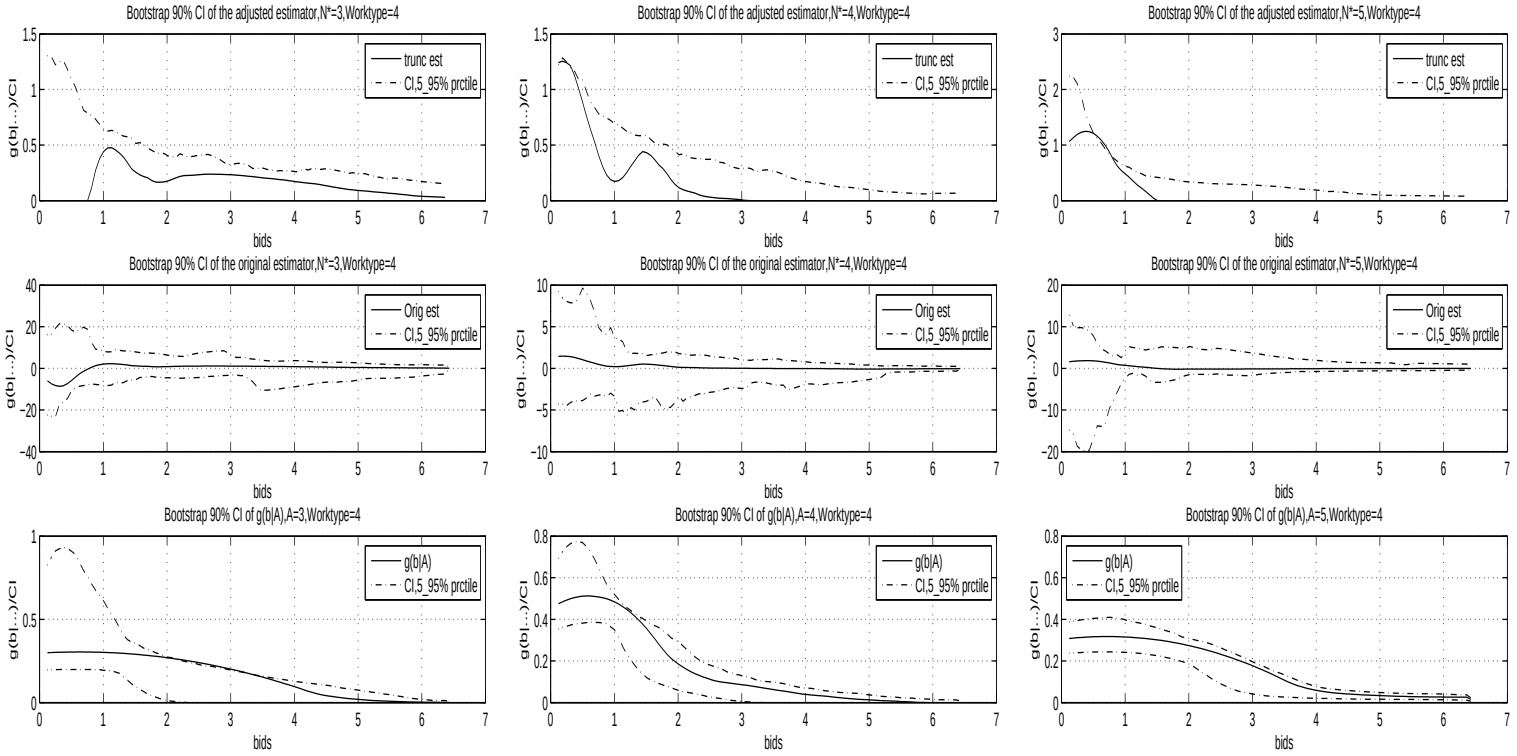
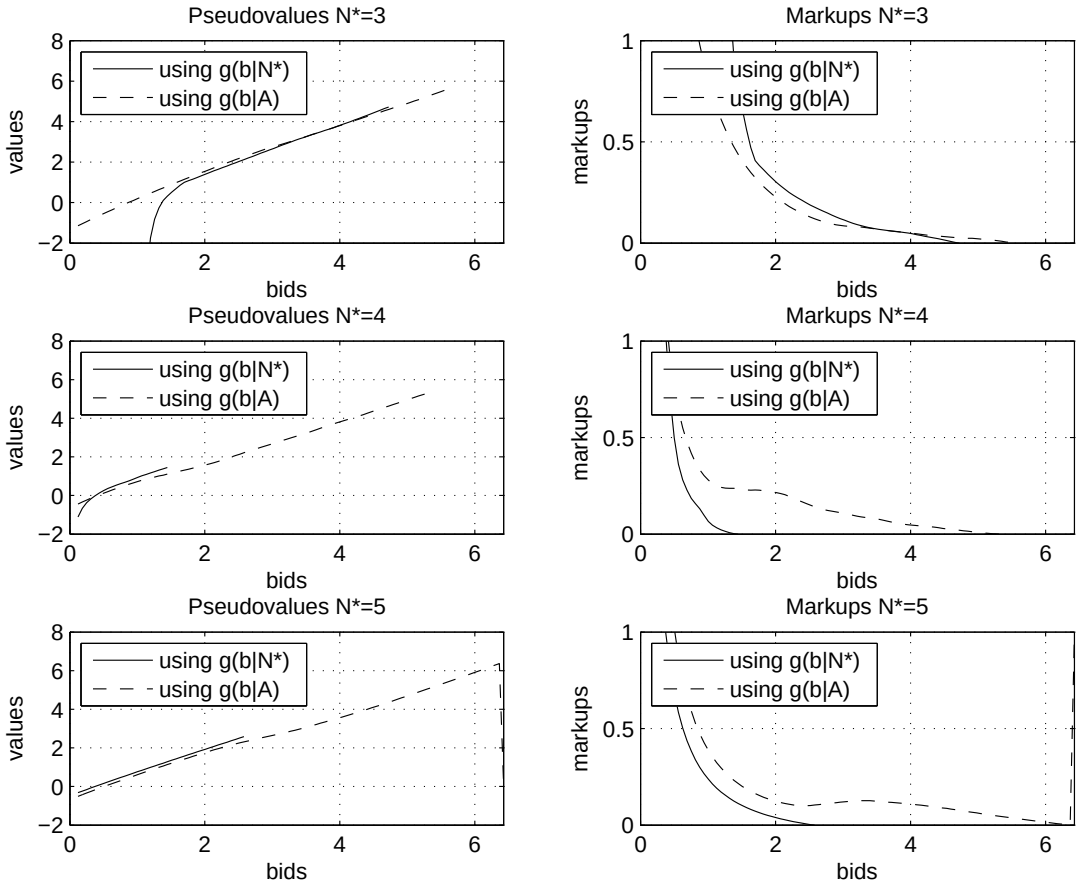


