Artificial Intelligence for Structural Estimation

UC San Diego - Econometrics Seminar

T. Kaji E. Manresa G. Pouliot

University of Chicago New York University University of Chicago

San Diego, October 22nd - 2019

Introduction

Structural Estimation

- Structural estimation allows us to learn about the effects of policies that have not yet been implemented.
- In spite of fully specified models, the likelihood is often intractable and does not exist in closed form.
- *Simulated Minimum Distance*, methods minimize a user-specified distance between observed data and data generated according to the model. E.g. simulated method of moments, or indirect inference.
- The choice of moments is partly informed by the theory.

Simulation-based methods

- Advantages
  - Conceptually straightforward
  - Freedom of the researcher to emphasize the features of the data upon to which base estimation
- Drawbacks
  - Curse of dimensionality
  - Large number of moment biases
  - It might not be obvious what features to match

This Paper (I)

- In this paper we propose a new estimator based on the Generative Adversarial Network (GANs) framework (Goodfellow et al. 2014) for structural estimation of economic models.

- The method can be understood as minimizing a data-driven distance between the real data and the simulated data.

- GANs takes advantage of advances in supervised learning algorithms to train classifiers to distinguish draws from the distribution of the data, from simulated data.

- The estimator is defined as the parameter value for which the discriminator cannot distinguish between simulated data and real data.

This Paper (II)

- We show there is a formal connection between the proposed adversarial estimator and two widely used estimators:
  1. Optimally weighted simulated method of moments
  2. Non-parametric Simulated Maximum Likelihood
- The method preserves the flexibility of indirect inference methods in that it allows the researcher to choose the aspects of the data to use in estimation.
- Recent results on adaptivity of neural networks as sieve non-parametric estimators provide a rationale for the use of these methods as classifiers.
- We use the estimator in two different empirical context:
  1. Roy Model with two location and two periods (simulated data).
  2. Dynamic optimization framework: Why do the elderly save (De Nardi, French and Jones, 2010) using HRS data

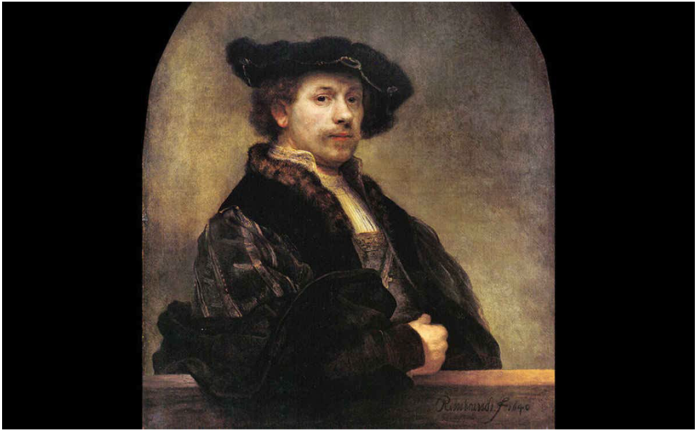
Pattern Recognition and Supervised Learning

- *Classifiers* are often a key building block of many AI algorithms.

- The success of AI technology partly hinges on the ability of machine learning algorithms to “uncover” what features of the data are useful for classification, as opposed to hard-coding what characterizes an object.

- These algorithms are trained using *supervised learning*, that is, repeated exposure of correctly labeled data.

Rembrandt



A Few of His Paintings

Rembrandt 1632

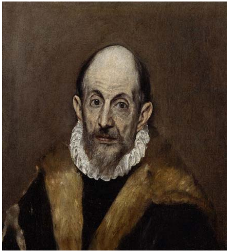
Rembrandt 1595

Rembrandt 1606

Rembrandt 1630

A Few of Other's Paintings

Da Vinci 1503

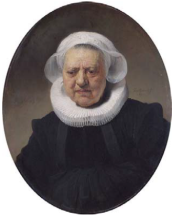


El Greco 1559

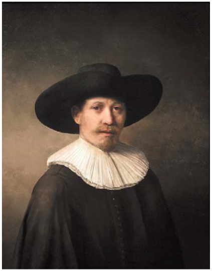
Picasso 1939

Tiziano 1567

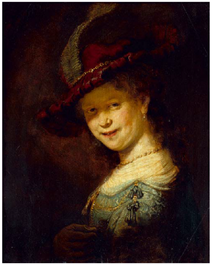
Which Is the Impostor Painting?



