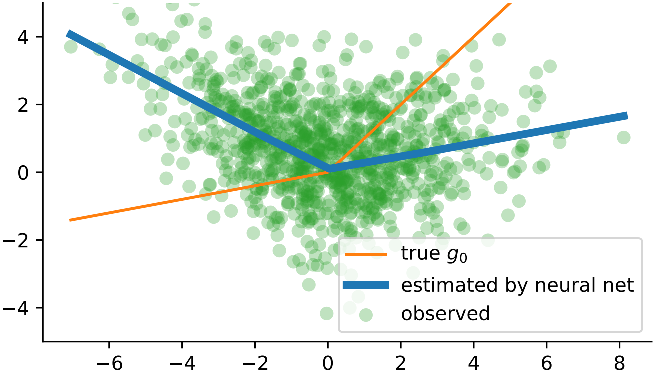
Deep Generalized Method of Moments for Instrumental Variable Analysis

Andrew Bennett, Nathan Kallus, Tobias Schnabel

Endogeneity

- $g_0(x) = \max(x,\, x/5)$
- $Y = g_0(X) - 2\epsilon + \eta$
- $X = Z + 2\epsilon,\; Z, \epsilon, \eta \sim \mathcal{N}(0,1)$



IV Model

- $Y = g_0(X) + \epsilon$
  - $\mathbb{E}\epsilon = 0$, $\mathbb{E}\epsilon^2 < \infty$
  - $\mathbb{E}\left[\epsilon \mid X\right] \neq 0$
    - Hence, $g_0(X) \neq \mathbb{E}\left[Y \mid X\right]$
- Instrument Z has
  - $\mathbb{E}\left[\epsilon \mid Z\right] = 0$
  - $\mathbb{P}\left(X \mid Z\right) \neq \mathbb{P}\left(X\right)$
- If had additional endogenous context L, include it in both X and Z
- $g_0 \in \mathcal{G} = \{g(\,\cdot\,;\theta) : \theta \in \Theta\}$
  - $\theta_0 \in \Theta$ is such that $g_0(x) = g(x;\theta_0)$

IV is Workhorse of Empirical Research

Outcome Variable	*Endogenous Variable*	*Source of Instrumental Variable(s)*	*Reference*
		1. Natural Experiments	
Labor supply	Disability insurance replacement rates	Region and time variation in benefit rules	Gruber (2000)
Labor supply	Fertility	Sibling-Sex composition	Angrist and Evans (1998)
Education, Labor supply	Out-of-wedlock fertility	Occurrence of twin births	Bronars and Grogger (1994)
Wages	Unemployment insurance tax rate	State laws	Anderson and Meyer (2000)
Earnings	Years of schooling	Region and time variation in school construction	Duflo (2001)
Earnings	Years of schooling	Proximity to college	Card (1995)
Earnings	Years of schooling	Quarter of birth	Angrist and Krueger (1991)
Earnings	Veteran status	Cohort dummies	Imbens and van der Klaauw (1995)
Earnings	Veteran status	Draft lottery number	Angrist (1990)
Achievement test scores	Class size	Discontinuities in class size due to maximum class-size rule	Angrist and Lavy (1999)
College enrollment	Financial aid	Discontinuities in financial aid formula	van der Klaauw (1996)
Health	Heart attack surgery	Proximity to cardiac care centers	McClellan, McNeil and Newhouse (1994)
Crime	Police	Electoral cycles	Levitt (1997)
Employment and Earnings	Length of prison sentence	Randomly assigned federal judges	Kling (1999)
Birth weight	Maternal smoking	State cigarette taxes	Evans and Ringel (1999)

From Angrist & Krueger 2001

Going further

- Standard methods like 2SLS and GMM and more recent variants are significantly impeded when:
  - X is structured high-dimensional (*e.g.*, image)?
  - and/or Z is structured high-dimensional (*e.g.*, image)?
  - and/or g_0 is complex (*e.g.*, neural network)?
- (As we'll discuss)

DeepGMM

- We develop a method termed DeepGMM
  - Aims to addresses IV with such high-dimensional variables / complex relationships
  - Based on a new variational interpretation of optimally-weighted GMM (inverse-covariance), which we use to efficiently control very many moment conditions
  - DeepGMM given by the solution to a smooth zero-sum game, which we solve with iterative smooth-game-playing algorithms (à la GANs)
  - Numerical results will show that DeepGMM matches the performance of best-tuned methods in standard settings and continues to work in high-dimensional settings where even recent methods break

