Sensitivity Analysis in Some Econometric Models

Elie Tamer

Cowles Lecture

2015 World Congress of the Econometric Society - Montreal

Problem

We are interested in inference on θ based on a random sample of observations using the *parametric* likelihood

$$P \in \mathcal{F}_\theta$$

Issues:

- Standard inference is possible here under the usual conditions.
- With current data sets with large sample sizes, standard errors tend to be small and the main issues become one of misspecification.
- Acemoglu ...
- So, we want to formally account for this model uncertainty.
- So, heuristically, in addition to standard confidence intervals that people report, we want to develop *model uncertainty intervals* that give an idea of the *sensitivity* of our preferred estimates to worrisome assumptions.
- The research agenda, described in this talk, tries to make more progress on this.

Background, Motivation and Examples

- By now, it is well understood that inferences in models may be sensitive to the assumptions made.
- Early advocates for formal sensitivity analysis such as Leamer (*"Sensitivity Analysis Would Help"*) have understood that estimates that are fragile "are not worth making"...
- This has led to a massive literature in econometrics on semiparametric models that essentially considered almost the weakest models (semiparametric nonlinear model, GMM with unknown functions, ...) under conditions guaranteeing point identification.
- The success of this literature did not translate into wide use of these models.
- Instead, empirical work continues to either simple linear models (or variants thereoff), or more complicated structural and *parametric* models.

NOT SURE I LIKE THIS

- For Semiparametric models, we are told: 1) the assumptions are hard to understand, 2) require restrictions that may not hold in commonly used data sets, 3) not easy to compute, 4) resulting estimands difficult to interpret,....
- On the other hand, fancy parametric models are easy to interpret and simpler to compute (though not always), and sensitivity when done, is "checked" by adding *higher order terms, etc...*

- The semiparametrics literature has focused on point identification. Without strong assumptions, conditions for point identification there entail *strong requirements* on the data: Large support, continuous regressors, etc and/or some *limits on* heterogeneity, correlation, etc. This has limited its empirical appeal (do we really have data sets with continuous regressors that admit large support conditional on other variable...)
- Another literature has emerged (early with Manski's work) that gives up on point identification and the focus becomes getting the identified set (or the tightest set) given a set of conditions.

- The research I will describe today takes this *partial identification approach* from the *top down*:
  - start with a (fully) parametric model (typically a structural model),
  - designate the parameter of interest (policy response, elasticity etc).
  - identify what part of the parametric model that you are most worried about? a parametric distribution or a functional form restriction that is not motivated by economic theory, or a behavioral assumption (like say Markov Perfect, or rational expectation)
  - Conduct *formal sensitivity analysis* with respect to that class: See whether the parametric conclusions rely on the specification of those suspect assumptions.
  - The main issue is that if one were to relax the worrisome assumption, point identification is likely to fail.
  - So, the statistical work entails the construction of valid inference procedures that work whether or not point identification holds.

This is appealing because:

- The starting point of this is an *applied economist's* preferred model (can be a fancy structural model, etc)
- This is in contrast with starting with the most nonparametric model
- We add to it formal -and valid- statistical approaches to account for *model uncertainty* for that part of the model that the analyst is most worried about,
- so, in addition to standard confidence intervals that people report, think of this as a way to report *model uncertainty* intervals that give an idea of the *sensitivity* of our preferred estimates to worrisome assumptions.

Challenges of this program

- This should be done in a computationally tractable way so it can be used; this seems to be the biggest challenge
- Must be theoretically attractive (and valid)
- Can this framework, developed initially for a likelihood setup be used in moment based models?

Outline of Lecture

1. examples
2. inference approaches
3. more examples
4. variations
5. future work

Example 1: Dynamic Binary Choice Panel

Consider the model (studied in Honoré and Tamer (2006) and Chen, Tamer, and Torgovitsky (2013))

$$y_{it} = 1\left\{x_{it}'\beta + y_{i,t-1}\gamma + \eta_i + \varepsilon_{it} \geq 0\right\}$$

- Observe a sample of N individuals for T time periods starting with $t = 1$, to get a sample of$\{(y_{it}, x_{it})\}_{t=1}^{T}$ where we use the notation $x_i^T = (x_{i1}, ..., x_{iT})$.
- Lagged dependent variable creates the *Initial Condition Problem*, so let the function $g(\eta_i, x_i^T) \equiv P(y_{i0} = 1|\eta_i, x_i^T)$
- Let η_i be independent of x_i^T and ε_{it} for all t, and is distributed as $N(0, \sigma^2)$ and let ε_{it} be a standard normal random variable that is i.i.d. across t and i and statistically independent of x_i^T
- The parameter of interest can be a marginal effect, or other parameters.

The population log likelihood is

$$E\left[\log\left(p\left(y_i;\theta,g\left(\cdot,x_i^T\right),x_i^T\right)\right)\right]$$

$$= E\left[\log\left(\int\left\{g\left(\eta,x^T\right)^{y_{i1}}\left(1-g\left(\eta,x^T\right)\right)^{1-y_{i1}}\prod_{t=2}^{T}P\left(y_{it}|\,x_i^T,y_{it-1};\theta,\eta\right)\right\}dF_{\eta|x^T}\left(\eta|\right.\right.\right.$$

where

$$P\left(y_{it}=0|\,x_i^T,y_{it-1};\theta,\eta\right)=\Phi(-x_{i,t}'\beta-\gamma y_{i,t-1}-\eta)$$

This is a fully parametric likelihood (except for the initial condition problem), and we are worried about specifying the initial condition distribution.
We want to see whether our estimates of the parameters *are sensitive* to specification of the condition distribution.

notes on this

- Without specifying the initial condition distribution, we know that the parameters are not (point) identified.
- The *identified set* here can be defined through the likelihood above.
- Why are we just worried about the initial condition distribution, and not everything else?

Women's Labor Supply: (more on this later)

- Sample consists of 1812 women who were aged 18 60 in 1980 and continuously married to an employed husband during each year 1979 1985.
- Covariates: transitory and permanent nonlabor income, number of children aged 0-2, 3-5 and 6-17, lagged number of children aged 0-2, age and its square, highest reported level of education over the sample period, and race.
- Permanent nonlabor income is defined as the average of the husbands log earnings over the sample period. Transitory nonlabor income in each period is defined as the deviation of husbands log earnings in that period from permanent nonlabor income.