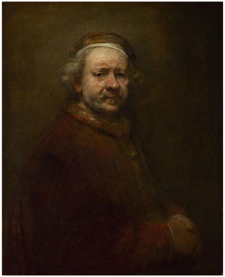
Rembrandt?



Rembrandt?



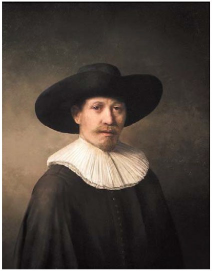
Rembrandt?



Rembrandt?

Which Is the Impostor Painting?

Rembrandt 1634



~~Rembrandt~~

AI

Rembrandt 1633

Rembrandt ~ 1669

The Model and Estimator

The Model

- Individual outcomes, $y_i \in \mathbb{R}^L$, are a function, G, of a finite dimensional vector of parameters $\theta_G \in \Theta \subset \mathbb{R}^K$, exogenous variables, $x_i \in \mathbb{R}^M$, and an error, $\varepsilon_i \in \mathbb{R}^L$ whose distribution is known. G is *not necessarily* available in closed form.
- Given θ_G, and $\{\varepsilon_i\}_{i=1}^H$, we assume it is computationally feasible to obtain a large sample of size H, $(\tilde{y}_i(\theta_G))_{i=1}^H$ of generated/synthetic data:

$$\tilde{y}_i(\theta_G) = G(\theta_G; x_i, \varepsilon_i).$$

- Object of interest is θ_G or a function of it.
- We assume the model is correctly specified.

The Estimator

Let $h_i = h(y_i, x_i) \in \mathbb{R}^d$ be a d-dimension vector of functions of the data, and $\tilde{h}_i(\theta_G) = h(\tilde{y}_i(\theta_G), x_i)$ be the same functions using simulated data.

For example, h_i could be a subvector of (y_i, x_i).

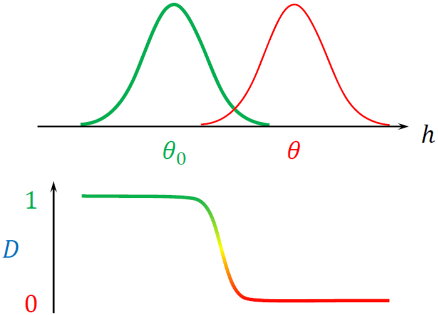
Let $D : \mathbb{R}^D \to [\varepsilon, 1-\varepsilon]$ belong to a pre-specified class of functions. We define the estimator $\widehat{\theta}_G$ as the solution to the following program:

$$\min_{\theta_G} \max_{D \in \mathcal{D}} \frac{1}{N} \sum_{i=1}^{N} \log(D(h_i)) + \frac{1}{H} \sum_{i=1}^{H} \log(1 - D(\tilde{h}_i(\theta_G))).$$

- D provides a prediction of the probability that a vector h was drawn from the true distribution or the simulated distribution for a given θ_G. The function D acts as a critic.
- G maximizes missclassification of simulated data into true data. The economic model acts like a forger.

In Pictures

$$\max_{D} \mathbb{E}[\log D(h)] + \mathbb{E}[\log(1-D)(h_{\theta})]$$

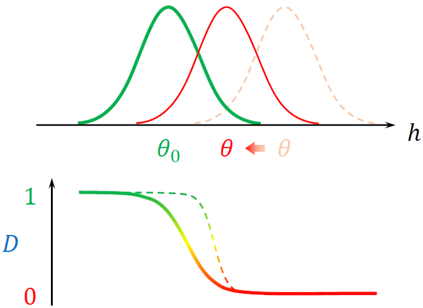


⇨

Objective function is close to 0

In Pictures

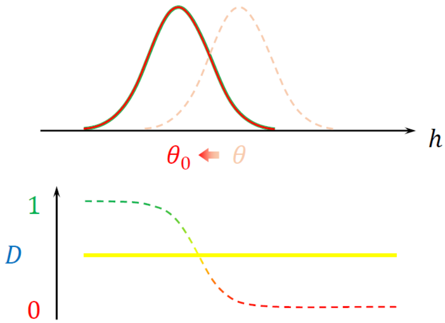
$$\max_{D} \mathbb{E}[\log D(h)] + \mathbb{E}[\log(1 - D)(h_{\theta})]$$