This talk

1 Introduction

2 Background

3 Methodology

4 Experiments

Two-stage methods

- $\mathbb{E}[\epsilon \mid Z] = 0$ implies

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[g_0(X) \mid Z] = \int g_0(x) d\mathbb{P}(X = x \mid Z)$$

- If $g(x;\theta) = \theta^T \phi(x)$: becomes $\mathbb{E}[Y \mid Z] = \theta^T \mathbb{E}[\phi(X) \mid Z]$
  - Leads to 2SLS: regress $\phi(X)$ on Z (possibly transformed) by least-squares and then regress Y on $\hat{\mathbb{E}}[\phi(X) \mid Z]$
  - Various methods that find basis expansions non-parametrically (*e.g.*, Newey and Powell)
- In lieu of a basis, DeepIV instead suggests to learn $\mathbb{P}(X = x \mid Z)$ as NN-parameterized Gaussian mixture
  - Doesn't work if X is rich
  - Can suffer from "forbidden regression"
    - Unlike least-squares, MLE doesn't guarantee orthogonality irrespective of specification

Moment methods

- $\mathbb{E}[\epsilon \mid Z] = 0$ implies $\mathbb{E}[f(Z)(Y - g_0(X))] = 0$
  - For any $f_1, \ldots, f_m$ implies the moment conditions $\psi(f_j; \theta_0) = 0$ where $\psi(f; \theta) = \mathbb{E}[f(Z)(Y - g(X; \theta))]$
  - GMM takes $\psi_n(f; \theta) = \hat{\mathbb{E}}_n[f(Z)(Y - g(X; \theta))]$ and sets

$$\hat{\theta}^{\text{GMM}} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \|(\psi_n(f_1; \theta), \ldots, \psi_n(f_m; \theta))\|^2$$

    - Usually: $\|\cdot\|_2$. Recently, AGMM: $\|\cdot\|_\infty$
- Significant inefficiencies with many moments: wasting modeling power to make redundant moments small
  - Hansen et al: (With finitely-many moments) this norm gives the minimal asymptotic variance (efficiency) for any $\tilde{\theta} \to_p \theta_0$:

$$\|v\|^2 = v^T C_{\tilde{\theta}}^{-1} v, \;\; [C_\theta]_{jk} = \frac{1}{n} \sum_{i=1}^n f_j(Z_i) f_k(Z_i)(Y_i - g(X_i; \theta))^2.$$

  - E.g., two-step/iterated/cts GMM. Generically OWGMM.

Failure with Many Moment Conditions

- When $g(x;\theta)$ is a flexible model, many – possibly infinitely many – moment conditions may be needed to identify θ_0
  - But both GMM and OWGMM will fail if we use too many moments

This talk

1 Introduction

2 Background

3 Methodology

4 Experiments

Variational Reformulation of OWGMM

- Let $\mathcal{V}$ be vector space of real-valued fns of Z
  - $\psi_n(f;\theta)$ is a linear operator on $\mathcal{V}$
  - $\mathcal{C}_\theta(f,h) = \frac{1}{n}\sum_{i=1}^n f(Z_i)h(Z_i)(Y_i - g(X_i;\theta))^2$ is a bilinear form on $\mathcal{V}$
- Given any subset $\mathcal{F} \subseteq \mathcal{V}$, define

$$\Psi_n(\theta;\mathcal{F},\tilde{\theta}) = \sup_{f\in\mathcal{F}} \psi_n(f;\theta) - \frac{1}{4}\mathcal{C}_{\tilde{\theta}}(f,f)$$

Theorem

Let $\mathcal{F} = \mathrm{span}(f_1,\ldots,f_m)$ *be a subspace. For OWGMM norm:*

$$\left\|(\psi_n(f_1;\theta),\ \ldots,\ \psi_n(f_m;\theta))\right\|^2 = \Psi_n(\theta;\mathcal{F},\tilde{\theta}).$$

Hence: $\hat{\theta}^{\mathit{OWGMM}} \in \mathrm{argmin}_{\theta\in\Theta}\,\Psi_n(\theta;\mathcal{F},\tilde{\theta})$.