Women's Labor Supply: ctd

- There are a few benchmark parametric models that are used.
- The first approach is to simply assume that y1 is exogenous.
- The second approach, proposed by Heckman (1981), is to model the initial period as

$$y_{i1} = 1[x_{i1}'\lambda + \rho\eta_i + \epsilon_{i1} \geq 0],$$

- The third approach, due to Wooldridge (2005), is to treat the first period outcome as an explanatory variable:

$$y_{it} = 1[\gamma y_{i,t-1} + x_{it}'\beta + \psi y_{i1} + u_i + \epsilon_{it} \geq 0]$$

Ex 1: preview

Benchmark with $T = 3$

	Exogenous	Wooldridge	Heckman
y_{it-1}	1.909 [1.800,2.018]	.819 [.433,1.204]	1.054 [.781,1.326]
log ymp	−.107 [−.208,−.00709]	−0.145 [−0.302,0.0108]	−0.289 [−0.462,−0.116]
log ymt	−.201 [−0.369,−0.0340]	−0.230 [−0.456,−0.00449]	−0.241 [−0.450,−0.0328]
$chi02$	−.255 [−.402,−.107]	−0.359 [−0.562,−0.157]	−0.289 [−0.462,−0.116]
⋮			

Standard parametric estimates using MLE with 95% CIs.

Ex1: preview

	Benchmark with $T = 3$		
Pref. Model	Exogenous	Wooldridge	Heckman
y_{it-1}	1.909 [1.800,2.018]	.819 [.433,1.204]	1.054 [.781,1.326]
	[.435, 1.288]	[.435, 1.288]	[.435, 1.288]
log *ymp*	$-.107$ [$-.208$,$-.00709$]	-0.145 [-0.302,0.0108]	-0.289 [-0.462,-0.116]
log *ymt*	$-.201$ [-0.369,-0.0340]	-0.230 [-0.456,-0.00449]	-0.241 [-0.450,-0.0328]
*chi*02			
⋮			

We add here a *third row* to account for misspecification of initial condition distribution:

$$P\left(y_{it} = 0 |\, x_i^T, y_{it-1}; \theta, \eta \right) = \Phi(-x'_{i,t}\beta - \gamma y_{i,t-1} - \eta)$$

Ex1: preview

Benchmark with $T = 3$			
Pref. Model	Exogenous	Wooldridge	Heckman
y_{it-1}	1.909 [1.800,2.018]	.819 [.433,1.204]	1.054 [.781,1.326]
	[.435, 1.288]	[.435, 1.288]	[.435, 1.288]
$\log ymp$	$-.107$ $[-.208,-.00709]$	-0.145 $[-0.302,0.0108]$	-0.289 $[-0.462,-0.116]$
$\log ymt$	$-.201$ $[-0.369,-0.0340]$	-0.230 $[-0.456,-0.00449]$	-0.241 $[-0.450,-0.0328]$
$chi02$			
$\vdots$			

We add here a *third row* to account for misspecification of initial condition distribution:

$$P\left(y_{it} = 0 | \, x_i^T, y_{it-1}; \theta, \eta \right) = \Phi(-x'_{i,t}\beta - \gamma y_{i,t-1} - \eta)$$

Ex1: preview

Benchmark with $T = 3$			
Pref. Model	Exogenous	Wooldridge	Heckman
y_{it-1}	1.909 [1.800,2.018]	.819 [.433,1.204]	1.054 [.781,1.326]
	[.435, 1.288]	[.435, 1.288]	[.435, 1.288]
log *ymp*	$-.107$ [$-.208$,$-.00709$]	-0.145 [-0.302,0.0108]	-0.289 [-0.462,-0.116]
	[-0.679, -.382]	[-0.679, -.382]	[-0.679, -.382]
log *ymt*	$-.201$ [-0.369,-0.0340]	-0.230 [-0.456,-0.00449]	-0.241 [-0.450,-0.0328]
*chi*02			
⋮			

We add here a *third row* to account for misspecification of initial condition distribution:

$$P\left(y_{it} = 0 |\, x_i^T, y_{it-1}; \theta, \eta\right) = \Phi(-x'_{i,t}\beta - \gamma y_{i,t-1} - \eta)$$

Ex1: preview

	Benchmark with $T = 3$		
Pref. Model	Exogenous	Wooldridge	Heckman
y_{it-1}	1.909	.819	1.054
	[1.800,2.018]	[.433,1.204]	[.781,1.326]
	[.435, 1.288]	[.435, 1.288]	[.435, 1.288]
$\log ymp$	$-.107$	-0.145	-0.289
	[$-.208$,$-.00709$]	[-0.302,0.0108]	[-0.462,-0.116]
	[-0.679, -.382]	[-0.679, -.382]	[-0.679, -.382]
$\log ymt$	$-.201$	-0.230	-0.241
	[-0.369,-0.0340]	[-0.456,-0.00449]	[-0.450,-0.0328]
	[-1.051, .324]	[-1.051, .324]	[-1.051, .324]
$chi02$			
⋮			

We add here a *third row* to account for misspecification of initial condition distribution:

$$P\left(y_{it} = 0 |\, x_i^T, y_{it-1}; \theta, \eta \right) = \Phi(-x'_{i,t}\beta - \gamma y_{i,t-1} - \eta)$$

Ex1: preview

	Benchmark with $T = 7$		
Pref. Model	Exogenous	Wooldridge	Heckman
y_{it-1}	1.733 [1.586,1.880]	1.122 [1.012,1.232]	1.167 [1.065,1.269]
	[1.010, 1.246]	[1.010, 1.246]	[1.010, 1.246]
$\log ymp$	$-.273$ [$-.368$,$-.179$]	-0.283 [-0.401,-0.165]	-0.432 [-0.562,-0.303]
	[-0.571, -0.427]	[-0.571, -0.427]	[-0.571, -0.427]
$\log ymt$	$-.162$ [-0.262,-0.0620]	-0.200 [-0.307,-0.0931]	-0.187 [-0.294,-0.0809]
	[-0.332,-0.115]	[-0.332,-0.115]	[-0.332,-0.115]
$chi02$			
$\vdots$			