Objective function is less than 0

In Pictures

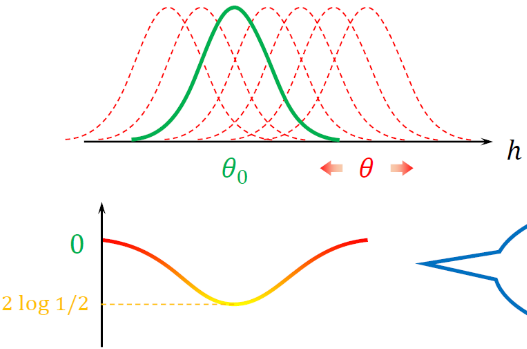
$$\max_{D} \mathbb{E}[\log D(h)] + \mathbb{E}[\log(1-D)(h_{\theta})]$$



Objective function
is equal to $2\log 1/2$

In Pictures

$$\max_{D} \mathbb{E}[\log D(h)] + \mathbb{E}[\log(1-D)(h_{\theta})]$$



Objective function is minimized at $\theta = \theta_0$!!

Comments

- This estimation approach is referred as “adversarial” in the Machine Learning community.

- This approach was first introduced by Goodfellow et al. (2014) to estimate models to generate images. In that case, both the discriminator and the structural models are deep neural networks.

- The function $\widehat{D}(\theta_G)$ induces a data-driven distance between the distribution of the synthetic data and the distribution of the true data. The estimator minimizes this distance.

- Different choices of D and h_i produce different estimators of θ_G.

Example 1: Logit Discriminator

Let $D(t) = \frac{1}{1+e^{-t}}$. The estimator is then:

$$\widehat{\theta}_G = \underset{\theta_G}{\operatorname{argmin}} \max_{\theta_D} \frac{1}{N} \sum_{i=1}^{N} \log(D(\theta_D' \cdot h_i)) + \frac{1}{H} \sum_{i=1}^{H} \log(1 - D(\theta_D' \cdot \tilde{h}_i(\theta_G))).$$

- Interpretation: when $H = N$, the inner maximization is a logit maximum likelihood estimation problem, where the outcome variable is 1 if data is true, and 0 otherwise.
- The F.O.C. of the inner maximization problem for a given θ_G is:

$$\frac{1}{N} \sum_{i=1}^{N} (1 - \widehat{D}(h_i)) h_i - \frac{1}{H} \sum_{i=1}^{H} \widehat{D}(\tilde{h}_i(\theta_G)) \tilde{h}_i(\theta_G) = 0,$$

- When $\theta_G = \theta_G^0$, $\widehat{\theta}_D = o_p(1)$, hence:

$$\frac{1}{N} \sum_{i=1}^{N} h_i - \frac{1}{H} \sum_{i=1}^{H} \tilde{h}_i(\theta_G^0) \approx 0.$$

Example 2: Oracle Discriminator

When N and H go to infinity and D is any continuous differentiable function, the solution to the inner maximization problem is:

$$D^*(h, \theta_G) = \frac{f_{\theta_G^0}(h)}{f_{\theta_G^0}(h) + f_{\theta_G}(h)}.$$

where $f_{\theta_G^0}(h)$ is the pdf of h_i, and $f_{\theta_G}(h)$ is the pdf of $\tilde{h}_i(\theta_G)$.

Hence, $\widehat{\theta}_G$ **minimizes the Jensen-Shannon distance** between the distribution of h_i and the distribution of $\tilde{h}_i(\theta_G)$

$$\min_{\theta_G} \int \log\left(\frac{f_{\theta_G^0}(h)}{f_{\theta_G^0}(h) + f_{\theta_G}(h)}\right) f_{\theta_G^0}(h)dh + \int \log\left(\frac{f_{\theta_G}(h)}{f_{\theta_G^0}(h) + f_{\theta_G}(h)}\right) f_{\theta_G}(h)d$$

The oracle estimator is **efficient**.

Example 3: AI/ML Discriminators

- Mutli-layer Neural Networks (e.g. 2 layers):

$$D^N(h, \theta_D^N) = S(\alpha_0 + \sum_{k=1}^{d_2} \gamma_k S(\alpha_{0,k} + \sum_{j=1}^{d_1} \gamma_j^k S(\alpha_{0,j}^k + \lambda_j^{k\prime} h)))$$

where S is an activation functions and $\theta_D^N = (\alpha_{0,j}^k, \lambda_j^k, \gamma_k, \alpha_0)$.

- The Universal approximation properties of 1-hidden NN are well known now for decades (e.g. Chen and White, 1999).

- Recent results show that the convergence rate of multilayer NN in different estimation context depend on $d^* < d = dim(h_i)$ (e.g. Mhaskar and Poggio (2017), Bach (2017), Bauer and Kohler (2019)).

