The Econometrics of Auctions with Asymmetric Anonymous Bidders: how to test symmetry ?

Laurent Lamy

Paris School of Economics

Outline

- A short summary of the structural econometrics literature of auction data with a special emphasis on how to deal with **incomplete data**
- then I will consider **anonymous data** which is the specific form of incomplete data I am interested in.
- I will present mostly the unpublished parts (and related research) of a previous working paper entitled "The Econometrics of Auctions with Asymmetric Anonymous Bidders", part of which just appeared under the same title in the Journal of Econometrics.
- More precisely: the part on testing procedures.
- I will talk a bit about identification (but not on estimation which is the main part of the published paper).

Pure private value auction model

A model is a pair $(\mathbb{F}, \gamma)$, where $\mathbb{F}$ is a family of joint distributions (corresponding to the latent random private values), γ is a mapping (capturing the equilibrium prediction) $\gamma : \mathbb{F} \longrightarrow \mathbb{H}$ where $\mathbb{H}$ is a family of joint distributions (corresponding to the observable bids).
First issue : can we identify (nonparametrically) private values from the bids ? In other words: is the function γ injective ?

- Estimation of F (identification can be vacuous if estimators performs very poorly as it can be the case with nonparametric setup due to the curse of dimensionality)
- Tests on F (e.g., are some bidders symmetric ? or are some bidders stronger than others ?)
- Testing the model itself (is the function γ surjective ?)

Two kinds of difficulties

- The link between bids and values, i.e. we may not observe bidders values directly as in first-price auctions. Seems the most problematic point for identification but it is not necessarily the case (however it can be highly problematic for estimation)
- In second-price or English (ascending) auctions, bids are equal to values but we often do not observe the full set of bids: this is the **incomplete data** problem which raises identification issues.

Next, I consider only second-price auction to focus solely on this last aspect.

Some practical example

In many environment we do not observe all bids and all identities:

- In open auction (intrinsic incompleteness): in the Dutch auction we just observe the winning bid, in the English auction the bid that the winner would have made is not observed.
- In some auction data set (especially procurement), only the two highest bids are disclosed. Furthermore, we may not observe also the number of participant

First lesson: we often observe only a limited set of order-statistics of the bids.

- Bidders identities may be not recorded in the data set: confidential data or lost. Some bids remains confidential: on eBay the bid of the winner is not disclosed!
- the identities of non-winning bidders are often not disclosed to weaken the sustainability of cooperative agreements. (French timber auctions by ONF but also typically the case in daily auctions e.g. in electricity markets)
- Bids may be structurally anonymous due to the vacuous nature of bidders' identities, e.g. in internet auctions. (if the seller bids on his own item, it will be through a false-name since it is forbidden)

Second lesson: bids are often anonymous

Identification of the symmetric IPV model in the second price auction: some (basic) results

IPV: bidders valuations are i.i.d.

- If the number of bidders is fixed. The model is identified from a single order-statistic $B^{i:n}$.

$$F_{B^{i:n}}(y) = \frac{n!}{(n-i)!(i-1)!} \int_0^{F_B(y)} u^{n-i}(1-u)^{i-1} du$$

- If the number of bidders is unknown and stochastic. The model is identified from two order-statistic (Song, 2004). Let $k_2 > k_1$ the order statistic that we observe, then denote by $p_{k_2|k_1}(y|x)$ be the density of the k_1^{th} highest value Y conditional on X the k_2^{th} highest value. It corresponds to the k_1 highest draw among $k_2 - 1$ variables that are drawn according to the distribution $\frac{F(y)-F(x)}{1-F(x)}$ on $[x, \overline{x}]$. We are back to the previous case.

Identification of the asymmetric IPV model in the second price auction: some results

AIPV: bidders valuations are drawn independently but may be asymmetric.

- If the set of bidders is known (and bidders supports are the same) then the model is identified from a single order-statistic (useful for Dutch auctions e.g.).
- This is an application of the competing risk problem (Meilijson 81, Journal of Applied Probability). A machine is made of n components $i = 1, \ldots, , n$. The lifetime of each component is X_i is not observable to the statistician. Let W be the lifetime of the machine and I the set of component that are broken at time W (the result of the autopsy of the machine).

Illustration: the English auction

Let $F^j_{2:n}$ the CDF of the price conditional on the winner of the auction being j. For $j = 1, \cdots, n$, those CDFs are observed.

- We obtain a system of 'Pfaffian integral equations' : for any $j \in N$

$$\prod_{i \neq j} F_i(w) = \int_0^w (1 - F_j(u))^{-1} dF^j_{2:n}(u)$$

- This system has a unique solution.
- Identification from a single order-statistic relies crucially on the fact that bids are non-anonymous !
- Comment: difficulty to estimate the lower tail of the bid distribution if we observe only top order statistics.

Identification of the asymmetric IPV model in the second price auction: anonymous data

- **My main identifiability result:** the asymmetric IPV model is identified from the full vector of bids without observing the identities of the bidders.
- A non linear inverse problem (non standard a priori?)

Preliminary intuition: The two coins example.

Consider two different coins with probability p_1, p_2 for Head ($(1-p_1)$, $(1-p_2)$ for Tail) respectively. Consider we observe a infinite sequence of joint realizations where each coin realization is supposed to be independent from the other one.

- We observe $p(H,H)$, $p(T,T)$ and $p(H,T)$ with the constraint that $p(H,H)+p(T,T)+p(H,T)=1$
- By independence, we have: $p(H,H)=p_1 \cdot p_2$ and $1+p(H,H)-p(T,T)=p_1+p_2$
- A non-linear system of two equations-two unknowns.
- The system has a unique solution: the roots of the polynomial $X \rightarrow a \cdot X^2 + b \cdot X + c$ where $a=1$, $b=1+p(H,H)-p(T,T)$ and $c=p(H,H)$
- the model is thus identified.

The asymmetric IPV model is identified: two reparametrizations.

Notation: let $F_B^{k:p}$ the CDF of the k order statistic of a vector of p elements taken at random among the n elements taken from the vector of bids B. E.g. $F_B^{k:n}$ is the CDF of k^{th} order statistic. Let F_{B_i} the CDF of bidder i values.