We add here a *third row* to account for misspecification of initial condition distribution:

$$P\left(y_{it} = 0 | x_i^T, y_{it-1}; \theta, \eta\right) = \Phi(-x'_{it}\beta - \gamma y_{it-1} - \eta)$$

Ex 2: Discrete Game

Consider the following game:

$$y_1 = 1\left[x_1'\beta_1 + \Delta_1 y_2 + \epsilon_1 \geq 0\right]$$
$$y_2 = 1\left[x_2'\beta_2 + \Delta_2 y_1 + \epsilon_2 \geq 0\right]$$

We observe (y_1, x_1', y_2, x_2) and given a set of assumptions (...) the likelihood:

$$P_y(\mathbf{x}, \theta, g(.)) = \sum_{e \in \mathcal{E}(x,\theta)} g^e(x, \theta) P_y^e(\epsilon, x, \theta) dF(\epsilon, \theta)$$

We can use (standard MLE). But, what if we are worried about sensitivity of our estimates to specification of the *equilibrium selection function* $g^e(x, \theta)$?
When using a semiparametric version of this likelihood (g^e flexible) we loose identification in general.

Ex2 Preview

Data: second quarter of 2010s Airline Origin and Destination Survey (DB1B) - see Kline and Tamer (2015) (more on this later)

Pref. Model	Logit Selection	Fixed Probability selection
Market Presence - large	1.724 [1.112,1.233]	2.272 [2.112,2.432]
Market Size - large	.23 [0.21,0.25]	0.85 [0.83,0.87]
Correlation	0.366 [.310,.422]	1.152 [0.144,0.156]
⋮		

This is estimated using (standard MLE) where *correlation* is the correlation between ϵ_1 and ϵ_2, assumed jointly normal.

Ex2 Preview

Pref. Model	Logit Selection	Fixed Probability selection
Market Presence - large	1.724 [1.112,1.233]	2.272 [2.112,2.432]
	[.85, 2.201]	[.85, 2.201]
Market Size - large	.23 [0.21,0.25]	0.85 [0.83,0.87]
	[.012, .865]	[.012, .865]
Correlation	0.366 [.310,.422]	1.152 [0.144,0.156]
	[.652, .99]	[.652, .99]
⋮		

Ex 3: Trade Model of HMR (2008)

- In an influential paper, Helpman, Melitz and Rubenstein (2008) examine the extensive margin of trade using a structural model estimated with current trade data.
- introduce a "cut-off" a_L for productivity: if a random draw for productivity from country i to j, $a_{ij} > a_L$ then i trades with j. Otherwise, no trade occurs. So, in their notation:

$$m_{ij} = \beta_{01} + \lambda_{j1} + \chi_{i1} + ... + f(z^*_{ij}) + u_{ij}$$

$$z^*_{ij} = \beta_{02} + \lambda_{j2} + \chi_{i2} + \beta_{12} \log dist_{ij} + ... + u_{ij} + \nu_{ij}$$

where $\lambda^1_i, \lambda^2_i, \chi^1_i, \chi^2_j$ are country fixed effects (> 200), (u_{ij}, ν_{ij}) is jointly normal and $f(.)$ is a known *nonlinear function*. We only observe m_{ij} when $z^*_{ij} > 1$.

Ex3 assumptions

This is a structural model that is fully parametrized:

- Pareto distribution on productivity a (which leads to a particular functional form for $f(.)$)
- joint normality of the errors
- heteroskedasticity - which is well known to be a problem in trade data (and is especially problematic in nonlinear selection like models)

As above, we report below HMRs favorite model and also formal sensitivity intervals.

The inference approach here will be described below.

Ex3 results

bivariate selection model estimated via MLE (similar estimates using NLS -)

Pref. Model	Outcome Equation	Selection Equation
log distance	$-.03265$ [−0.19,0.13]	-0.165 [−0.212,−0.1125]
border	1.9548 [1,4348,2.4747]	0.2527 [0.0927,0.41]
legal system	0.1747 [0.0369,0.3107]	-0.0532 [−0.0932,−0.0132]
⋮		

Standard 95% confidence intervals (? fixed effects and ? parameters)

Ex3 results

bivariate selection model estimated via MLE (similar estimates using NLS -)

Pref. Model	Outcome Equation	Selection Equation
log distance	$-.03265$	-0.165
	[-0.19,0.13]	[-0.212,-0.1125]
	[-0.1089, 0.0911]	[-0.517, -0.8122]
border	1.9548	0.2527
	[1.4348,2.4747]	[0.0927,0.41]
	[0.262, 1.954]	[-0.022, 1.073]
legal system	0.1747	-0.0532
	[0.0369,0.3107]	[-0.0932,-0.0132]
	[-0.549, -0.048]	
⋮		

Formal problem, Inference and Computations

Let the data $\{Z_i = (Y_i, X_i)\}_{i=1}^{n}$ be a random sample of $Z = (Y, X)$ that has true (but unknown) density $p_0 \in \mathcal{P}_0$

Here, the applied economist uses (standard inference) his/her preferred model (parametric):

$$p(\theta) \equiv p(\theta, g)$$

$g \in \mathcal{G}$ represents the piece of the parametrization that this applied economist is most worried about.

Are estimates of θ sensitive to the particular specification used for g (especially when there is loss of identification)?

The issue here and throughout the work is that there might be pairs (θ_1, g_1) and (θ_2, g_2) s.t.

$$p(.;\theta_1, g_1) = p(.;\theta_2, g_2) = p_0$$

So, all inference methods should allow for lack of identification.

notation

- *The identified set* for θ is

$$\Theta_I = \{\theta \in \Theta \subset \Re^{d_\theta} : \quad p(.;\theta, g) = p_0 \quad \text{for some} \quad g \in \mathcal{G}\}$$

  This is the set that characterizes the **sensitivity** of the model to the inclusion of the infinite dimensional nuisance parameter g. Similarly,

- The identified set for $\alpha = (\theta, g) \in \mathcal{A} = \Theta \times \mathcal{G}$ is

$$\mathcal{A}_I \equiv \{\alpha = (\theta, g) \in \mathcal{A} : p(\cdot;\alpha) = p_0\}$$

- So, if $\mathcal{G}$ is small (parametric), then above becomes a parametric likelihood with partial identification.
- When $\mathcal{G}$ is large (non-parametric), then this becomes a semiparametric likelihood with partial identification.