- Other algorithms such as Random Forest (e.g. Athey, Tibshirani, Wager (2019)) or k-means also show adaptivity properties.

Implementation

Computation

Recall the optimization problem:

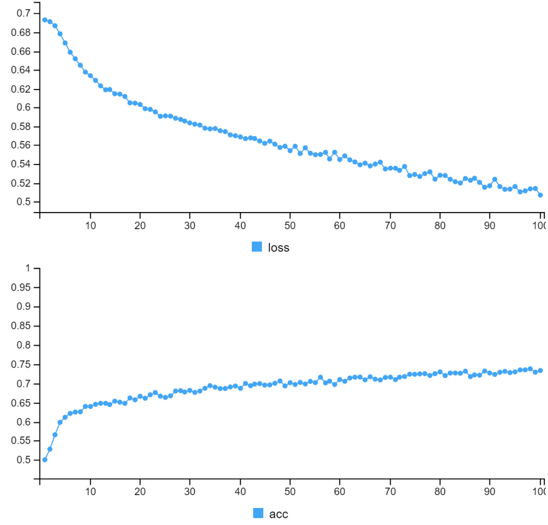
$$\min_{\theta_G} \max_{D \in \mathcal{D}} \frac{1}{N} \sum_{i=1}^{N} \log(D(h_i)) + \frac{1}{H} \sum_{i=1}^{H} \log(1 - D(\tilde{h}_i(\theta_G))).$$

We implement the following iterative algorithm. Starting with a random initial $\theta = \theta^{(0)}$:

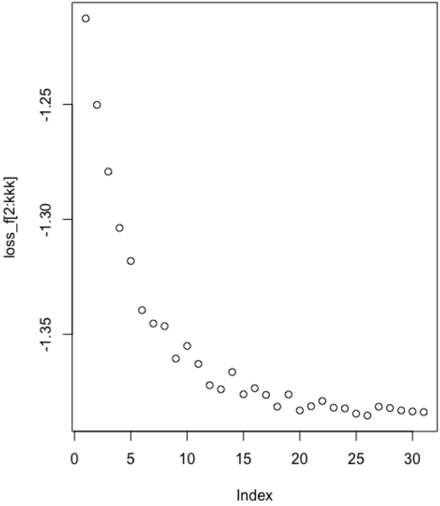
1. For given $\theta = \theta^{(s)}$, we solve the inner maximization problem to obtain $\widehat{D}^{(s)}$ using h_i and $\tilde{h}_i(\theta^{(s)})$. E.g. Estimate a logit regression by maximum likelihood, or a NN.
2. Given $\widehat{D}^{(s)}(\cdot)$ we compute the gradient of the objective function (numerically) with respect to θ and update $\theta^{(s+1)}$ with 1 step gradient descent.

When D is a NN we make use of off-the-shelf routines relying on stochastic gradient descent and back-propagation algorithms for efficient computation of $\widehat{D}$.

Step 1: Training NN using off-the-shelf optimization routine



Step 2: Gradient descent



Misspecification

- When the model is misspecified the criterion can still be interpreted as a minimization of a distance.

- The discriminator will always be able to “distinguish” between the two distributions.

- The adversarial framework can be easily combined with robust inference proposed in Bonhomme and Weidner (2019), where the estimator is adjusted in the direction of the score of a *larger reference model*.

- The score of the model can be obtained as a by-product of estimating of the oracle discriminator:

$$\frac{\partial log(f_\theta(y))}{\partial \theta_G} = \frac{1}{1 - D^*(y,\theta)} \frac{\partial (log(1 - D^*(y,\theta_G)))}{\partial \theta_G}$$

Statistical Properties

Logit Discriminator

Assume the following high-level conditions:

1. $\sup_{\theta_G} \|\hat{\theta}_D(\theta_G) - \theta_D^0(\theta_G)\| = o_p(1)$
2. $\theta_D^0(\theta_G) = 0$ if and only if $\theta_G = \theta_G^0$.
3. $\sqrt{N}(\hat{\theta}_D(\theta_G^0) - \theta_D^0(\theta_G^0)) \to N(0, V_{\theta^0})$

where $\theta_D^0(\theta_G) = \operatorname*{plim}_{N,H} \hat{\theta}_D(\theta_G)$.