Under independence, we have the following nonlinear system of n equations with n unknowns:

$$\begin{aligned}
F_{\mathbf{B}}^{(1:1)}(b) &= \frac{1}{n} \cdot && \sum_{i=1}^{n} F_{\mathbf{B_i}}(b) \\
F_{\mathbf{B}}^{(2:2)}(b) &= \frac{1}{n(n-1)} \cdot && \sum_{i_1, i_2, i_1 \neq i_2} F_{\mathbf{B_{i_1}}}(b) \cdot F_{\mathbf{B_{i_2}}}(b) \\
&.. && . \\
F_{\mathbf{B}}^{(r:r)}(b) &= \frac{1}{n(n-1)\cdots(n-r+1)} \cdot && \sum_{i_1, \cdot, i_r, i_k \neq i_{k'},\ i_k \in \{i_1, \cdot, i_n\}} \prod F_{\mathbf{B_{i_k}}}(b) \\
&.. && . \\
F_{\mathbf{B}}^{(n:n)}(b) &= \frac{1}{n!} \cdot && \sum_{i_1, \cdot, i_n, i_k \neq i_{k'},\ i_k \in \{i_1, \cdot, i_n\}} \prod F_{\mathbf{B_{i_k}}}(b)
\end{aligned} \tag{1}$$

where $F_{\mathbf{B}}^{(k:k)}(b)$ is obtained by a recursive use of the formula (valid for exchangeable variables)

$$\frac{n-r}{n} F_{\mathbf{B}}^{(r:n)}(u) + \frac{r}{n} F_{\mathbf{B}}^{(r+1:n)}(u) = F_{\mathbf{B}}^{(r:n-1)}(u), \forall u, r \leq n-1 \tag{2}$$

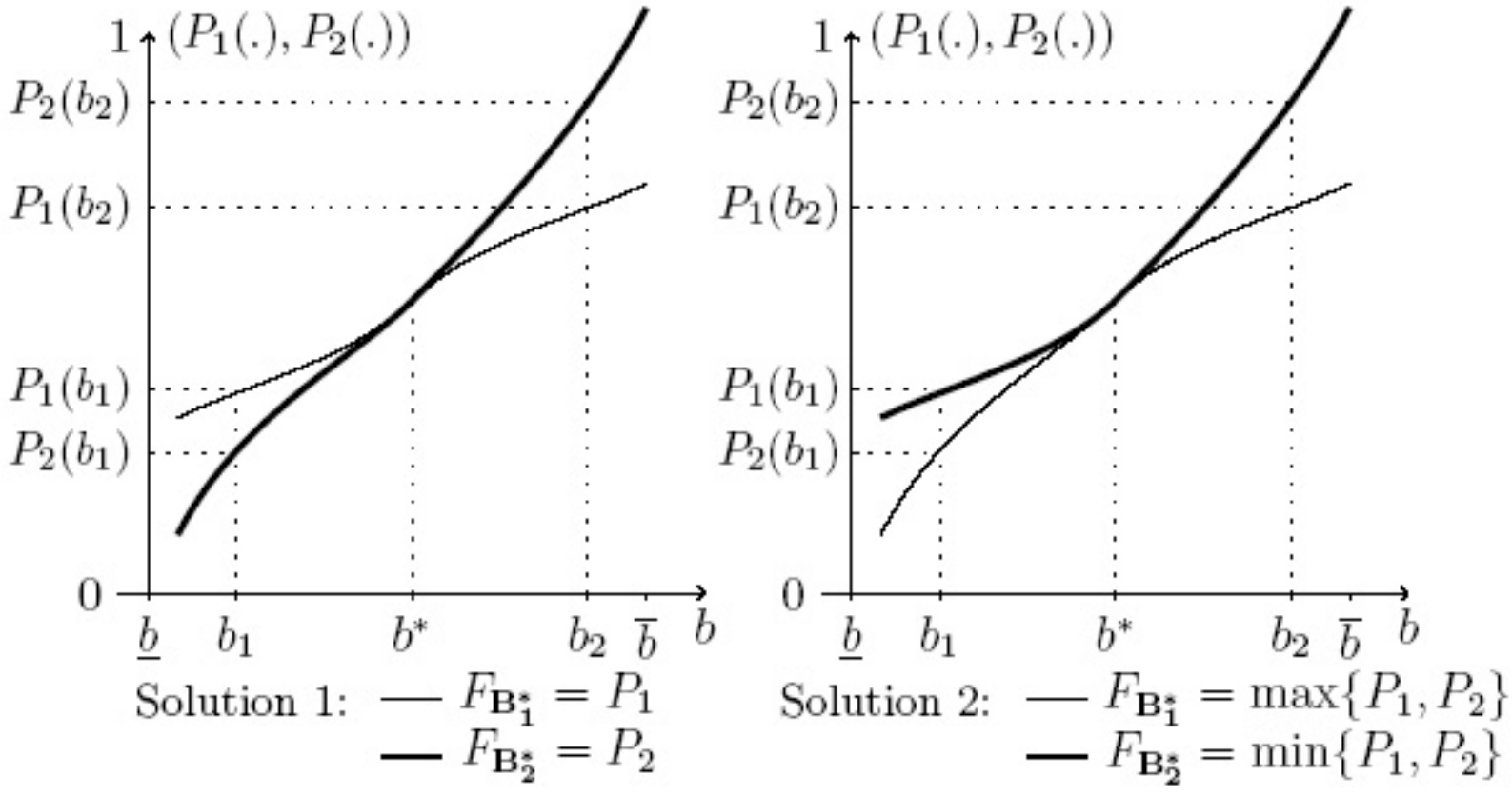
The two reparametrizations.

- The second one is linear. The corresponding matrix is triangular with positive number on the diagonal. Thus invertible [and also differentiable]
- The first reparametrization is non linear. Nevertheless it is invertible [and differentiable only on points where "bidders fully asymmetric"].
  $(F_{\mathbf{B_i}}(b))_{i=1,\cdots,n} = \Upsilon(c_n.F_{\mathbf{B}}^{(n:n)}(b), \cdots, c_1.F_{\mathbf{B}}^{(1:1)}(b))$ where Υ corresponds to the bijection between the n roots or a monic polynomial of degree n and its n coefficients.
- Let $\Upsilon : [0,1]^n \to \mathbb{Z}^n$ be the function such that $(\omega_1, \cdots, \omega_n) = \Upsilon(a_0, \cdots, a_{n-1})$ (where $\omega_1 \geq \cdots \geq \omega_n$) is the ordered vector of the roots (possibly complex number) counted with their order of multiplicity of the polynomial $Q(u) = u^n + \sum_{i=0}^{n-1} a_i \cdot (-1)^{n-i} u^i$, i.e. $Q(u) = \prod_{i=1}^{n} (u - \omega_i)$.