Figure 8: Estimated Valuations and Markups from Highway work projects, controlling for Auction-Specific Heterogeneity



6 Extension: Only Winning Bids are Recorded

In some first-price auction settings, only the winning bid is observed by the researcher. This is particularly likely for the case of descending price, or *Dutch* auctions, which end once a bidder signals his willingness to pay a given price. For instance, Laffont, Ossard, and Vuong (1995) consider descending auctions for eggplants where only the winning bid is observed, and van den Berg and van der Klaauw (2007) estimate Dutch flower auctions where only a subset of bids close to the winning bid are observed. Within the symmetric IPV setting considered here, Guerre, Perrigne, and Vuong (2000) and Athey and Haile (2002) argue that observing the winning bid is sufficient to identify the distribution of bidder valuations, provided that N^* is known. Our estimation methodology can be applied to this problem even when the researcher does not know N^*, under two scenarios.

First Scenario: Non-Binding Reserve Price In the first scenario, we assume that there is no binding reserve price, but the researcher does not know N^*. (Many Dutch auctions take place too quickly for the researcher to collect data on the number of participants.) Because there is no binding reserve price, the winning bid is the largest out of the N^* bids in an auction. In this case, bidders' valuations can be estimated in a two-step procedure.

In the first step, we estimate $g_{WB}(\cdot|N^*)$, the equilibrium density of winning bids, conditional on N^*, using the methodology above. In the second step, we exploit the fact that in this scenario, the equilibrium CDF of winning bids is related to the equilibrium CDF of the bids by the relation:

$$G_{WB}(b|N^*) = G(b|N^*)^{N^*}.$$

This implies that the equilibrium bid CDF can be estimated as $\hat{G}(b|N^*) = \hat{G}_{WB}(b|N^*)^{1/N^*}$, where $\hat{G}_{WB}(b|N^*)$ denotes the CDF implied by our estimates of $\hat{g}_{WB}(b|N^*)$. Subsequently, upon obtaining an estimate of $\hat{G}(b|N^*)$ and the corresponding density $\hat{g}(b|N^*)$, we can evaluate Eq. (3) at each b to obtain the corresponding value.