Then, $\sqrt{N}\left(\widehat{\theta}_G - \theta_G^0\right) \to N(0, 2 \cdot V)$, where

$$V = \left[\left(\operatorname*{plim}_{H\to\infty} \frac{1}{H}\sum_{i=1}^{H} \frac{\partial \tilde{h}_i(\theta_G^0)}{\partial \theta_G}'\right)\left(\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} h_i' \cdot h_i\right)^{-1}\left(\operatorname*{plim}_{H\to\infty} \frac{1}{H}\sum_{i=1}^{H} \frac{\partial \tilde{h}_i(\theta_G^0)}{\partial \theta_G}\right)\right]^{-1}$$

- It can be seen that this estimator has the same asymptotic distribution (to first order) to optimally weighted Simulated Method of Moments (SMM) with $\frac{1}{N}\sum_{i=1}^{N} h_i$ as moments.

The Oracle Case: $D = D^*$

Recall

$$D^*(h,\theta) = \frac{f_{\theta^0}(h)}{f_{\theta^0}(h) + f_{\theta}(h)}.$$

Under regularity conditions we have the following expansion:

$$\sqrt{N}\left(\widehat{\theta}_G - \theta_G^0\right) = M^{-1}\left[\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial \log f_{\theta_G^0}(h_i)}{\partial \theta_G} - \frac{\sqrt{N}}{H}\sum_{i=1}^{H}\frac{\partial \log f_{\theta_G^0}(\tilde{h}_i(\theta_G^0))}{\partial \theta_G}\right)\right]$$

where

$$M \longrightarrow \mathbb{E}_{\theta_G^0}\left(\left(\left.\frac{\partial \log f_{\theta_G}(h)}{\partial \theta_G}\right|_{\theta_G^0}\right)^2\right)$$

with $\mathbb{E}_{\theta_G^0}$ denoting the expectation taken with respect to the true distribution of the data.

- If $N/H \to 0$ and $\frac{1}{\sqrt{H}}\sum_{i=1}^{H}\frac{\partial \log f_{\theta_G^0}(\tilde{h}_i(\theta_G^0))}{\partial \theta_G} = O_p(1)$ the oracle estimator is efficient.

The Non-parametric case

We analyze the properties of the GANs estimator in the context of a parametric generative model. A set of sufficient conditions is:

1. Entropy conditions on the family of D
2. Support conditions on D
3. Correct specification and identification conditions
4. Rate of convergence of discriminator in bce metric is $o_p(N^{-1/4})$
5. Orthogonality condition to obtain $\sqrt{N}$ estimable θ_G
6. Differentiability of f_θ as well as $G(\theta;\varepsilon)$
7. $N/H \to 0$

Then

$$\sqrt{N}(\widehat{\theta}_G - \theta_G^0) \to N(0, I_{\theta_0}^{-1})$$

where $I_{\theta_0}^{-1}$ is the information matrix.

Monte Carlo Simulation

Roy Model with two locations

- Model with individual and sector-specific comparative advantage, as well as sector-specific returns to experience.
- Individuals choose to work among two sectors in exchange of a salary.
- There are two periods, and location choice in each period is denoted d_{i1} and d_{i2} respectively.
- Salary in the first period:

$$\log w_{i1} = \mu_{d_{i1}} + \sigma_{d_{i1}} \varepsilon_{i d_{i1} 1}$$

- Salary in the second period:

$$\log w_{i2} = \mu_{d_{i2}} + \gamma_{d_{i2}} 1\{d_{i1} = d_{i2}\} + \sigma_{d_{i2}} \varepsilon_{i d_{i2} 2},$$

- Individuals choose locations d_{i1} and d_{i2} to maximize their discounted stream of expected wages over the two periods.
- Data is wages and location choices. Structural parameters are μ_1, μ_2, γ_1, γ_2, σ_1, σ_2, discount factor β and correlation between shocks ρ_s and ρ_t.

Estimators (I)

Logit-Discriminator We choose the following 9-dimensional vector of predictors:

$$h_i = \left(d_{i1}, d_{i2}, w_{i1}, w_{i2}, d_{i1} \cdot d_{i2}, d_{i1} \cdot w_{i1}, d_{i2} \cdot w_{i2}, w_{i1}^2, w_{i2}^2\right).$$

The predictors aim at capturing first and second moments of the wage functions as well as dependence between choice of location driven by returns to experience.