DeepGMM

- Idea: use this reformulation and replace $\mathcal{F}$ with a rich set
  - But *not* with a hi-dim subspace (that'd just be GMM)
  - Let $\mathcal{F} = \{f(z;\tau) : \tau \in \mathcal{T}\}$, $\mathcal{G} = \{g(x;\theta) : \theta \in \Theta\}$ be all networks of given architecture with varying weights τ, θ
    - (Think about it as the union the spans of the penultimate layer functions)
- DeepGMM is then given by the solution to the smooth zero-sum game (for any data-driven $\tilde{\theta}$)

$$\hat{\theta}^{\text{DeepGMM}} \in \operatorname*{argmin}_{\theta\in\Theta} \sup_{\tau\in\mathcal{T}} \; U_{\tilde{\theta}}(\theta,\tau)$$

$$\text{where} \quad U_{\tilde{\theta}}(\theta,\tau) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} f(Z_i;\tau)(Y_i - g(X_i;\theta)) - \tfrac{1}{4n}\textstyle\sum_{i=1}^{n} f^2(Z_i;\tau)(Y_i - g(X_i;\tilde{\theta}))^2.$$

Consistency of DeepGMM

- Assumptions:
  - Identification: θ_0 uniquely solves $\psi(f;\theta) = 0\ \forall f \in \mathcal{F}$
  - Complexity: $\mathcal{F}, \mathcal{G}$ have vanishing Rademacher complexities (alternatively, can use a combinatorial measure like VC)
  - Absolutely star shaped: $f \in \mathcal{F}, |\lambda| \leq 1 \implies (\lambda f) \in \mathcal{F}$
  - Continuity: $g(x;\theta), f(x;\tau)$ are continuous in θ, τ for all x
  - Boundedness: $Y, \sup_{\theta\in\Theta} |g(X;\theta)|, \sup_{\tau\in\mathcal{T}} |f(Z;\tau)|$ bounded

Theorem

Let $\tilde{\theta}_n$ by any data-dependent sequence with a limit in probability. Let $\hat{\theta}_n, \hat{\tau}_n$ be any approximate equilibrium of our game, i.e.,

$$\sup_{\tau\in\mathcal{T}} U_{\tilde{\theta}_n}(\hat{\theta}_n, \tau) - o_p(1) \leq U_{\tilde{\theta}_n}(\hat{\theta}_n, \hat{\tau}_n) \leq \inf_{\theta} U_{\tilde{\theta}_n}(\theta, \hat{\tau}_n) + o_p(1).$$

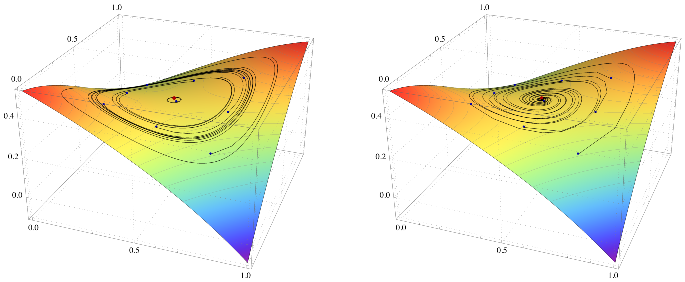
Then $\hat{\theta}_n \to_p \theta_0$.

Consistency of DeepGMM

- Specification is much more defensible when use such a rich $\mathcal{F}$
- Nonetheless, if we drop specification we instead get

$$\inf_{\theta:\psi(f;\theta)=0\ \forall f\in\mathcal{F}} \|\theta - \hat{\theta}_n\| \to_p 0$$

Optimization



- Thanks to surge of interest in GANs, lots of good algorithms for playing smooth games
- We use OAdam by Daskalakis et al.
  - Main idea: use updates with *negative* momentum

Choosing $\tilde{\theta}$

- Ideally $\tilde{\theta} \approx \theta_0$
- Can let it be $\hat{\theta}^{\text{DeepGMM}}$ using another $\tilde{\theta}$
  - Can repeat this
- To simulate this, at every step of the learning algorithm, we update it to be the last θ iterate