Approaches to inference

I will explain three approaches to inference.

1. CIs based on inverting semiparametric LR statistics under lack of point identification. This work, used in Example 1, is based on work joint with Xiaohong Chen and Alex Torgovitsky.
2. In certain "separable" models, a Bayesian approach in finite support models. This approach, used in Example 2, is based on work with Brendan Kline.
3. A more general pseudo-Bayes (in progress) approach based on establishing a type of Bernstein-VonMises Theorem for sets. This more general approach, used in Example 3, is based on ongoing work with Xiaohong Chen and Tim Christensen.

inference approach 1: intuition and some details

- Given a density $p(\theta, g)$ with $\theta \in \Theta \subset \mathcal{R}^{d_\theta}$ and $g \in \mathcal{G}$, where for a given n *approximate* this $\mathcal{G}$ by some (parametric) set $\mathcal{G}_k$.
- Invert the following LR statistic

$$LR(\theta_0) \equiv 2 \left[\sup_{(\theta, g_k) \in \Theta \times \mathcal{G}_{k(n)}} \sum_{i=1}^{n} \log p(Z_i; (\theta, g_k)) - \sup_{g \in \mathcal{G}_{k(n)}} \sum_{i=1}^{n} \log p(Z_i; (\theta_0, g)) \right]$$

- how to show that LR above has a tight limiting distribution under the null hypothesis $H_0 : \theta = \theta_0$

- Typically, for a regular parametric likelihood model $p(\cdot, \theta)$ distribution of LR statistic under the null of $\theta = \theta_0$ uses a quadratic approximation to the sample log-likelihood $L_n(\theta) = \sum_{i=1}^{n} \log p(Z_i; \theta)$ in a Euclidean $n^{-1/2}$ neighborhood of the "true parameter" θ_0; see, e.g. Chernoff (1954).
- When the model is not point identified, this quadratic approximation around the true value is not natural (there are more than one observationally equivalent parameters θ_0 under the null).

- Even though it is possible for given n, θ is identified in the sieve space, but in the limit, we have partial identification that needs to be accounted for.
- We use the following insight: whereas the parameter θ is not unique under the null, the true probability density, p_0, is unique (the density of the data). So, we can derive the distribution of the LR statistic by *expanding and centering around p_0 appropriately*. See Liu and Shao (2003).

Asymptotic Null Dist. of Sieve LR Statistic

Theorem 1

Under the null hypothesis $\theta = \theta_0$ (and assumptions ...)

$$
\begin{aligned}
LR(\theta_0) &\equiv 2\left[\sup_{(\theta,g)\in\Theta\times\mathcal{G}_{k(n)}} \sum_{i=1}^{n} \log p(Z_i;(\theta,g)) - \sup_{g\in\mathcal{G}_{k(n)}} \sum_{i=1}^{n} \log p(Z_i;(\theta_0,g))\right] \\
&= \sup_{d\in\mathcal{D}^{eff}_{k(n)}} \left(\max\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d(Z_i), 0\right\}\right)^2 + o_{P_Z}(1) \\
&= \sup_{d\in\mathcal{D}^{eff}} \left(\max\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d(Z_i), 0\right\}\right)^2 + o_{P_Z}(1) \\
&\Rightarrow \sup_{d\in\mathcal{D}^{eff}} \left(\max\{W(d), 0\}\right)^2 \text{ in distribution,}
\end{aligned}
$$

where $\{W(d) : d \in \mathcal{D}^{eff}\}$ is a tight centered Gaussian process with variance one and covariance function $\Gamma(d_1, d_2) = \int d_1 d_2 p_0 \mu$ defined on $\mathcal{D}^{eff} \times \mathcal{D}^{eff}$ and $\mathcal{D}^{eff}$ is and appropriately defined set of efficient scores.

comments

- distribution above is complicated (think of mixture models). So, it is generally hard to use it directly.
- simplifies to the usual χ^2 limit in regular cases.
- test can be shown to be consistent.

Weighted Bootstrap to do inference

Assumption 1

Let the following hold. (i) $\{\omega_i\}_{i=1}^n$ *is a positive, i.i.d. sequence drawn from the distribution of a positive random variable* ω *with* $E[\omega] = 1$, $Var[\omega] = \sigma_\omega^2 \in [0, \infty)$ *and* $\|\omega\|_{2,1} \equiv \int_0^\infty \sqrt{\Pr(\omega > t)} dt < \infty$; *(ii)* $\{\omega_i\}_{i=1}^n$ *is independent of* $\{Z_i\}_{i=1}^n$.

Inference: Weighted Bootstrap

Theorem 2

Let $\widehat{p} \equiv p(\cdot; \widehat{\theta}, \hat{g}) = \arg\sup \sum_{i=1}^{n} \log p(Z_i; (\theta, g))$ *be the sieve MLE.*

Let $\{\omega_i\}_{i=1}^n$ *be a positive, i.i.d. sequence drawn from the distribution of a positive r.v.* ω *with* $E[\omega] = 1$, $Var[\omega] = 1$. *Then: conditional on the data* $\{Z_i\}_{i=1}^n$ *satisfying the null hypothesis of* $\mathcal{A}_I^r \neq \varnothing$,

$$\begin{aligned} LR^{\omega}(\widehat{r}) &\equiv 2\left[\sup_{(\theta,g)\in\Theta\times\mathcal{G}_{k(n)}} \sum_{i=1}^{n} \omega_i \log p(Z_i; (\theta, g)) - \sup_{g\in\mathcal{G}_{k(n)}} \sum_{i=1}^{n} \omega_i \log p(Z_i; (\hat{\theta}, g \right. \\ &\Rightarrow \sup_{d\in\mathcal{D}^{eff}} \left(\max\{W(d), 0\}\right)^2 \text{ in distribution.} \end{aligned}$$

where $\{W(d) : d \in \mathcal{D}^{eff}\}$ *is a tight centered Gaussian process with variance one and covariance function* $\Gamma(d_1, d_2) = \int d_1 d_2 p_0 \mu$.