The asymmetric IPV model is identified.

- We have proved that for any bid b, the vector $F_{B_i}(b)$, $i \in [1, n]$ is identified up to a permutation (generalization of the result with two coin to n coins). **Furthermore, it gives a path for estimation**
- If the maps F_{B_i} do not cross and since those map have to be continuous, then there is a unique solution up to a permutation for the CDFs F_{B_i}, $i \in [1, n]$
- In general, i.e. with some intersections (or crossing points), the observation of the CDFs $F_B^{(k:n)}$, $k \in [1, n]$ is not sufficient to recover the CDFs F_{B_i}, $i \in [1, n]$ in a unique way.
- But if the joint vector of bids, F_B, is observed then we can distinguish the different solution generated by the intersections.

Figure 1: Identification of the asymmetric IPV model, $n = 2$



How to test the symmetry structure ?

Why ?: It is a key output before running the estimation procedures that I develop since they perform badly if some bidders are symmetric and that we do not take into account the symmetry (because the Jacobian of Υ is non invertible). **Testing is a key step before estimation !**

- A polynomial with real roots $P(X)$ of degree n has the root structure $(k_1, \cdots, k_{r(P)})$ where $\sum_{i=1}^{r(P)} k_i = n$ and $k_1 \geq \cdots \geq k_{r(P)} \geq 1$ if $P(X) = \prod_{i=1}^{r(P)} (X - x_i)^{k_i}$ for some $\{x_i\}_{i=1,\cdots,r(P)}$ such that $x_i \neq x_j$ for all i, j. The integer $r(P)$ is the number of distinct roots.
- there is a literature on applied math on testing roots (but without any statistical perspective). We will briefly recall useful elements from this literature and then put statistics on top of it.

Some elements on generalized discriminants

The Discrimination matrix of the monic polynomial $Q = X^p + \sum_{i=0}^{p-1} a_i.X^i$ is the $(2p+1) \times (2p+1)$ matrix:

$$Discr(Q) = \begin{bmatrix}
1 & a_{p-1} & a_{p-2} & \cdots & a_0 & & & & \\
0 & p & (p-1)a_{p-1} & \cdots & a_1 & & & & \\
 & 1 & a_{p-1} & \cdots & a_1 & a_0 & & & \\
 & 0 & p & \cdots & 2.a_2 & a_1 & & & \\
 & & & \cdots & \cdots & & & & \\
 & & & \cdots & \cdots & & & & \\
 & & & & 1 & a_{p-1} & \cdots & a_0 & \\
 & & & & 0 & p & \cdots & a_1 & \\
 & & & & & 1 & a_{p-1} & \cdots & a_0
\end{bmatrix}.$$

Some elements on generalized discriminants

- For $k \in [1, n]$, let $\Delta(P, k)$ denote the determinant of the submatrix formed by the first $2k$ rows and the first $2k$ columns of $Discr(P)$. The numbers $\Delta(P, k)$ are also called generalized discriminants.
- Denote by $P^{(i)}$ the i^{th} derivative of the polynomial P (with $P^{(0)} = P$).
- Note that the generalized discriminants $\Delta(P^{(i)}, k)$ are polynomial functions of the coefficients of the primitive polynomial P.

Lemma 1: Corollary of Theorem 2.1 in Yang (J. Symbolic Computation)

A polynomial with real roots P has $r(P)$ distinct real roots if and only if $\Delta(P,k) > 0$ for $k \leq r(P)$ and $\Delta(P,k) = 0$ for $k > r(P)$

Lemma 2

The polynomial P has the root structure $(k_1, \cdots, k_{r(P)})$ if and only if the number of distinct roots of the polynomials $P^{(i)}$ is given by $\rho(i) = n - i - \sum_{j=1}^{r(P)} (k_j - 1 - i)^+$ for $i = 1, \cdots, n-2$.

Combining the two previous lemmas we obtain:

Proposition

A polynomial with real roots P has the root structure $(k_1, \cdots, k_{r(P)})$ if and only if, for any $i \in [0, n-2]$,

$$\begin{cases} \Delta(P^{(i)}, k) > 0 & \text{for } 1 \leq k \leq \rho(i) \\ \Delta(P^{(i)}, k) = 0 & \text{for } \rho(i) < k \leq n - i, \end{cases} \tag{3}$$

where $\rho(i) = n - i - \sum_{j=1}^{r(P)} (k_j - i - 1)^+$.

Statistical tests on bidders private values

Coming back to our framework, the probabilities $(F_{\mathbf{B}_i^*,\mathbf{Z}}(b,z))_{i=1,\cdots,n}$ are corresponding exactly to the n roots of the polynomial $P_{(b,z)}$ of degree n: $u \to \sum_{i=0}^{n} a_i(b,z) \cdot (-1)^{n-i} \cdot u^i$, where $a_n(b,z) = 1$ and $a_i(b,z) = \frac{n(n-1)\cdots(i+1)}{(n-i)!} \cdot F_{\mathbf{B},\mathbf{Z}}^{(n-i:n-i)}(b,z) \cdot (f_{\mathbf{Z}}(z))^{n-i-1}$ for $i < n$. For a given bid b and a given set of covariates z, the root structure is characterized by the generalized discriminants $\Delta(P_{b,z}^{(i)}, k)$, which can be easily estimated by their sample analogs $\widehat{\Delta}(P_{b,z}^{(i)}, k) = \Delta(\widehat{P}_{b,z}^{(i)}, k)$ where $\widehat{P}_{b,z}$ is the sample analog of the polynomial $P_{b,z}$, i.e. with $a_i(b,z)$ being replaced (for $i < n$) by $\widehat{a}_i(b,z) = \frac{n(n-1)\cdots(i+1)}{(n-i)!} \cdot \widehat{F}_{\mathbf{B},\mathbf{Z}}^{(n-i:n-i)}(b,z) \cdot (\widehat{f}_{\mathbf{Z}}(z))^{n-i-1}$.