This talk

1 Introduction

2 Background

3 Methodology

4 Experiments

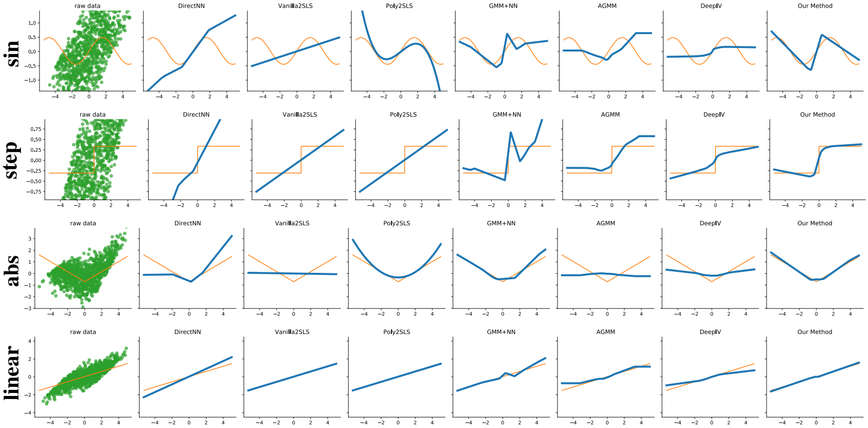
Overview

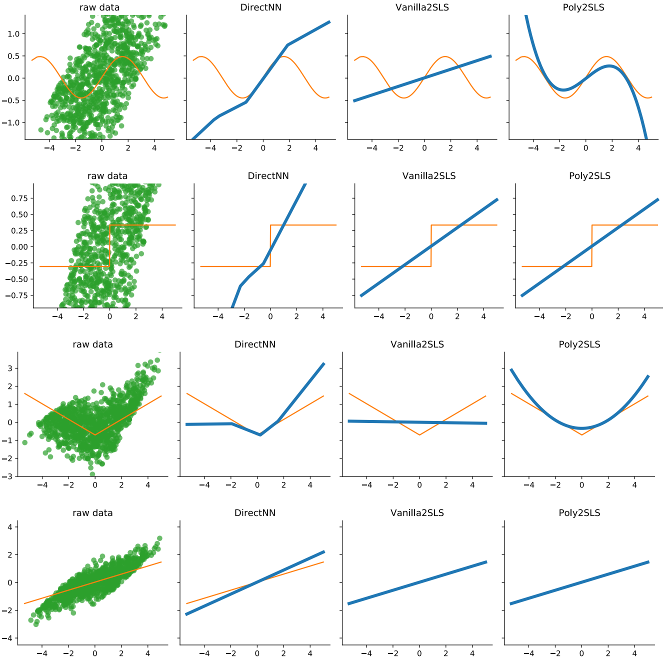
- Low-dimensional scenarios: 2-dim Z, 1-dim Z
- High-dimensional scenarios: Z, X, or both are images
- Benchmarks:
  - DirectNN: regress Y on X with NN
  - Vanilla2SLS: all linear
  - Poly2SLS: select degree and ridge penalty by CV
  - GMM+NN*: OWGMM with NN $g(x;\theta)$; solve using Adam
    - When Z is low-dim expand with 10 RBFs around EM clustering centroids. When Z is high-dim use raw instrument.
  - AGMM: `github.com/vsyrgkanis/adversarial_gmm`
    - One-step GMM with $\|\cdot\|_\infty$ + jitter update to moments
    - Same moment conditions as above
  - DeepIV: `github.com/microsoft/EconML`

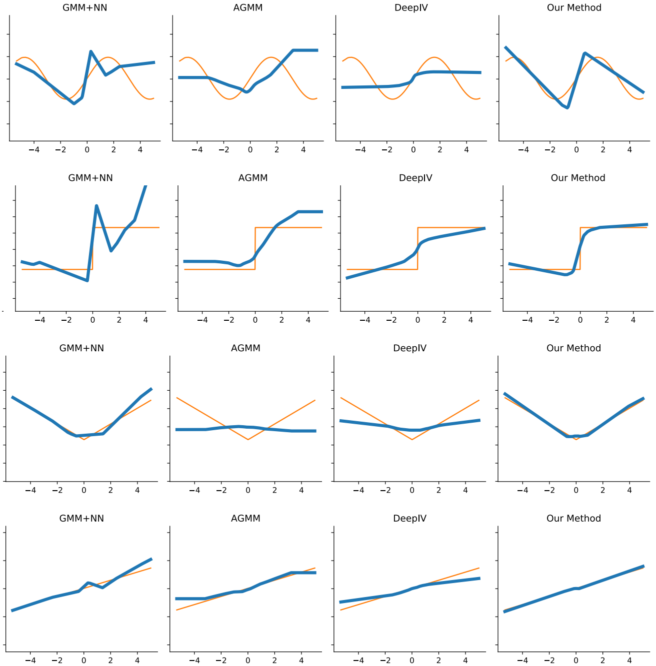
Low-dimensional scenarios

$$Y = g_0(X) + e + \delta \qquad X = 0.5\, Z_1 + 0.5\, e + \gamma$$
$$Z \sim \text{Uniform}([-3,3]^2) \qquad e \sim \mathcal{N}(0,1), \quad \gamma, \delta \sim \mathcal{N}(0, 0.1)$$