Weighted Bootstrap: the way it works

- For each $b = 1, \ldots, B$, generate a sample $(\omega_{ib})_{i=1}^{N}$ from an exponential distribution with mean 1 and variance 1.
- For each b sample, construct

$$LR_b = 2\left[\sup_{\theta, g \in \mathcal{G}_{k(n)}} \sum_{i=1}^{n} \omega_i \log p(Z_i, \theta, g) - \sup_{g \in \mathcal{G}_{k(n)}} \sum_{i=1}^{n} \omega_i \log p(Z_i; \hat{\theta}, g)\right]$$

  where $\hat{\theta}$ is the sieve MLE estimator of θ obtained from the data.
- Repeat above for every b to get the distribution of the LR statistic.

back to Ex1: labor supply

- to examine sensitivity, model the initial condition distribution flexibly

$$y_{i1} = g(x_{i1}'\beta, \alpha_i),$$

where $g(.)$ is approximatedf with a fully-interacted two-dimensional Bernstein polynomial of order 4 and x contains all covariates as the main model (except for lagged y).

Benchmark with $T = 3$

Pref. Model	Exogenous	Wooldridge	Heckman
y_{it-1}	1.909 [1.800,2.018]	.819 [.433,1.204]	1.054 [.781,1.326]
	[.435, 1.288]	[.435, 1.288]	[.435, 1.288]
log *ymp*	−.107 [−.208,−.00709]	−0.145 [−0.302,0.0108]	−0.289 [−0.462,−0.116]
log *ymt*	−.201 [−0.369,−0.0340]	−0.230 [−0.456,−0.00449]	−0.241 [−0.450,−0.0328]
*chi*02			
⋮			

comments on inference approach 1:

- coverage for the *true parameter* (that is not identified)
- computationally hard to implement due to inversion of test statistic

Inference Approach 2

- In certain classes of models, there is a more direct approach to inference that is *computationally very attractive.*
- Examples:
  1. Separable models: Example 2 above (more next)
  2. Finite support models
- Approach taken is Bayesian - but that is not as important (computationally tractable)

Inference Approach 2: examples

Consider the following game

	$a_2 = 0$	$a_2 = 1$
$a_1 = 0$	$0, 0$	$0, x_2\beta_2 + \epsilon_2$
$a_1 = 1$	$x_1\beta_1 + \epsilon_1, 0$	$x_1\beta_1 + \Delta_1 + \epsilon_1,\ x_2\beta_2 + \Delta_2 + \epsilon_2$

Given data on (a_1, a_2, x_1, x_2) and assumptions, we get the following:

$$\underbrace{\begin{pmatrix} \Pr((0,0)) \\ \Pr((1,1)) \\ \Pr((1,0)) \\ \Pr((0,1)) \end{pmatrix}}_{\text{DATA}} = \begin{pmatrix} P(\epsilon \in S_1^{x,\theta}) \\ P(\epsilon \in S_2^{x,\theta}) \\ P(\epsilon \in S_3^{x,\theta}) + P(d = 1 | \epsilon \in S_5^{x,\theta}) P(\epsilon \in S_5^{x,\theta}) \\ P(\epsilon \in S_3^{x,\theta}) + (1 - P(d = 1 | \epsilon \in S_5^{x,\theta})) P(\epsilon \in S_5^{x,\theta}) \end{pmatrix}$$

- $(\epsilon_1, \epsilon_2) \sim N(0, \Omega)$
- $P(d = 1 | \epsilon \in S_5^{x,\theta})$ is the equilibrium selection probability.

Inference Approach 2: examples

$$\underbrace{\begin{pmatrix} \Pr((0,0)) \\ \Pr((1,1)) \\ \Pr((1,0)) \\ \Pr((0,1)) \end{pmatrix}}_{\text{DATA}} = \begin{pmatrix} P(\epsilon \in S_1^{x,\theta}) \\ P(\epsilon \in S_2^{x,\theta}) \\ P(\epsilon \in S_3^{x,\theta}) + P(d=1|\epsilon \in S_5^{x,\theta})P(\epsilon \in S_5^{x,\theta}) \\ P(\epsilon \in S_3^{x,\theta}) + (1 - P(d=1|\epsilon \in S_5^{x,\theta}))P(\epsilon \in S_5^{x,\theta}) \end{pmatrix}$$

- One can do standard MLE when the selection probability are parametrized.
- To construct sensitivity analysis with respect to the selection probabilities for example, we can exploit the *separability* between the "data parameters" and the structural parameters and then map inference on data parameters to θ: Construct standard inference on $(P(0,0), P(1,1), P(1,0), P(0,1))$ and map that via model above to θ.
- No need for inversion of any test statistic.
- Bayesian analysis has equivalent frequentist interpretations in relevant cases.

Idea

The data provides information on (reduced form)

$$\begin{pmatrix} \Pr(0,0) \\ \Pr(1,1) \\ \Pr(1,0) \\ \Pr(0,1) \end{pmatrix}$$

These can be conditional...

Idea: Getting a posterior distribution on this vector of multinomial probabilities is instantaneous (more on this next).

Then, for every draw from this posterior, we "solve" for the set of θs that satisfy the above model.

Posterior probability statements on *the identified set* are then simple to deduce.

more generally (idea)

for a moment equality (or inequality) problem, and in cases with *finite support*, the problem is similar:
Consider

$$E_F[m(x,\theta)] = 0$$

This is equivalent to:

$$\begin{aligned} E_F[m(x,\theta)] &= 0 \\ \equiv \sum_{j=1}^{K} m(x_j,\theta)p_j &= 0 \end{aligned}$$

above can be moment inequality.

Again, we can make *standard inference* on $(p_1,\ldots,p_K)$ (get a posterior) and map posterior draws to θ via above vector of moment conditions.