Statistical tests on bidders private values

Then various testing statistics can be build to test for some underlying root structure. Popular examples are:

- Kolmogorov-Smirnov-type (KS) tests based on suprema, i.e. based on $Sup_{b,z}\widehat{\Delta}(P^{(i)}_{b,z}, k)$
- Tests based on means, i.e. based on weighted expectations of $\widehat{\Delta}(P^{(i)}_{b,z}, k)$.

My Monte Carlo simulations have shown that tests based on means outperforms Kolmogorov-Smirnov-type (KS) tests.

Testing symmetry

We develop next a test for full symmetry against the alternatives of some asymmetry. In this case, the discrimination system reduces to a single equation as stated below and can thus be easily tested with standard one-sided tests (on the contrary, we can not provide analytic formulas for tests involving multiple nonlinear inequality constraints).

Corollary

A polynomial with real roots P of degree n has the root structure (n) if and only if $\Delta(P, 2) = 0$. If P has some distinct roots, then $\Delta(P, 2) > 0$.

Note that we are allowing covariates.

Testing symmetry

$$H_0 \text{ (full symmetry)} : F_{\mathbf{B}_1^*,\mathbf{Z}}(.,.) = \cdots = F_{\mathbf{B}_n^*,\mathbf{Z}}(.,.)$$

$$H_1 \text{ (some asymmetry)} : F_{\mathbf{B}_i^*,\mathbf{Z}}(b,z) \neq F_{\mathbf{B}_j^*,\mathbf{Z}}(b,z),$$

for some i and j on a positive measure of b and z. The discriminant $\Delta(P,2)$ is equal to $n^2(n-1)\left((F^{(1:1)}_{\mathbf{B},\mathbf{Z}}(b,z))^2 - F^{(2:2)}_{\mathbf{B},\mathbf{Z}}(b,z)\cdot f_{\mathbf{Z}}(z)\right)$. From corollary [1], our testing hypothesis can be written as:

$$H_0 \text{ (full symmetry)} : \mathcal{H} = 0$$

$$H_1 \text{ (some asymmetry)} : \mathcal{H} > 0,$$

where $\mathcal{H} = \int\int \left((F^{(1:1)}_{\mathbf{B},\mathbf{Z}}(b,z))^2 - F^{(2:2)}_{\mathbf{B},\mathbf{Z}}(b,z)\cdot f_{\mathbf{Z}}(z)\right) dF^{(1:1)}_{\mathbf{B},\mathbf{Z}}(b,z)$. A straightforward calculation gives $\mathcal{H}$ as a function of $(F_{\mathbf{B_i}^*,\mathbf{Z}}(b,z))_{i=1,\cdots,n}$:

$$\mathcal{H} = \int\int \frac{1}{2n^2(n-1)}.\sum_{i=1}^{n}\sum_{j=1}^{n}\left(F_{\mathbf{B_i}^*,\mathbf{Z}}(b,z) - F_{\mathbf{B_j}^*,\mathbf{Z}}(b,z)\right)^2 d\left(\frac{1}{n}.\sum_{i=1}^{n} F_{\mathbf{B_i}^*,\mathbf{Z}}(b,z)\right).$$

Testing symmetry

The sample analog of $\mathcal{H}$ is

$$\widehat{\mathcal{H}} = \frac{1}{Ln} \sum_{t=1}^{L} \sum_{k=1}^{n} \left([\widehat{F}^{(1:1)}(X_k^t, Z^t)]^2 - \widehat{F}^{(2:2)}(X_k^t, Z^t).\widehat{f}_{\mathbf{Z}}(Z^t) \right).$$

After a tedious calculation with U-statistics:

Proposition

Suppose that $K_{F_{\mathbf{B_p}|\mathbf{Z}}}$ and $K_{f_{\mathbf{Z}}}$ are kernels and $h_{F_{\mathbf{B_p}|\mathbf{Z}}}$ and $h_{f_{\mathbf{Z}}}$ converges to zero as $L \to \infty$.

$$\sqrt{L}.\left(\widehat{\mathcal{H}} - \mathcal{H}\right) \to_d \mathcal{N}(0, \Sigma).$$

Under H_0, we have $\Sigma^2 = E_{Z^t}[[f(Z^t)]^4]/(45n(n-1))$. Without covariates, the expression of Σ^2 is reduced to $\frac{1}{45n(n-1)}$.

Testing symmetry

- Define the test statistic: $t = \frac{\sqrt{L}.\widehat{\mathcal{H}}}{\sqrt{\widehat{\Sigma^2}}}$, where $\widehat{\Sigma^2} = \frac{1}{45n(n-1)} \cdot \frac{1}{L} \sum_{l=1}^{L} [\widehat{f}(Z_l)]^4$ is the sample analog of Σ^2.
- $\widehat{\Sigma^2}$ is a consistent estimate of Σ^2 converging at the rate $\sqrt{L}$ with some covariates. Without covariates, Σ^2 is known.

Two main questions:

1. Are asymptotic approximations accurate for small data sets under H_0?
2. Has the test enough power to reject the null under H_1 ?

The answer is yes in both cases.

Testing symmetry

	$L = 40$			$L = 200$		
n	2	4	6	2	4	6
share of p-values $< 10\%$	0.13	0.13	0.12	0.11	0.11	0.10
share of p-values $< 5\%$	0.06	0.05	0.06	0.05	0.06	0.05

Table: Performance of the asymptotic version of the test based on the Means. 5000 replications for the simulated statistics.

Testing symmetry

Under the alternative H_1, we have no tractable asymptotic approximation for the standard deviation of the test statistic. However, the median of the test statistic coincides asymptotically with the mean which is known. Finally, we obtain the following corollary about the way to reach the power 50%.

Corollary

Asymptotically, our test reject symmetry with a probability greater than one half if and only if the variable $\mathcal{H}$ is greater than $\frac{q_{1-\alpha}\Sigma}{\sqrt{L}}$.