Second Scenario: Binding Reserve Price, but A Observed In the second scenario, we assume that the reserve price binds, but that A, the number of bidders who are willing to submit a bid above the reserve price, is observed. The reason we require A to be observed is that when reserve prices bind, the winning bid is not equal to $b^{N^*:N^*}$, the highest order statistic out of N^* i.i.d. draws from $g(b|N^*, b > r)$, the equilibrium bid distribution truncated to $[r, +\infty)$. Rather, for a given N^*, it is equal to $b^{A:A}$, the largest

out of A i.i.d. draws from $g(b|N^*, b > r)$. Hence, because the density of the winning bid depends on A, even after conditioning on N^*, we must use A as a conditioning covariate in our estimation.

For this scenario, we estimate $g(b|N^*, b > r)$ in two steps. First, treating A as a conditioning covariate, we estimate $g_{WB}(\cdot|A, N^*)$, the conditional density of the winning bids conditional on both the observed A and the unobserved N^*. Second, for a fixed N^*, we can recover the conditional CDF $G(b|N^*, b > r)$ via

$$\hat{G}(b|N^*, b > r) = \hat{G}_{WB}(b|A, N^*)^{1/A}, \ \forall A.$$

(That is, for each N^*, we can recover an estimate of $G(b|N^*, b > r)$ for each distinct value of A. Since the model implies that these distributions should be identical for all A, we can, in principle, use this as a specification check of the model.)

In both scenarios, we need to find good candidates for the auxiliary variables N and Z. Since typically many Dutch auctions are held in a given session, one possibility for N could be the total number of attendees at the auction hall for a given session, while Z could be an instrument (such as the time of day) which affects bidders' participation for a specific auction during the course of the day.[20]

7 Conclusions

In this paper, we have explored the application of methodologies developed in the econometric measurement error literature to the estimation of structural auction models, when the number of potential bidders is not observed. We have developed a nonparametric approach for estimating first-price auctions when N^*, the number of potential bidders, is unknown to the researcher, and varies in an unknown way among the auctions in the dataset. To our knowledge, our approach is the first solution to estimating such a model. Accommodating unknown N^* is also important for the policy implications of auction estimates, and the Monte Carlo and empirical results illustrate that ignoring the problem can lead to economically meaningful in differences the estimates of bidders' markups.

One maintained assumption in this paper that N^* is observed and deterministic from bidders' point of view, but not known by the researcher. The empirical literature has also

[20] This corresponds to the scenario considered in the flower auctions in van den Berg and van der Klaauw (2007).

considered models where the number of bidders N^* is stochastic and unobserved from the bidders' perspective: e.g., Athey and Haile (2002); Hendricks, Pinkse, and Porter (2003); Bajari and Hortacsu (2003); Li and Zheng (2006); and Song (2006). It will be interesting to explore whether the methods used here can be useful for estimating these models.

More broadly, these methodologies developed in this paper may also be applicable to other structural models in industrial organization, where the number of participants is not observed by the researcher. These could include search models, or entry models. We are considering these possibilities in future work.

References

Adams, C. (2007): "Estimating Demand from eBay Prices," *International Journal of Industrial Organization*, forthcoming.

Athey, S., and P. Haile (2002): "Identification of Standard Auction Models," *Econometrica*, 70, 2107–2140.

Athey, S., J. Levin, and E. Seira (2005): "Comparing Open and Sealed Bid Auctions: Theory and Evidence from Timber Auctions," working paper, Harvard University.

Bajari, P., and A. Hortacsu (2003): "Winner's Curse, Reserve Prices, and Endogenous Entry: Empirical Insights from eBay Auctions," *RAND Journal of Economics*, 34, 329–355.

Bajari, P., S. Houghton, and S. Tadelis (2006): "Bidding for Incomplete Contracts," working paper, University of Minnesota.

Efromovich, S. (1999): *Nonparametric Curve Estimation: Methods, Theory, and Applications*. Springer-Verlag.

Guerre, E., I. Perrigne, and Q. Vuong (2000): "Optimal Nonparametric Estimation of First-Price Auctions," *Econometrica*, 68, 525–74.

——— (2005): "Nonparametric Identification of Risk Aversion in First-Price Auctions Under Exclusion Restrictions," Manuscript, Pennsylvania State University.