Bayesian Bootstrap idea here

- Under a limiting uninformative Dirichlet prior for p (where the parameters of this prior all approach 0), the posterior for p approaches the Dirichlet posterior $Dir(n_1, \ldots, n_J)$, where $n_j = \sum_{i=1}^{n} 1[T_i = T_j]$.
- Standard results that connect the Gamma and Dirichlet distributions, a draw $p^{(s)}$ from the posterior for p can be approximated by: for each j, $\tilde{p}_j^{(s)} \sim^{ind.} Gamma(n_j, 1)$ if $n_j > 0$ and $\tilde{p}_j^{(s)} = 0$ if $n_j = 0$; and then taking for each j, $p_j^{(s)} = \frac{\tilde{p}_j^{(s)}}{\sum_{j=1}^{J} \tilde{p}_j^{(s)}}$.
- But also (additivity of Gamma distribution) we have:

$$\sum_{j=1}^{J} m(x_j, \theta) p_j^{(s)} = \sum_{j:n_j>0} m(x_j, \theta) p_j^{(s)} =^{distr.} \frac{\sum_{i=1}^{n} m(x_i, \theta) w_i^{(s)}}{\sum_i w_i^{(s)}},$$

where $w_i^{(s)} \sim^{ind.} Gamma(1, 1)$!

See Chamberlain and Imbens (2004) for a point identified version of this.

Computation - Big Advantage for this Approach

1. Generate copies of identified set:
   - (Step 1) Take a draw (p^s) from posterior of reduced form parameter. This is instantaneous via say the bayesian bootstrap.
   - (Step 2) Compute the identified set at $p^{(s)}$, $\{\Theta_I(p^{(s)})\}$.
2. Based on $\{\Theta_I(p^{(s)})\}_{s=1}^{S}$, compute an approximation to the desired posterior probability over the identified set.

Step 2- (hard) How do we get $\Theta(p^{(s)})$ for a given $p^{(s)}$

- This is a nonstochastic problem. Essentially, it is one of finding *all solutions in* θ to a mapping $Q(\theta, p) = 0$ for a given p with $Q(.)$ being some nonnegative function (like $Q(\theta, p) = \|(E_p m(T, \theta))^2\|$ or $\|(E_p m(T, \theta))_+^2\|$).
- One simple and useful approach to this computational problem is to use stochastic simulation methods to draw from the *pseudo-density*:

$$\tilde{f}_{\Theta_I(p^{(s)}),T}(\theta) = \exp\left(\frac{-Q(\theta, p^{(s)})}{T}\right)$$

  with T a tuning parameter that goes to zero (we have used $T = 10^{-4}$).
- Hwang(82) shows that $\{\tilde{f}_{\Theta_I(p^{(s)}),T}(\theta)\}$ converges to a distribution that is *supported over the set of argmin of* Q.

comments:

In these “separable problems” or models with finite support the above provides:

1. Valid posterior probability statements about the *identified set*.
2. Easy to compute in standard empirical models with a large number of parameters
3. Priors are innocuous since they are concerned with reduced form parameters
4. Paper contains Bernstein vonMises type results for the Bayesian credible sets to have frequentist interpretation.

Ex3: empirical model of entry

- We use data from from the second quarter of 2010's Airline Origin and Destination Survey (DB1B) to estimate a version of the binary game above where the payoff for firm i from entering market m is

$$\beta_i + \beta_i^x x_{im} + \Delta_i y_{3-i} + \epsilon_{im} \quad i = 1, 2$$

  The data contains 7882 markets which are formally defined as trips between two airports...
- The kinds of firms are "LCC" (low cost airlines) and OA (other airlines).
- X's: *market presence* and *market size*. Market presence is market/airline specific and is important here since it acts as the "excluded" variable.

connecting reduced form to structure

$$\underbrace{\begin{pmatrix} \Pr((0,0)|x) \\ \Pr((1,1)|x) \\ \Pr((1,0)|x) \\ \Pr((0,1)|x) \end{pmatrix}}_{DATA} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & P(d=1|\epsilon \in S_5^{x,\theta}) \\ 0 & 0 & 0 & 1 & 1 - P(d=1|\epsilon \in S_5^{x,\theta}) \end{pmatrix} \begin{pmatrix} P(\epsilon \in S_1^{x,\theta}) \\ P(\epsilon \in S_2^{x,\theta}) \\ P(\epsilon \in S_3^{x,\theta}) \\ P(\epsilon \in S_1^{x,\theta}) \\ P(\epsilon \in S_5^{x,\theta}) \end{pmatrix}$$

$$= \begin{pmatrix} P(\epsilon \in S_1^{x,\theta}) \\ P(\epsilon \in S_2^{x,\theta}) \\ P(\epsilon \in S_3^{x,\theta}) + P(d=1|\epsilon \in S_5^{x,\theta}) P(\epsilon \in S_5^{x,\theta}) \\ P(\epsilon \in S_3^{x,\theta}) + (1 - P(d=1|\epsilon \in S_5^{x,\theta})) P(\epsilon \in S_5^{x,\theta}) \end{pmatrix}$$

$$= \begin{pmatrix} P(\epsilon_1 \geq -\beta_1^0 - x_1\beta_1^1 - \Delta_1;\ \epsilon_2 \geq -\beta_2^0 - x_2\beta_2^1 - \Delta_2) \\ P(\epsilon_1 \leq -\beta_1^0 - x_1\beta_1^1;\ \epsilon_2 \leq -\beta_2^0 - x_2\beta_2^1) \\ P(\epsilon \in S_3^{x,\theta}) + P(d=1|\epsilon \in S_5^{x,\theta}) P(\epsilon \in S_5^{x,\theta}) \\ P(\epsilon \in S_3^{x,\theta}) + (1 - P(d=1|\epsilon \in S_5^{x,\theta})) P(\epsilon \in S_5^{x,\theta}) \end{pmatrix}$$

Sensitivity with respect to what:

So, above is the applied economist's prefered specification. It can be estimated via MLE.

Assumptions it is built on:

1. Economics: linear variable profits, Nash equilibrium, ...
2. functional forms/distributional: joint normality on the errors, a parametrized form for *equilibrium selection*

So, we can choose what piece of the model is worrisome. Here, we usually have little information about *equilibrium selection*.