Equivalently, it says that for a given degree of asymmetry $\mathcal{H} > 0$, the necessary size L^* of the data to reject symmetry at the level α with probability at least one half is approximately $\left(\frac{q_{1-\alpha}\Sigma}{\mathcal{H}}\right)^2$. Without covariates, the expression simplifies to:

$$L^* = q_{1-\alpha}^2 \cdot \frac{4(n-1)}{45n} \cdot \left[\frac{1}{n^2} \cdot \sum_{i=1}^{n}\sum_{j=1}^{n} E\left[\left(F_{\mathbf{B_i^*}}(b) - F_{\mathbf{B_j^*}}(b)\right)^2\right]\right]^{-2},$$

where the expectation is for b distributed according to the CDF $\sum_{i=1}^{n} F_{\mathbf{B_j^*}}(.)/n$.

Application of the previous result

- To gauge, what kind of asymmetries are the most difficult to detect. Consider two kinds of bidders stang versus weak and a fixed number of bidders n.
- Let $k \in [1, n-1]$ be the number of Strong bidders. Then the term in bracket is equal to $\frac{2k(n-k)}{n^2} . \left(E[(F_S(b) - F_W(b))^2] \right)^{-2}$. As a function of k, this term is symmetric with respect to $k = n/2$: it is decreasing from 1 to $n/2$ and then increasing. The 'balanced' panel with $k = [n/2]$ is the best one to reject symmetry.
- Numerical application with data taken from Flambard and Perrigne (2006): we obtain $L^* = 190$ ($L^* = 115$) for the 5% level (10% level) when there are 3 weak and 3 strong bidders as in their dataset. Those figures do not vary much when we slightly perturbate the structure of the bidders. For 8 [resp. 4] bidders while 4 [resp. 2] being Strong bidders, we obtain $L^* = 199$ and $L^* = 120$ [$L^* = 171$ and $L^* = 104$] for the 5% and 10% levels.

Structure	1/1		1/2		1/3		2/2	
Range of L	40	200	40	200	40	200	40	2
Degree of asymetry								
Distribution, $\epsilon = \pm\frac{1}{2}$								
share of p-values $< 10\%$	0.20	0.44	0.16	0.31	0.15	0.21	0.16	0.
share of p-values $< 5\%$	0.12	0.31	0.09	0.19	0.08	0.12	0.09	0.
Distribution, $\epsilon = \pm\frac{3}{4}$								
share of p-values $< 10\%$	0.44	0.90	0.29	0.70	0.21	0.38	0.30	0.
share of p-values $< 5\%$	0.31	0.83	0.19	0.56	0.11	0.24	0.20	0.
Distribution, $\epsilon = \pm 1$								
share of p-values $< 10\%$	0.78	1.00	0.54	0.98	0.30	0.67	0.57	0.
share of p-values $< 5\%$	0.67	1.00	0.39	0.94	0.18	0.50	0.43	0.

Table: Monte Carlo Results. Test based on Means. 5000 replications for each experiment.

Maximum likelihood with categorical data

- At this stage, I have dealt only with continuous data.
- A similar analysis could be useful for categorical data (under categorical data, the model reduces to a finite set of parameter) and to understand how maximum likelihood performs.
- Anonymity is related to possible singularities....

Related Literature

Asymptotic properties of ML estimators for i.i.d. samples and related tests are well-known under some standard regularity assumptions (Aitchison and Silvey, Ann. Math. Stat. 58).
ML in Nonregular econometrics models

- Unbounded likelihood in finite normal-mixture model (Hathaway, Annals of Stat. 85) or in heteroscedastic regression models (Crisp and Burridge, Biometrika 94)
- Discontinuous likelihood in models with a jump in the conditional density (Smith, Biometrika 85, Chernozhukov and Hong, Econometrica 04)
- The true parameter is on the boundary of the parameter space: Constrained Statistical Inference (Silvapulle and Sen, Wiley 2004)
- The information matrix is singular (and the model is identifiable) in Stochastic Frontier Function Model (Lee, Econometric Theory 93). See also Sargan, Econometrica 83

A Maximum Likelihood perspective?

- In a regular problem, ML estimator is asymptotically unbiased and achieves the Cramer Rao lower bound. To test a 'richer' model 2 against another 1 by a difference of f parameter (degree of freedom), just take $2.[\lambda_2 - \lambda_1]$ and compare to the χ^2 statistic with f degrees of freedom.
- It seems a priori the best candidate for testing. In general, the big issue is its computational tractability.
- When the problem is not regular, ML estimators may perform badly and the related tests may not work....

Some notation

- A for asymmetric; S for symmetric, I for independent and C for correlated. We consider the AI, SI, AC and SC models.
- K is the number of categories. T is the sample size. n is the number of variables.
- The vector α labels the parameters of the underlying sampling scheme. E.g. in the AI model α_i^k is the theoretical probability that the variable i lie in the category k.
- The vector p labels the observed proportion in each realizations. E.g. under non-anonymous data, p_i^k is the observed proportion of realizations such that the variable i lies in the category k. Under anonymous data, $p^{(e_1,\cdots,e_K)}$ with $\sum_{k=1}^{K} e_k = n$ be the observed proportion of the draws with e_k variables in the category k.

The Likelihood under non-anonymous data: the AI model

$$\lambda_{AI}((\alpha_1^i, \cdots, \alpha_K^i)_{i=1,\cdots,n}) = T. \sum_{k=1}^{K} \sum_{i=1}^{n} p_i^k . \log[\alpha_k^i],$$

where $\alpha_K^i = 1 - \sum_{k=1}^{K-1} \alpha_k^i$ for all i.
The solution that maximizes the likelihood is immediately given by:

$$\widehat{\alpha}_k^i = p_k^i.$$

The estimated proportions under ML are corresponding to the empirical proportions.

The Likelihood under anonymous data

Let E be the set of the events $e = (e_1, \cdots, e_K)$ with $\sum_{k=1}^{K} e_k = n$ and denote by $Ano(K, n) + 1$ its cardinality. E is the set of feasible events.

$$\lambda_{AI}(\alpha) = T. \sum_{e \in E} p^e . \log[\sum_{\sigma \in \Sigma} \prod_{k=1}^{K} \frac{\prod_{j=1}^{e_k} \alpha_k^{j+\sigma(\sum_{s<k} e_s)}}{e_k!}],$$

where $\alpha_K^i = 1 - \sum_{k=1}^{K-1} \alpha_k^i$ for all i.
The sum in the Logarithm due to anonymity breaks the separability.
Standard algorithms may fail to find the maximum!