Haile, P., H. Hong, and M. Shum (2003): "Nonparametric Tests for Common Values in First-Price Auctions," NBER working paper #10105.

Hall, P., I. McKay, and B. Turlach (1996): "Performance of Wavelet Methods for Functions with Many Discontinuities," *Annals of Statistics*, 24, 2462–2476.

Hendricks, K., J. Pinkse, and R. Porter (2003): "Empirical Implications of Equilibrium Bidding in First-Price, Symmetric, Common-Value Auctions," *Review of Economic Studies*, 70, 115–145.

Hong, H., and M. Shum (2002): "Increasing Competition and the Winner's Curse: Evidence from Procurement," *Review of Economic Studies*, 69, 871–898.

Hu, Y. (2007): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: a General Solution," *Journal of Econometrics*, forthcoming.

Krasnokutskaya, E. (2005): "Identification and Estimation in Highway Procurement Auctions under Unobserved Auction Heterogeneity," Manuscript, University of Pennsylvania.

Krasnokutskaya, E., and K. Seim (2005): "Bid Preference Programs and Participation in Highway Procurement Auctions," working paper, University of Pennsylvania.

Laffont, J. J., H. Ossard, and Q. Vuong (1995): "Econometrics of First-Price Auctions," *Econometrica*, 63, 953–980.

Li, T. (2005): "Econometrics of First Price Auctions with Entry and Binding Reservation Prices," *Journal of Econometrics*, 126, 173–200.

Li, T., I. Perrigne, and Q. Vuong (2000): "Conditionally Independent Private Information in OCS Wildcat Auctions," *Journal of Econometrics*, 98, 129–161.

——— (2002): "Structural Estimation of the Affiliated Private Value Acution Model," *RAND Journal of Economics*, 33, 171–193.

Li, T., and X. Zheng (2006): "Entry and Competition Effects in First-Price Auctions: Theory and evidence from Procurement Auctions," working paper, Vanderbilt University.

Mahajan, A. (2006): "Identification and Estimation of Regression Models with Misclassification," *Econometrica*, 74, 631–665.

Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden. North Holland.

Paarsch, H. (1997): "Deriving an Estimate of the Optimal Reserve Price: An Application to British Columbian Timber Sales," *Journal of Econometrics*, 78, 333–357.

Paarsch, H., and H. Hong (2006): *An Introduction to the Structural Econometrics of Auction Data*. MIT Press, with M. Haley.

Pagan, A., and A. Ullah (1999): *Nonparametric Econometrics*. Cambridge University Press.

Song, U. (2004): "Nonparametric Estimation of an E-Bay Auction Model with an Unknown Number of Bidders," working paper, University of British Columbia.

——— (2006): "Nonparametric Identification and Estimation of a First-Price Auction Model with an Uncertain Number of Bidders," working paper, University of British Columbia.

van den Berg, G., and B. van der Klaauw (2007): "If Winning Isn't Everything, Why do they Keep Score? A Structural Empirical Analysis of Dutch Flower Auctions," mimeo, Free University Amsterdam.

A Appendix: asymptotic properties of the two step estimator

A.1 Uniform consistency

In the first step, we estimate $\widehat{G}_{N|N^*}$ from

$$\widehat{G}_{N|N^*} := \psi\left(\widehat{G}_{Eb,N,Z}\widehat{G}_{N,Z}^{-1}\right), \tag{29}$$

where $\psi(\cdot)$ is an analytic function as mentioned in Hu (2007) and

$$\begin{aligned}\widehat{G}_{Eb,N,Z} &= \left[\frac{1}{T}\sum_t \frac{1}{N_j}\sum_{i=1}^{N_j} b_{it}\mathbf{1}(N_t = N_j, Z_t = Z_k)\right]_{j,k}, \\ \widehat{G}_{N,Z} &= \left[\frac{1}{T}\sum_t \mathbf{1}(N_t = N_j, Z_t = Z_k)\right]_{j,k}.\end{aligned}$$

We summarize the uniform convergence of $\widehat{G}_{N|N^*}$ as follows:

Lemma 6 *Suppose that* $Var(b|N,Z) < \infty$. *Then,*

$$\widehat{G}_{N|N^*} - G_{N|N^*} = O_p\left(T^{-1/2}\right).$$

Proof. It is straightforward to show that $\widehat{G}_{Eb,N,Z} - G_{Eb,N,Z} = O_p\left(T^{-1/2}\right)$ and $\widehat{G}_{N,Z} - G_{N,Z} = O_p\left(T^{-1/2}\right)$. As mentioned in Hu (2007), the function $\psi(\cdot)$ is an analytic function. Therefore, the result holds. ■