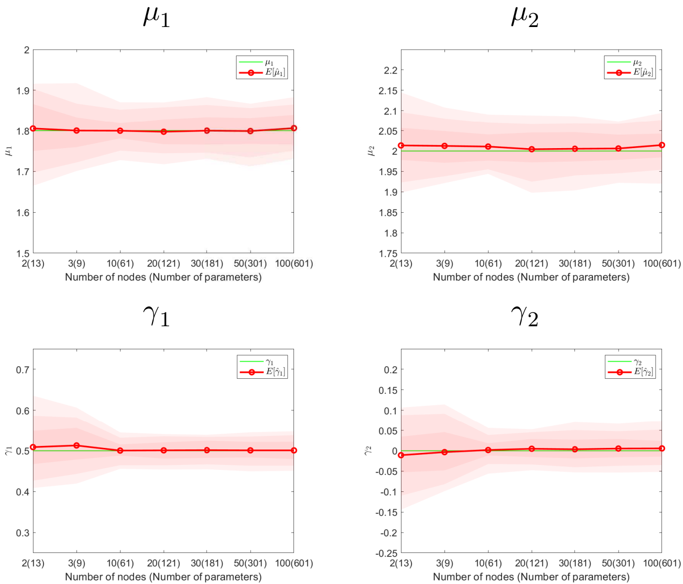
Indirect Inference We select the following 9 moments: the regression coefficients of w_{it} on a constant and d_{it-1} as well as the variance of the residuals, separately for observations in each location, the regression coefficients of d_{it} on a constant and d_{it-1}, and the correlation between the residuals of the first regression. We use the "Nelder-Mead" optimization algorithm. The synthetic moments are computed using 10 different sets of shocks and we take the sample mean across them.

Estimators (II)

NN-Discriminator Using as inputs $h_i = (w_1, d_1, w_2, d_2)$, we estimate the model using 1-hidden layer NN for nodes ranging from 2 to 100 as the discriminator. We compute the estimator with R using the keras package to define and optimize the NN. We use the iterative algorithm using 5 different initial conditions and we define the estimator as the one with lowest criterion.

Random Forest Using as inputs $h_i = (w_1, d_1, w_2, d_2)$, we estimate the model with random forest with 100 trees as discriminator. In each tree 2 variables chosen at random are used for classification. We use the default settings of the R package randomForest to choose the depth of the trees.

Results with different NN (I)



μ_1

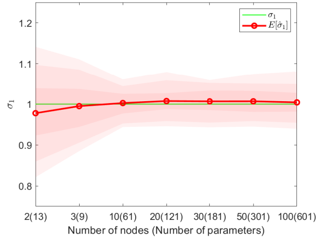
μ_2

γ_1

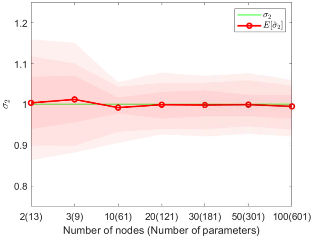
γ_2

Results with different NN (II)

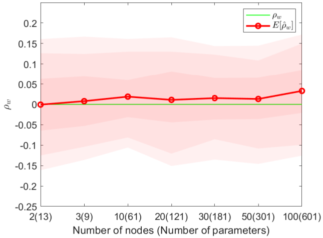
σ_1



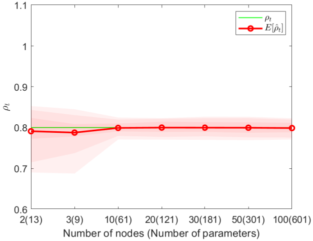
σ_2



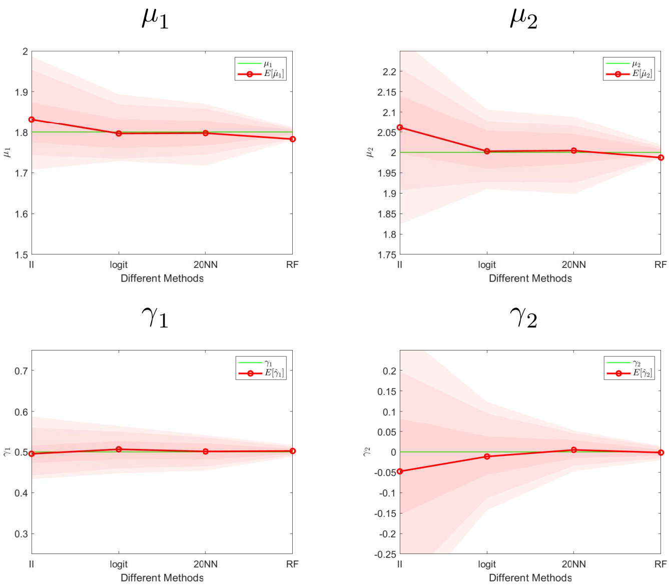
ρ_w



ρ_t



Results with different methods (I)