Ex2: equilibrium selection sensitivity

Pref. Model	Logit Selection	Fixed Probability selection
Market Presence - large	1.724 [1.112,1.233]	2.272 [2.112,2.432]
	[.85, 2.201]	[.85, 2.201]
Market Size - large	.23 [0.21,0.25]	0.85 [0.83,0.87]
	[.012, .865]	[.012, .865]
Correlation	0.366 [.310,.422]	.151 [0.144,0.156]
	[.652, .99]	[.652, .99]
⋮		

we can transform these interval estimates into partial effects.

sensitivity of correlation coefficient to equilibrium selection

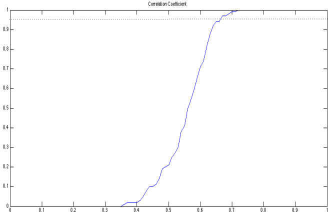

Figure: Posterior Probability that a particular parameter value belongs to the identified set - general equilibrium selection

other results

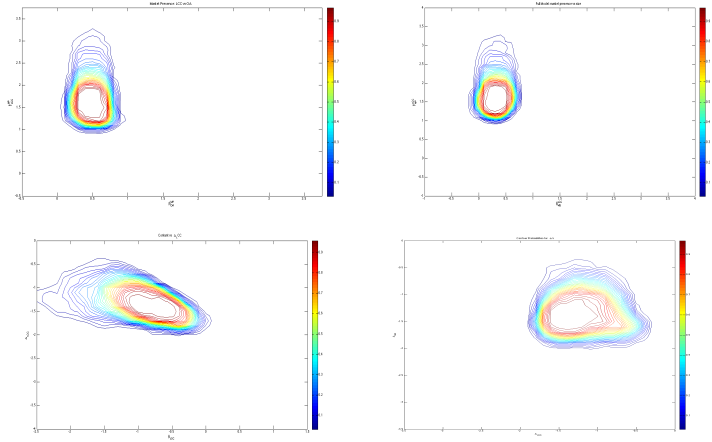

Figure: Posterior Probability that a particular parameter value belongs to the identified set

inference approach 3

this is work in progress with Xiaohong Chen and Tim Christensen.

we combine ideas from top two projects to propose an inference theory for *identified sets* based on simulations in semiparametric likelihood models.

question: given a likelihood function where the parameter is not identified, can we use simulations from this likelihood to build confidence regions for the identified set?

again,

we are interested in inference on θ in $p(\theta, g)$.

in particular, when g is parametric, standard inference is possible, but happens to our baseline estimates of θ as less assumptions are made on g? i.e., in the baseline model, $g \in \mathcal{G}$, a parametric class; what happens as we *enlarge* $\mathcal{G}$?

idea: just embed/approximate $\mathcal{G}$ with a sieve space $\mathcal{G}^{(k)}$ and then use draws from the overall density to build a confidence region for θ.

interesting because:
start with a fully parametric model, provide misspecification robust confidence regions with respect to a subset of suspect assumptions (that we pick)

can harness the recent advances in computational statistics (and economics) to simulate possibly complicated models with a large number of parameters.

ideas

Inference on a set Θ_I that maximizes (minimizes) a (general) objective function $Q(\theta)$ has been examined in Chernozhukov, Hong and Tamer (2007) and Romano and Shaikh (2010) among others.

in the case when $Q(\theta)$ is a likelihood, and we draw a large sample from "$Q_n(\theta)$" (more on "" $Q_n(.)$" later), can we show that this sequence of simulated θ's are useful in building a confidence region for the identified set Θ_I? (why? ... easy to compute)

The question is: can draws from a (flat in the limit) posterior have a frequentist interpretation in terms of coverage of the identified set?

even in parametric problems, we do not know of Bernstein VonMises like results for partially identified likelihoods. We try and do this here.

some technical insights

(some definitions): parametric model

$$\mathcal{P} = \{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$$

$$X_1, \ldots, X_n \sim p_0 \in \mathcal{P} \quad iid$$

$$\theta \mapsto p_\theta \quad \text{not injective}$$

$$\Theta_I = \{\theta \in \Theta : \; p_\theta = p_0\} \quad \text{identified set}$$

$$\mathbb{P}_n f = n^{-1} \sum_{i=1}^{n} f(X_i) \quad \text{and} \quad L_n(p) \sim \quad \text{log likelihood}$$

$$\widehat{\Theta}_\alpha^{CHT} = \{\theta \in \Theta : Q_n(\theta) \leq \xi_{n,\alpha}\}$$

where $\xi_{n,\alpha}$ quantile of LR statistic (which has a nonstandard and tedious distribution...)

what we do and show in parametric settings (first)

- use following "posterior" to generate (via simulations) sample of θ's $(\theta^1, \ldots, \theta^S)$

$$\Pi_n(d\theta|X_1, \ldots, X_n) = \frac{e^{L_n(p_\theta)}\Pi(d\theta)}{\int_\Theta e^{L_n(p_\theta)}\Pi(d\theta)}$$

- Show that this simulated sample of θ's, under some conditions, provide a confidence like region$\hat{\Theta}_\alpha$ (frequentist sense) for Θ_I. Think of it as

$$\hat{\Theta}_\alpha = \{\theta \in \Theta : \; Q_n(\theta) \leq \xi^{mc}_{n,\alpha}\}$$

simple to show in standard point identified models

since heuristically,

$$L_n(p_0) = L_n(\hat{\theta}) - \frac{1}{2}n(\theta_0 - \hat{\theta})' I_0(\theta_0 - \hat{\theta}) + o_{P_0}(1)$$

This expansion is not available here since *argmax* of likelihood is not unique.

idea for our approach

- (from inference method 2), if we have a separable model (a standard reduced form connected to the structural parameters), then inference on Θ_I can be easily obtained through the model by mapping standard confidence regions on reduced form to confidence regions on Θ_I.
- here, the reduced form is $p_0 \in \mathcal{P}$, the true density of X_i; this density is well identified.
- we know that the posterior contracts to p_0 (in the KL divergence metric) and so even if θ is not point identified, $p_{\hat{\theta}}$ approaches p_0 (appropriately).
- So, we embed the parametric model class $\mathcal{P}$ into a manageable class of parametric *pseudo-true models* for which we derive a Bernstein Von Mises type result for the likelihood ratio of this pseudo true model (this is simpler)... and then revert back to $\mathcal{P}$ to get correct coverage.
- key difficulty: ensuring that the map of the (flat) prior π in the pseudo-true space is "proper..."

result

Result 1

under ...,

$$\lim_{n \to \infty} \mathbb{P}(\Theta_I \subseteq \widehat{\Theta}_\alpha) = \alpha$$

Here, $\widehat{\Theta}_\alpha$ is easily constructed using simulations.

same approach for semiparametric models

which is the important piece of this research agenda... as the main requirement for applications is how user friendly the methods are.

world of mcmc here

it matters... and we are still working on this but for complicated likelihoods with many parameters, tuning the sampler and using clever combinations of various mcmc procedures is key. more on this in the context of our application is next.