In the second step, we have

$$\widehat{g}(b|N^*) := e_{N^*}^T \left[\widehat{G}_{N|N^*}^{-1} \left(\widehat{G}_{b,N,Z}(b) \, \widehat{G}_{N,Z}^{-1} \right) \widehat{G}_{N|N^*} \right] e_{N^*},$$

where

$$\begin{aligned} \widehat{G}_{b,N,Z}(b) &= \left[\widehat{g}_{b,N,Z}(b, N_j, Z_k) \right]_{j,k}, \\ \widehat{g}_{b,N,Z}(b, N_j, Z_k) &= \frac{1}{Th} \sum_t \frac{1}{N_j} \sum_{i=1}^{N_j} K\left(\frac{b - b_{it}}{h} \right) \mathbf{1}(N_t = N_j, Z_t = Z_k). \end{aligned}$$

Let $\omega := (b, N, Z)$. Define the norm $\|\cdot\|_\infty$ as

$$\|\widehat{g}(\cdot|N^*) - g(\cdot|N^*)\|_\infty = \sup_b \left| \widehat{g}_{b|N^*}(b|N^*) - g_{b|N^*}(b|N^*) \right|.$$

The uniform convergence of $\widehat{g}(\cdot|N^*)$ is established as follows:

Lemma 7 *Suppose:*

(7.1) $\omega \in \mathcal{W}$ *and* $\mathcal{W}$ *is a compact set.*

(7.2) $g_{b,N,Z}(\cdot, N_j, Z_k)$ *is continuously differentiable to order* R *with bounded derivatives on an open set containing* $\mathcal{W}$*.*

(7.3) $K(u)$ *is differentiable of order* R*, and the derivatives of order* R *are bounded.* $K(u)$ *is zero outside a bounded set.* $\int_{-\infty}^{\infty} K(u)du = 1$*, and there is a positive integer* m *such that for all* $j < m$*,* $\int_{-\infty}^{\infty} K(u)u^j du = 0$*. The characteristic function of* K *is absolutely integrable.*

(7.4) $h \to 0$ *and* $nh \to \infty$*, as* $n \to \infty$*.*

Then, for all j*,*

$$\|\widehat{g}(\cdot|N^*) - g(\cdot|N^*)\|_\infty = O_p\left[\left(\frac{T}{\ln T} h^{1+2R} \right)^{-1/2} + h^m \right]. \tag{30}$$

The most important assumption for Lemma 7 is (7.2), which places smoothness restrictions on the joint density $g(b, N, Z)$. Via Eq. (7), this distribution is a mixture of conditional distributions $g(b|N^*)$, which possibly have a different support for different N^*. When the supports of $g(b|N^*)$ are known, condition (7.2) only requires the smoothness of $g(b|N^*)$ on its own support $[r, u_{N^*}]$ because the distribution $g(b|N, Z)$ can be estimated piecewise on $[r, u_2], [u_2, u_3], ..., [u_{K-1}, u_K]$. When the supports of $g(b|N^*)$ are unknown, condition (7.2)

would require that the density $g(b|N^*)$ for each value of N^* to be smooth at the upper boundary.[21]

Proof. By Lemma 6, it is straightforward to show that

$$\widehat{g}(b|N^*) = e_{N^*}^T \left[G_{N|N^*}^{-1} \left(\widehat{G}_{b,N,Z}(b) G_{N,Z}^{-1} \right) G_{N|N^*} \right] e_{N^*} + O_p\left(T^{-1/2}\right).$$

In order to show the consistency of our estimator $\widehat{g}(b|N^*)$, we need the uniform convergence of $\widehat{g}_{b,N,Z}(\cdot, N_j, Z_k)$. The kernel density estimator has been studied extensively. Following results from Lemma 8.10 in Newey and McFadden (1994), we have for all j and k

$$\sup_b |\widehat{g}_{b,N,Z}(\cdot, N_j, Z_k) - g_{b,N,Z}(\cdot, N_j, Z_k)| = O_p\left[\left(\frac{T}{\ln T} h^{1+2R} \right)^{-1/2} + h^m \right]. \tag{31}$$

The uniform convergence of $\widehat{g}_{b|N^*}$ then follows. ■

Remark: Another technical issue pointed out in Guerre, Perrigne, and Vuong (2000) is that the density $g(b|N^*)$ may not be bounded at the lower bound of its support, which is the reserve price r. They suggest using the transformed bids $b_\dagger \equiv \sqrt{b-r}$. Our identification and estimation procedures remain the same if b replaced by $b_\dagger$, where an estimate of the reserve price r could be the lowest observed bid in the dataset (given our assumption that the reserve price is fixed in the dataset).