μ_1

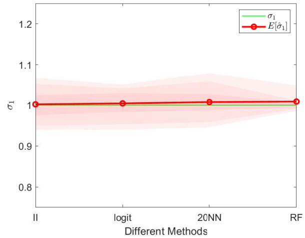
μ_2

γ_1

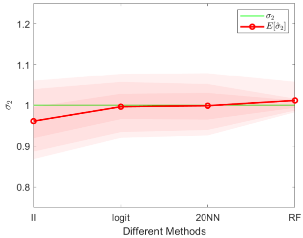
γ_2

Results with different methods (II)

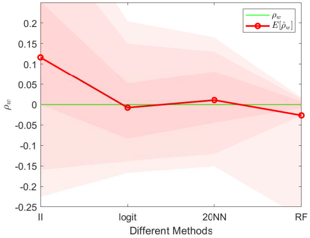
σ_1



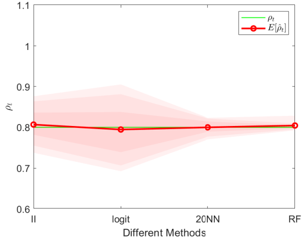
σ_2



ρ_w



ρ_t



Empirical Application

Why do the Elderly Save? (I)

De Nardi, French, Jones (JPE, 2010)

- We explore the usefulness of the adversarial estimation framework in rationalizing the saving patterns of the old in the AHEAD data.

- DFJ emphasize three different motives for precautionary savings:
  1. Lifespan risk
  2. Out-of-pocket medical expense risk (towards end of life)
  3. Bequest motives

- Disentangling the different channels is important in order to evaluate programs such as Medicare and Medicaid, for example.

- The author's estimation strategy is based on matching median assets by cohort, permanent income quintile, and period of observation. There is a total of 120 moments and 2968 individuals.

Why do the Elderly Save? (II)

- The matching moment estimator is uninformative about the importance of the bequest motive.

- We consider the adversarial framework estimation for two different choices of the discriminator:

  1. NN discriminator based on individual profiles of assets, lifespan profiles, permanent income rank, and age in 1996.
  2. NN discriminator based on the same variables as before as well as individual profile of health status, and gender.

- As we will see, health status profiles and gender are an important source of identification of medical expense and bequest motives.

- Including health status profiles is infeasible in a SMM framework due to the curse of dimensionality.

The Model

- Model for heterogeneous single retirees, aged 72 and above, who are out of the labor force.
- Agents obtain utility from consuming and leaving bequests:

$$u(c) = \frac{c^{1-\nu}}{1-\nu} \qquad \phi(e) = \theta \frac{(e+k)^{1-\nu}}{1-\nu}.$$

- Agents take optimal savings and consumption decisions by maximizing the sum of the discounted stream of utilities, subject to their budget constraint
- Agents face three different types of uncertainty: health status, medical expenses, and survival. These shocks are conditional on gender, age, health status and permanent income.
- In addition, there is an endogenous consumption floor ($\underline{c}$), by which the government ensures at least a level of consumption equal to $\underline{c}$.

Data

- Data is from the Assets and Health Dynamics of the Oldest Old (AHEAD) collected by the University of Michigan from 1994 to 2006 every 2 years. The sample consists of non-institutionalized individuals, aged 70 or above in 1994.

- We follow the authors and focus on single and retired individuals, which are 3,200 individuals (600 males and 2600 females).

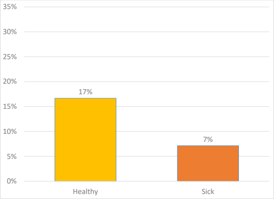
- The survey collects information on age, financial wealth, non-asset income, and medical expenses. The mean yearly spending is $\$3,700$, with standard deviation $\$13,400$.

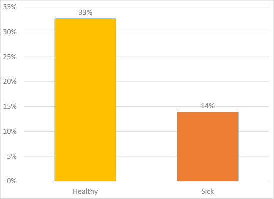
- A measure of permanent income ranking is inferred out of the ranking on individual average income.