Ex 3: Application: are HMR's results sensitive to heteroskedasticity?

- again, HMR study the extensive margin of trade using a structural model estimated with current trade data.
- introduce a "cut-off" a_L for productivity: if a random draw for productivity from country i to j, $a_{ij} > a_L$ then i trades with j. Otherwise, no trade occurs. So, in their notation:

$$m_{ij} = \beta_{01} + \lambda_{j1} + \chi_{i1} + ... + f(z_{ij}^*) + u_{ij}$$

$$z_{ij}^* = \beta_{02} + \lambda_{j2} + \chi_{i2} + \beta_{12} \log dist_{ij} + ... + u_{ij} + \nu_{ij}$$

where $\lambda_i^1, \lambda_i^2, \chi_i^1, \chi_j^2$ are country fixed effects (> 200), (u_{ij}, ν_{ij}) is jointly normal and $f(.)$ is a known *nonlinear function*. We only observe m_{ij} when $z_{ij}^* > 1$.

Ex3 assumptions

This is a structural model that is fully parametrized:

- Pareto distribution on productivity a (which leads to a particular functional form for $f(.)$)
- joint normality of the errors
- heteroskedasticity - which is well known to be a problem in trade data (and is especially problematic in nonlinear selection like models)

As above, we report below HMRs favorite model and also formal sensitivity intervals.
The inference approach here will be described below.

Ex3 results

bivariate selection model estimated via MLE (similar estimates using NLS -)

Pref. Model	Outcome Equation	Selection Equation
log distance	$-.03265$ [-0.19,0.13]	-0.165 [-0.212,-0.1125]
border	1.9548 [1,4348,2.4747]	0.2527 [0.0927,0.41]
legal system	0.1747 [0.0369,0.3107]	-0.0532 [-0.0932,-0.0132]
⋮		

Standard 95% confidence intervals (? fixed effects and ? parameters)

our mcmc procedure

here, we can write down the full likelihood of the model.

we are worried about heteroskedasticity. so we allow for *multiplicative heteroskedasticity* as a flexible function (use a polynomial of order 4).

there is evidence in the literature that generally, parametric selection models may be sensitive to heteroskedasticity and that also point identification is not simple to get.

our mcmc procedure

- similar to HMR, we have *27 parameters*, in addition to a full set of fixed effects (ran country fixed effects ¿200 and then continent fixed effects).
- We tried a set of various implementations of mcmc algorithms starting with *random block random walk metropolis hastings* algorithm.
- But, we ended up using a Equi-Energy mcmc method proposed by Kou, Zhou and Wong (2006) with random block random walk metropolis-hasting local moves. The *EE-MCMC* algorithm is designed to overcome simulating a target distribution when it is multidimensional and multimodal where the modes can be far away from each other.

comments on computational mcmc

- requires many tuning parameters - energy levels (H's) at which density (energy function) is truncated, temperatures (used to "flatten" truncated densities) proposal density for local MH algorithm, and jump probabilities. The tuning seems to be problem specific and it may be that other mcmc algorithm may work better in different problems.
- The EE algorithm we use ....

Ex3 results WHAT IS LEGAL SYSTEM HERE

bivariate selection model estimated via MLE (similar estimates using NLS -)

Pref. Model	Outcome Equation	Selection Equation
log distance	$-.03265$	-0.165
	$[-0.19,0.13]$	$[-0.212,-0.1125]$
	[-0.1089, 0.0911]	[-0.517, -0.8122]
border	1.9548	0.2527
	[1.4348,2.4747]	[0.0927,0.41]
	[0.262, 1.954]	[-0.022, 1.073]
legal system	0.1747	-0.0532
	[0.0369,0.3107]	$[-0.0932,-0.0132]$
	[-0.549, -0.048]	
⋮		

conclusion and future directions

- the message is that incorporating model uncertainty in a theoretically valid and computationally attractive way is worthy.
- Empirical economists recognize the fact that data alone are not informative. Data + model are.
- Then, our job is to understand the sensitivity of our result at least to the part of the model that some experts and/or policy makers are most worried about.
- the agenda formalizes inference approaches in semiparametric models with lack of point identification.
- This is another approach in that direction.

likelihoods, moment conditions,...

- why likelihood?: starting point, provides sharp characterizations, etc
- possible to extend this to moment conditions. some work in this direction is: Chen, Pouzo and Tamer (2013) and Tao (2014).
- other relevant areas for future development is work at the intersection of stochastic computations/algorithms and inference and statistics.

Future Work

- Chen, Pouzo and Tamer extend the methods here to moment (equality) based models with unknown functions and partial identification.
- Need to investigate whether our inference procedures are uniformly valid over any DGP p_0.
- Also need to address choice of sieve dimensions (smoothing parameters)
- It might be possible to use *parametric bootstrap* for inference where we generate data from the estimated (sieve) density under the null. This procedure might have better small sample performance.

Takeways for Empirical work with Semiparametric Likelihoods

- Using **Criterion Based Sieve LR statistic** as the basis for inference (tests and CI's) has a robustness property to point identification, non-regular rates.
- Negatives: 1) Computational (perhaps just stick to a few extra elements in the expansion; 2) gives impression that identification does not matter - you get what you have flavor. But, this is dangerous because of potential misspecification + costly computation!