A.2 Asymptotic Normality

In this section, we show the asymptotic normality of $\widehat{g}(b|N^*)$ for a given value of b. Define $\gamma_0(b) = vec\{G_{b,N,Z}(b)\}$, a column vector containing all the elements in the matrix $G_{b,N,Z}(b)$. Similarly, we define $\widehat{\gamma}(b) = vec\left\{\widehat{G}_{b,N,Z}(b)\right\}$. The proof of Lemma 7 suggests that

$$\widehat{g}(b|N^*) = \varphi\left(\widehat{\gamma}(b)\right) + O_p\left(T^{-1/2}\right)$$

where

$$\varphi\left(\widehat{\gamma}(b)\right) \equiv e_{N^*}^T \left[G_{N|N^*}^{-1} \left(\widehat{G}_{b,N,Z}(b) G_{N,Z}^{-1} \right) G_{N|N^*} \right] e_{N^*}.$$

[21] In ongoing work, we are exploring alternative methods, based on wavelet methods (eg. Hall, McKay, and Turlach (1996)), to estimate the joint density $g(b, N, Z)$ when there are unknown points of discontinuity, which can be due to the non-smoothness of the individual densities $g(b|N^*)$ at the upper boundary of their supports.

Notice that the function $\varphi(\cdot)$ is linear in each entry of the vector $\widehat{\gamma}(b)$. Therefore, we have

$$\widehat{g}(b|N^*) - g(b|N^*) = \left(\frac{d\varphi}{d\gamma}\right)^T (\widehat{\gamma}(b) - \gamma_0(b)) + o_p\left(1/\sqrt{Th}\right),$$

where $\frac{d\varphi}{d\gamma}$ is nonstochastic because it is a function of $G_{N|N^*}$ and $G_{N,Z}$ only. The asymptotic distribution of $\widehat{g}(b|N^*)$ then follows that of $\widehat{\gamma}(b)$. We summarize the results as follows:

Lemma 8 *Suppose that assumptions in Lemma 7 hold with $R = 2$ and that*

1. *there exists some δ such that $\int |K(u)|^{2+\delta} du < \infty$,*
2. *$(Th)^{1/2} h^2 \to 0$, as $T \to \infty$.*

Then, for a given b and N^*,

$$(Th)^{1/2} [\widehat{g}(b|N^*) - g(b|N^*)] \xrightarrow{d} N(0, \Omega),$$

where

$$\begin{aligned} \Omega &= \left(\frac{d\varphi}{d\gamma}\right)^T V(\widehat{\gamma}) \left(\frac{d\varphi}{d\gamma}\right), \\ V(\widehat{\gamma}) &= \lim_{T\to\infty} (Th) E\left[(\widehat{\gamma} - E(\widehat{\gamma}))(\widehat{\gamma} - E(\widehat{\gamma}))^T\right]. \end{aligned}$$

Proof. As discussed before Lemma 8, the asymptotic distribution of $\widehat{g}(b|N^*)$ is derived from that of $\widehat{\gamma}(b)$. In order to prove that the asymptotic distribution of the vector $\widehat{\gamma}(b)$ is multivariate normal $N(0, V(\widehat{\gamma}))$, we show that the scalar $\lambda^T \widehat{\gamma}(b)$ for any vector λ has a normal distribution $N\left(0, \lambda^T V(\widehat{\gamma}) \lambda\right)$. For a given value of b, it is easy to follow the proof of Theorems 2.9 and 2.10 in Pagan and Ullah (1999) to show that

$$(Th)^{1/2} \left[\lambda^T \widehat{\gamma}(b) - \lambda^T \gamma_0(b)\right] \xrightarrow{d} N\left(0, Var\left(\lambda^T \widehat{\gamma}(b)\right)\right),$$

where $Var\left(\lambda^T \widehat{\gamma}(b)\right) = \lambda^T V(\widehat{\gamma}(b)) \lambda$ is the variance of the scalar $\lambda^T \widehat{\gamma}(b)$. The asymptotic distribution of $\widehat{g}(b|N^*)$ then follows. ■