Some remarks

- The intuition is that we would like to set our estimator $\widehat{\alpha}$ such that:

$$\sum_{\sigma\in\Sigma}\prod_{k=1}^{K}\frac{\prod_{j=1}^{e_k}\widehat{\alpha}_k^{j+\sigma(\sum_{s<k}e_s)}}{e_k!}=p^e,\forall e\in E$$

- $Ano(K,n)$ is the dimension of the observable variables whereas we need to identify $n.(K-1)$ variables.
- We have $Ano(K,n)\geq n.(K-1)$ with equality only in the case $K=2$.
- Thus for $K>2$, the likelihood is 'overidentified' and maximum likelihood is hardly tractable.
- For $K=2$, the above intuition works and can be used to construct maximum-likelihood based estimator for $K>2$.

The Likelihood under anonymous data and $K = 2$

Let $p^j, j \in [0, n]$ be the observed proportion of the draws with j variables in the first category. Let α_i be the theoretical proportion of the first category for the variable i.

$$\lambda(\alpha) = T.[p^n.\log[\prod_{i=1}^{n} \alpha_i] + \cdots +$$

$$+ \cdots + p^k.\log[\sum_{i_1,\cdot,i_r,i_k \neq i_{k'}} \prod_{i_k \in \{i_1,\cdot,i_n\}} \alpha_{i_k}. \prod_{i_k \notin \{i_1,\cdot,i_n\}} (1 - \alpha_{i_k})] + \cdots +$$

$$+ \cdots + p^1.\log[\sum_{i=1}^{n} \alpha_i \prod_{j \neq i} (1 - \alpha_j)] + p^0.\log[\prod_{i=1}^{n} (1 - \alpha_i)]]$$

The Likelihood in general with 2 categories

The information matrix $I_\alpha = \{a_{ij}\}_{1 \leq i,j \leq n}$ where $a_{ij} = [\frac{\partial \lambda(\alpha, p)}{\partial \alpha_i . \alpha_j}]_{p=p_\alpha}$ is singular if $\alpha_i = \alpha_j$. Is it immediate from the symmetry of the likelihood function such that the scores of linearly dependent. More generally, the information matrix has the same rank as the VanderMonde matrix with the coefficients $\{\alpha_1, \cdots, \alpha_n\}$.

Proposition

The information matrix I_α is of rank k where k is the number of distinct elements in $\{\alpha_1, \cdots, \alpha_n\}$. In case of the strict asymmetric model, the information matrix is non-singular.

With a suitable reparametrization of the likelihood (given that we know the symmetry structure), we can avoid the singularity.

We would like to choose the ML-estimator $\widehat{\alpha}_i$ such that:

$$\prod_{i=1}^{n} \widehat{\alpha}_i = p^n$$

$$..$$

$$..$$

$$\sum_{i_1,\cdot,i_r,i_k \neq i_{k'}} \prod_{i_k \in \{i_1,\cdot,i_n\}} \widehat{\alpha}_{i_k} . \prod_{i_k \notin \{i_1,\cdot,i_n\}} (1 - \widehat{\alpha}_{i_k}) = p^k \qquad (4)$$

$$..$$

$$\sum_{i=1}^{n} \widehat{\alpha}_i . \prod_{j \neq i} (1 - \widehat{\alpha}_j) = p^1$$

It guarantees that $\prod_{i=1}^{n} (1 - \widehat{\alpha}_i) = p^0$.

The issue

- Does a feasible solution with $\widehat{\alpha}_i \in [0, 1]$ exists?
- If yes, it is the ML-estimor and the maximum of the likelihood is equal to

$$\sum_{j=0}^{n} p^j . \log[p^j]$$

- Can we 'easily' solve the nonlinear system (tractable)? Is there a unique solution (identifiability)?

First Transformation (linear)

- Let $\alpha^{r:m}$ denotes the probability to have r variables in the first category in a random draw of m variables among the n variables for the random process with α.
- The left-hand terms of the previous system are corresponding exactly to $\alpha^{r:n}$ for $r = 1, \cdots, n$.
- We have the following recursive relationship (similar to the one for CDF derived in Athey and Haile, Econometrica 2002, p.2128)

$$\frac{m-r}{m}\alpha^{(r:m)}(u)+\frac{r+1}{m}\alpha^{(r+1:m)}(u) = \alpha^{(r:m-1)}(u), \forall u, r \leq m-1 \tag{5}$$

- Thus there is a bijection H such that $(\alpha^{i:i})_{i=1,\cdots,n} = H[(\alpha^{i:n})_{i=1,\cdots,n}]$ (linear equation with a triangular matrix)

First Transformation (linear)

- Let $(p^{i:i})_{i=1,\cdots,n} = H[(p^i)_{i=1,\cdots,n}]$. It corresponds to the 'estimated' probability to have r variables in the first category in a random draw of m variables among the n variables.
- The previous system of equation is then equivalent to:

$$\begin{aligned}
\prod_{i=1}^{n} \widehat{\alpha}_i &= p^{n:n} \\
&.. \\
&.. \\
\sum_{i_1,\cdot,i_r,i_k \neq i_{k'}} \prod_{i_k \in \{i_1,\cdot,i_r\}} \widehat{\alpha}_{i_k} &= p^{k:k} \qquad (6) \\
&.. \\
\sum_{i=1}^{n} \widehat{\alpha}_i &= p^{1:1}
\end{aligned}$$

Second Transformation (nonlinear)

- From the definition of $\alpha^{i:i}$, we also have:

$$\prod_{i=1}^{n}(X-\alpha_i)=\sum_{i=1}^{n}\alpha^{i:i}.(-1)^{n-i}.X^i$$

- Conclusion: the previous system of equation is thus equivalent to $(\widehat{\alpha}_i)_{i=1,\cdots,n}$ being the n-roots of the polynomial $X \to \sum_{i=1}^{n} p^{i:i}.(-1)^{n-i}.X^i$.

Two possibilities:

- All the roots are real numbers. Then the job is done provided that all the roots belongs to $[0,1]$. It is guaranteed by the fact that $p^{1:1} \geq p^{2:2} \geq \cdots \geq p^{n:n}$ which is easily verified.
- Some of the roots are complex number. Then the maximum of the likelihood is reached at a boundary solution, i.e. a solution with some multiple roots.