- **abs**: $g_0(x) = |x|$
- **linear**: $g_0(x) = x$
- **sin**: $g_0(x) = \sin(x)$
- **step**: $g_0(x) = \mathbb{I}_{\{x \geq 0\}}$


sin
step
abs
linear
raw data
DirectNN
Vanilla2SLS
Poly2SLS
GMM+NN
AGMM
DeepIV
Our Method

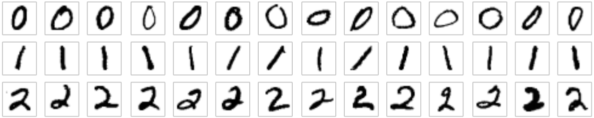

raw data
DirectNN
Vanilla2SLS
Poly2SLS


GMM+NN
AGMM
DeepIV
Our Method
GMM+NN
AGMM
DeepIV
Our Method
GMM+NN
AGMM
DeepIV
Our Method
GMM+NN
AGMM
DeepIV
Our Method
−4 −2 0 2 4

	abs	**linear**	**sin**	**step**
DirectNN	$.21 \pm .00$	$.09 \pm .00$	$.26 \pm .00$	$.21 \pm .00$
Vanilla2SLS	$.23 \pm .00$	$.00 \pm .00$	$.09 \pm .00$	$.03 \pm .00$
Poly2SLS	$.04 \pm .00$	$.00 \pm .00$	$.04 \pm .00$	$.03 \pm .00$
GMM+NN	$.14 \pm .02$	$.06 \pm .01$	$.08 \pm .00$	$.06 \pm .00$
AGMM	$.17 \pm .03$	$.03 \pm .00$	$.11 \pm .01$	$.06 \pm .01$
DeepIV	$.10 \pm .00$	$.04 \pm .00$	$.06 \pm .00$	$.03 \pm .00$
Our Method	$\mathbf{.03} \pm \mathbf{.01}$	$\mathbf{.01} \pm \mathbf{.00}$	$\mathbf{.02} \pm \mathbf{.00}$	$\mathbf{.01} \pm \mathbf{.00}$

High-dimensional scenarios

- Use MNIST images: $28 \times 28 = 784$



- Let RandImg(d) return random image of digit d
- Let $\pi(x) = \mathrm{round}(\min(\max(1.5x + 5, 0), 9))$
- Scenarios:
  - $\mathbf{MNIST_Z}$: X as before, $Z \leftarrow$ RandImg$(\pi(Z_1))$.
  - $\mathbf{MNIST_X}$: $X \leftarrow$ RandImg$(\pi(X))$, Z as before.
  - $\mathbf{MNIST_{X,\,Z}}$: $X \leftarrow$ RandImg$(\pi(X))$, $Z \leftarrow$ RandImg$(\pi(Z_1))$.

	MNIST_z	MNIST_x	$\text{MNIST}_{x,z}$
DirectNN	$.25 \pm .02$	$.28 \pm .03$	$.24 \pm .01$
Vanilla2SLS	$.23 \pm .00$	> 1000	> 1000
Ridge2SLS	$.23 \pm .00$	$.19 \pm .00$	$.39 \pm .00$
GMM+NN	$.27 \pm .01$	$.19 \pm .00$	$.25 \pm .01$
AGMM	–	–	–
DeepIV	$.11 \pm .00$	–	–
Our Method	$\mathbf{.07 \pm .02}$	$\mathbf{.15 \pm .02}$	$\mathbf{.14 \pm .02}$

DeepGMM

- We develop a method termed DeepGMM
  - Aims to addresses IV with such high-dimensional variables / complex relationships
  - Based on a new variational interpretation of optimally-weighted GMM (inverse-covariance), which we use to efficiently control very many moment conditions
  - DeepGMM given by the solution to a smooth zero-sum game, which we solve with iterative smooth-game-playing algorithms (à la GANs)
  - Numerical results will show that DeepGMM matches the performance of best-tuned methods in standard settings and continues to work in high-dimensional settings where even recent methods break