Health shocks × gender as a source of identification

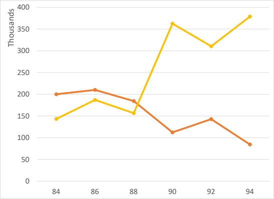
Men **Women**

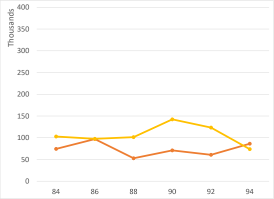
5-year survival at 85





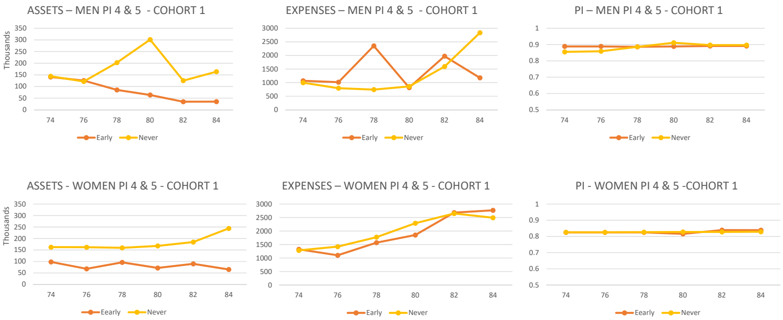
Asset Profiles (Cohort 3, 40% upper PI)



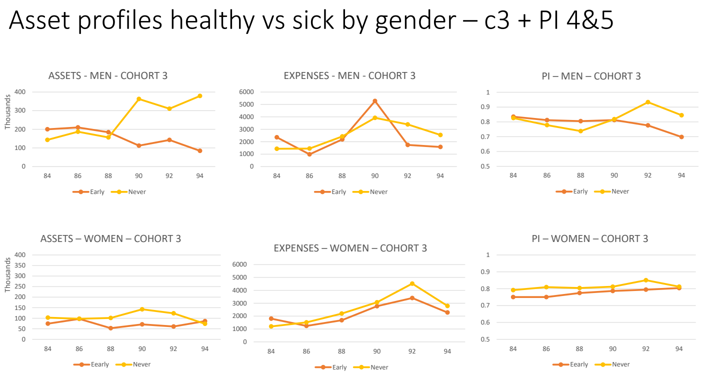


Cohort 1

Asset profiles healthy vs sick by gender – c1 + PI 4&5



Cohort 3



Asset profiles healthy vs sick by gender – c3 + PI 4&5

ASSETS - MEN - COHORT 3

EXPENSES - MEN - COHORT 3

PI – MEN – COHORT 3

ASSETS – WOMEN – COHORT 3

EXPENSES – WOMEN – COHORT 3

PI – WOMEN – COHORT 3

Parameter Estimates

	DFJ (2010)		Adversarial		Adversarial $h_t \times g$	
	(1)	(2)	(3)	(4)	(5)	(6)
β	$.97$	$.97$	$.97$	$.97$	$.97$	$.97$
ν	3.81 $(.50)$	3.84 $(.55)$	2.8 $(.039)$	5.00 $(.040)$	6.50 $(.016)$	5.50 $(.014)$
$\underline{c}$ (\$)	$2,663$ (346)	$2,665$ (353)	$2,838$ (119.48)	$4,475$ (180.16)	$4,797$ (14.51)	$4,475$ (19.84)
θ	0	$2,360$ $(8{,}122)$	0	8.48 (6.83)	0	119.16 (6.56)
MPC	1	0.12 (NA)	1	0.4 $(.039)$	1	0.3 $(.002)$
k (000)	–	273 (446)	–	10	–	10
$\underline{a}$ (\$)	–	$36,215$	–	$6,534$	–	$4,200$
$Loss$			-1.0819	-1.0763	-1.0808	-1.0832

Preliminary results. Results from specifications (1) and (2) are taken from DFJ (2010), using a SMM estimator. Specifications (3) and (4) use the Adversarial estimator with discriminator 20 neuron 1 hidden layer network with 14 inputs. Specifications (5) and (6) use the same discriminator with 21 inputs. Specifications (3) to (6) use a more balanced panel subsample of the original DFJ sample. Standard errors in (3) to (6) are computed as the variance of the estimated score.

Conclusion

Conclusion

- We investigate the use of the Adversarial estimation framework first introduced in Computer Science for structural estimation.
- The method shows close connections with other indirect inference methods that attain efficiency, such as the ones proposed by Nickl and Potscher (2010) or Fermanian and Salanie (2004).
- The method can be thought as a way to choose a data-driven distance between the distribution of the data and the generated data.
- The use of NN as discriminators have the potential to deal with the curse of dimensionality and uncover an "efficient" way to compare the two distributions.
- Computation is transparent.