The reparametrized likelihood maximization in the AI model

- $$\lambda((\alpha^{i:n})_{i=1,\cdots,n}) = \sum_{j=0}^{n} p^j . \log[\alpha^{j:n}]$$

  with the constraint that the roots of the polynomial $X \to \sum_{i=1}^{n} \alpha^{i:i}.(-1)^{n-i}.X^i$ are all real.
- The real roots constraints can be expressed as $n-1$ discriminants being positive which are polynomials of the coefficients. (see Yang and Xia MM Research Preprints 97, Yang J. Symbolic Computation 99, Basu et al Springer 04)
- The (nonlinear) constraints are Chernoff regular, i.e. can be approximated by a cone.

Likelihood under anonymous data: summary

- The AC model is not identified [the likelihood is maximized for a SC model among AC models].
- The SI and SC likelihood are not altered and the ML-estimator still corresponds to empirical proportions as under non-anonymous data.
- Only the AI model needs care.
- The key element for the tractability of the likelihood maximization: separability in the variables!
- Then the goal is: test H_{SI} versus H_{AI} or a partial test for H_{AI} versus H_{AC}. We also need a complete estimation procedure of the AI model and the corresponding large sample distributions.

The different likelihood maximizations

The likelihood is concave and should be maximized under some additional set of constraints.

$$\lambda((\alpha^{i:n})_{i=1,\cdots,n}) = \sum_{j=0}^{n} p^j . \log[\alpha^{j:n}]$$

- SC model: no additional constraint
- AI model:
  with the constraint that the roots of the polynomial $X \to \sum_{i=1}^{n} \alpha^{i:i}.(-1)^{n-i}.X^i$ are all real.
- SI model: with the constraint $\alpha^{i:n} = \binom{n}{i}\alpha^i.(1-\alpha)^{n-i}$

Intermediate symmetry assumption discussed later and maximum-likelihood based estimators are developed.

The SC versus the AI model

In the SC model, what is the interpretation of the additional constraints the roots of the polynomial $X \rightarrow \sum_{i=1}^{n} \alpha^{i:i}.(-1)^{n-i}.X^{i}$ are all real.

- For $n = 2$, it is equivalent to negative correlation.
- Corollary: we can distinguish the symmetric positively correlated model from the AI model!

Aim of the reparametrization

- The original likelihood is singular under symmetry, hardly tractable under standard algorithms
- After the reparametrization, the likelihood is not singular, tractable but the true parameter can lie on the boundary of the parameter space (under symmetry)

Extensions: maximum likelihood based approaches

Consistent Estimator that root-finding of polynomials (instead of complex optimization...)

- For $K > 2$ the general maximum likelihood is not tractable but an estimator can be constructed that uses the idea for $K = 2$ as in Lamy (mimeo 07) for continuous variables.
- For general symmetry structures. E.g. suppose a double root then first estimate the roots of the derivative polinomial P' and for each roots β compute the roots of the polynomial $Q/(X-\beta)^2$ where Q is the primitive of P' such that β is a double root. Pick the solution among the different candidates that maximizes the likelihood.
- With incomplete data set and some symmetry assumption.

The Sylvester Matrix

The Sylvester matrix of P and Q, denoted by $Syl(P, Q)$, is the $(p + q) \times (p + q)$ matrix:

$$\begin{bmatrix}
a_p & \cdots & \cdots & \cdots & \cdots & a_0 & 0 & \cdots & 0 \\
0 & \ddots & & & & & \ddots & \ddots & 0 \\
\vdots & \ddots & \ddots & & & & & \ddots & 0 \\
0 & \cdots & 0 & a_p & \cdots & \cdots & \cdots & \cdots & a_0 \\
b_q & \cdots & \cdots & \cdots & b_0 & 0 & \cdots & \cdots & 0 \\
0 & \ddots & & & & \ddots & \ddots & & \vdots \\
\vdots & \ddots & \ddots & & & & \ddots & \ddots & \vdots \\
\vdots & & \ddots & \ddots & & & & \ddots & 0 \\
0 & \cdots & \cdots & 0 & b_q & \cdots & \cdots & \cdots & b_0
\end{bmatrix},$$

The Sylvester Matrix

- where $P = \prod_{i=1}^{p} (X - x_i) = \sum_{i=0}^{p} a_i X^i$ and $Q = \prod_{i=1}^{q} (X - y_i) = \sum_{i=0}^{q} b_i X^i$
- $$det(Syl(P, Q)) = a_p^q b_q^p \prod_{i=1}^{p} \prod_{j=1}^{q} (x_i - y_j)$$
- $det(Syl(P, Q)) \neq 0$ if and only if P and Q are coprime.

The Sylvester Matrix

- $$\begin{array}{rccc} m: & \mathcal{C}^q \times \mathcal{C}^p & \rightarrow & \mathcal{C}^{p+q} \\ & (Q, P) & \rightarrow & QP \end{array}$$
- The Jacobian matrix of m is the Sylvester matrix of P and Q and the Jacobian of m is the resultant.
- Denote by $IndSyl(k, P, Q)$ the matrix:

$$\begin{bmatrix} I_k & \mathbf{0} \\ \mathbf{0} & Syl(P, Q) \end{bmatrix}$$

The Sylvester Matrix: corollary

$$\begin{array}{rccc} m: & \mathcal{C}^{q_1} \times \cdots \times \mathcal{C}^{q_l} & \rightarrow & \mathcal{C}^{q_1+\cdots+q_l} \\ & (Q_1, \cdots, Q_l) & \rightarrow & \prod_{i=1}^{l} Q_i \end{array}$$

m is regular at any point $(Q_1, \cdots, Q_l)$ such that each couple (Q_i, Q_j), $j \neq i$ are coprime. The jacobian matrix of m is given by:

$$J_m = \prod_{k=0}^{l-2} IndSyl(\sum_{j=1}^{k} q_j, Q_{k+1}, \prod_{i=k+2}^{l} Q_k)$$

since $m = m_{l-1} \circ m_k \circ m_1$ where

$$\begin{array}{rccc} m_k: & \mathcal{C}^{q_1} \times \cdots \times \mathcal{C}^{q_k} \times \mathcal{C}^{\sum_{i=k+1}^{l} q_i} & \rightarrow & \mathcal{C}^{q_1} \times \cdots \times \mathcal{C}^{q_{k-1}} \times \mathcal{C}^{\sum_{i=k}^{l} q_i} \\ & (Q_1, \cdots, , Q_{k+1}) & \rightarrow & (Q_1, \cdots, Q_{k-1}, Q_k.Q_{k+1}) \end{array}$$

- Let

$$\begin{array}{rccc} f^y : & \mathcal{C}^n & \rightarrow & \mathcal{C}^n \\ & (\alpha_{n-1}, \cdots, \alpha_0) & \rightarrow & (a_{n-1}, \cdots, a_0) \end{array}$$

such that the polynomials $X^n + a_{n-1}.X^{n-1} + \cdots + a_0$ and $(X-y)^n + \alpha_{n-1}.(X-y)^{n-1} + \cdots + \alpha_0$. f^y corresponds indeed to a change of base and is regular. The jacobian matrix $(f_{ij}^y)_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n}}$ is given by

$$\begin{array}{rcll} f_{ij}^y = & 0 & if & i > j \\ f_{ij}^y = & 1 & if & i = j \\ f_{ij}^y = & \binom{j-1}{i-1}(-y)^{j-i} & if & i < j \end{array}$$

Proposition

The linear function f is invertible.

Use of the previous results

Corollary

If all the roots are distinct, then the Jacobian of the map linking the coefficient of a polynomial to its roots is inversible.

Next when we consider some multiplicity of the roots (e.g. some symmetry)

Links between the roots of a polynomial and the roots of the derivatives

Lemma

Let P a polynomial of degree p with p real roots (counted with the order of multiplicity) $x_1 \leq \cdots \leq x_p$, then the polynomial P' has $p-1$ real roots $x'_1 \leq \cdots \leq x'_{p-1}$ such that $x'_i \in [x_i, x_{i+1}]$, for $i = 1, \cdots, p$

Corollary

Suppose that P of degree p is a polynomial with p real roots. The i^{th} derivative $P^{(i)}$ of the polynomial P has a root x of multiplicity $k > 1$ if and only if x is a root of P with multiplicity $k + i$.

To test that the highest multiplicity of the roots of P is k, then we can derive P $k-2$ times and test multiplicity of order 2!

Estimation of the roots

- The estimated coefficient $\alpha^{i:i}$ of the polynomial are normally distributed (asymptotic of a multinomial distribution + the linear transformation H)
- The tricky point is to go from the estimation of the coefficient to the roots of the polynomials. This point has been lead by Pantula and Fuller (Biometrika 93) who were motivated by the analysis of an autoregressive polynomial for time-series analysis. Their analysis is restricted to polynomials of order 2. Ours is general!
- Thus our analysis may be useful outside the anonymous data framework!

Large Sample Distribution of the roots

Proposition

Let $P = \prod_{i=1}^{l} (X - \alpha_i)^{k_i}$ with $\sum_{i=1}^{l} k_i = n$ and $\alpha_1 < \cdots < \alpha_l$ be the true value of the polynomial. Then the maximum likelihood estimator of the AI model without the constraints and denoted by $\widehat{\alpha}_1^1, \cdots, \cdots, \widehat{\alpha}_1^{k_1}, \cdots, \widehat{\alpha}_l^1, \cdots, \widehat{\alpha}_l^{k_l}$ is distributed such:

$$\sqrt{T}. \begin{pmatrix} (\widehat{\alpha}_1^1 - \alpha_1)^{k_1} \\ \vdots \\ (\widehat{\alpha}_1^{k_1} - \alpha_1)^{k_1} \\ \vdots \\ \vdots \\ (\widehat{\alpha}_l^1 - \alpha_l)^{k_l} \\ \vdots \\ (\widehat{\alpha}_l^{k_l} - \alpha_l)^{k_l} \end{pmatrix} \rightarrow_d \mathcal{N}(0, \Sigma),$$

where the matrix Σ can be expressed as $\Sigma = \Pi J H \Sigma_0 H' J' \Pi'$

Decomposition of the covariance matrix Σ

- $$\sqrt{T}.(\widehat{p}_m - p_m) \rightarrow_d \mathcal{N}(0, \Sigma_0)$$

  where $\Sigma_0 = \{p^i_m(\delta_{ij} - p^j_m)\}_{1 \leq i,j \leq n}$ and δ_{ij} is Kronecker's delta (asymptotic of a multinomial distribution).
- H is the linear transformation to go from p_m to $p^{i:i}$
- J is the jacobian of the nonlinear bijection that transform the coefficients $p^{i:i}$ into the coefficient of the 'polynomial decomposition', i.e. the coefficients of the l polynomials with degree k_j denoted by $X^{k_l} + \sum_{i=0}^{k_j - 1} a^i_j$. [it is the inverse of the determinant of some sylvester matrix]
- Π is the projection that keeps only the coefficient corresponding to the lowest order term of each polynomial, i.e. the terms a^0_j.

Comments

- If all roots of P are single roots, then the convergence rate is in $\sqrt{T}$.
- For a multiple root of order k_i, the convergence rate is $T^{\frac{1}{2k_i}}$. The roots of α_i are distributed according to the k_i^{th} roots of a normal distribution.
- Corollary: If $k_i > 2$, the probability that all estimated roots are real converge to zero as T goes to infinity.
- The estimation procedure is 'bad' if the true sampling scheme contains some symmetry. Need for a testing/estimation procedure such that we estimate the order of multiplicity of each roots and then estimates the roots. A procedure that always converges in $\sqrt{T}$

ML-based estimator under some symmetry

Let $P = \prod_{i=1}^{l} (X - \alpha_i)^{k_i}$ with $\sum_{i=1}^{l} k_i = n$ and $\alpha_1 < \cdots < \alpha_l$ be the true value of the polynomial which is now known to the econometrician. The following algorithm builds a $\sqrt{T}$-consistent estimator.

- Pick the estimated polynomial. Take the maximum of the k_i.
- Derive $k_i - 1$ times the polynomial and pick all the roots. Successively consider the $n - k_i + 1$ roots and do
- Integrate the polynomial such that the chosen root is of order k_i and repeat the algorithm for the quotient that is theoretically equal to $P = \prod_{j \neq i} (X - \alpha_j)^{k_j}$.
- At the end consider among all the candidate solution the one that maximize the likelihood.

Comments

- Note that the $k_i - 1$ lowest order term of the original estimated polynomial have not been used except at the last stage!
- Application for incomplete data set...