Multi-layer ANN Efficient Estimation in Nonparametric Instrumental Variables (NPIV): A Case Study

Jiafeng Chen
Harvard

Xiaohong Chen
Yale

Elie Tamer
Harvard

Big Data Conference 2022, CMSA, Harvard, August 26, 2022
https://arxiv.org/abs/2110.06763

This Paper: ANN Efficient Estimation of Average Derivative in NPIVs

- Nonparametric Instrumental Variables model: $\mathbb{E}[Y_1 - h_0(Y_2) \mid X] = 0$
- Para. of interest: $\beta_0 \equiv \mathbb{E}[a(Y_2)\nabla_1 h_0(Y_2)]$
  - (weighted $a(\cdot) > 0$) average partial derivative of $h(Y_2)$ wrt its first argument of Y_2.
  - a causal/policy parameter with continuous endogenous treatment.
- Difficulty: unknown NPIV function h_0 depends on moderately high-dim and endogenous Y_2 without known sparsity.
- iid sample: $\{Z_i = (Y_{1i}, Y_{2i}, X_i)\}_{i=1}^n$ for typical economic survey data size n
- Aim: ANN sieve efficient estimation and inference for β_0

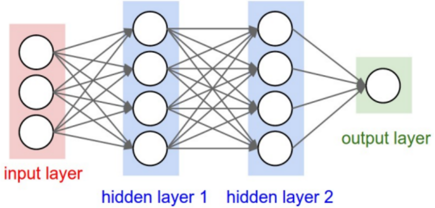
Why ANNs ?

- ANNs are compositions of simple functions $\sigma_j(W_j x + b_j)$, $j = 1, \ldots, L$, activation $\sigma_j(\cdot)$ is nonlinear, e.g., ReLU $\max(t, 0)$.

- Hornik, Stinchcombe and White (1989): universal denseness property of multi-layer ANN. DNN = ANN with hidden layer $L > 1$

- Deep learning $L >> 1$ is very successful in image processing, natural language processing,... many areas with huge + high quality data.

- Enthusiasm for using ReLU-ANN with $L \geq 2$ for average treatment effects: e.g., Farrell, Liang and Misra (2021), Athey, Imbens, Metzger and Munro (2021), ...



Questions that motivate our paper

For NPIV models $\mathbb{E}[Y_1 - h_0(Y_2)|X] = 0$ with high-dim continuous endogenous/exogenous regressors Y_2,

- will deep-layer/overparameterized/overfitted ANNs be advantageous?
- will multi-layer ReLU-ANNs perform better than other ANNs?
- will ANNs (nonlinear sieves) perform better than splines (linear sieves)?
- using ANNs, which procedure might perform better in finite samples: efficient score equation vs optimal criterion ?

Rest of the Talk

- Recall semiparametric efficiency bound for $\beta_0 = \mathbb{E}[a(Y_2)\nabla_1 h_0(Y_2)]$
- Two types of efficient estimation for β_0:
  - efficient score/influence estimation
  - optimal minimum distance (MD) estimation.
- ANN approximation error rates for function h_0.
- Monte Carlo comparisons of many inefficient and efficient estimators for β_0:
  - various ANN sieve MD estimators
  - ANN sieve MD vs spline sieve MD vs AGMM estimators
  - ANN sieve MD vs sieve score vs cross-fit sieve score estimators
  - various ways to compute standard errors.
- Empirical illustrations: averaged price elasticity in endogenous demand curves
- Conclusion and extension.

Rest of the Talk

- Recall semiparametric efficiency bound for $\beta_0 = \mathbb{E}[a(Y_2)\nabla_1 h_0(Y_2)]$
- Two types of efficient estimation for β_0:
  - efficient score/influence estimation
  - optimal minimum distance (MD) estimation.
- ANN approximation error rates for function h_0.
- Monte Carlo comparisons of many inefficient and efficient estimators for β_0:
  - various ANN sieve MD estimators
  - ANN sieve MD vs spline sieve MD vs AGMM estimators
  - ANN sieve MD vs sieve score vs cross-fit sieve score estimators
  - various ways to compute standard errors.
- Empirical illustrations: averaged price elasticity in endogenous demand curves
- Conclusion and extension.

Recall semiparametric efficiency bound for β_0

- Efficiency bound for β_0 in sequential moments Ai and Chen (2012):

$$\mathbb{E}[Y_1 - h_0(Y_2) \mid X] = 0, \quad \beta_0 = \mathbb{E}[a(Y_2)\nabla_1 h_0(Y_2)].$$

- The semiparametric efficient influence function (IF) for β_0 is

$$\psi_e(Z, \beta_0) = \underbrace{a(Y_2)\nabla_1 h_0(Y_2) - \beta_0 - \Gamma(X)[Y_1 - h_0(Y_2)]}_{\text{orthogonalized residuals}} + \underbrace{\frac{\mathbb{E}[v_e^\star(Y_2)|X]}{\Sigma(X)}}_{\alpha_e(X),\ \text{Riesz}}(Y_1 - h_0(Y_2)), \tag{1}$$

where

$$\Gamma(X) \equiv \frac{\mathrm{Cov}(a(Y_2)\nabla_1 h_0(Y_2) - \beta_0, Y_1 - h_0(Y_2) \mid X)}{\Sigma(X)}, \quad \Sigma(X) \equiv \mathrm{Var}(Y_1 - h_0(Y_2) \mid X)$$

and $\mathbb{E}[v_e^\star(Y_2)|X]$ is one solution to an optimization problem. Definition of $v_e^\star$

- The efficient variance for β_0 is: $\mathrm{Var}(\psi_e(Z, \beta_0))$.

Rest of the Talk

- Recall semiparametric efficiency bound for $\beta_0 = \mathbb{E}[a(Y_2)\nabla_1 h_0(Y_2)]$
- Two types of efficient estimation for β_0:
  - efficient score/influence estimation
  - optimal minimum distance (MD) estimation.
- ANN approximation error rates for function h_0.
- Monte Carlo comparisons of many inefficient and efficient estimators for β_0:
  - various ANN sieve MD estimators
  - ANN sieve MD vs spline sieve MD vs AGMM estimators
  - ANN sieve MD vs sieve score vs cross-fit sieve score estimators
  - various ways to compute standard errors.
- Empirical illustrations: averaged price elasticity in endogenous demand curves
- Conclusion and extension.

Efficient Score/IF Based Estimation for $\beta_0 = \mathbb{E}[a(Y_2)\nabla_1 h_0(Y_2)]$

- Efficient IF moment: $\mathbb{E}[\psi_e(Z,\beta_0)] = 0$ (Neyman orthogonal moment) [ES]

$$\sum_{i=1}^{n} \hat{\psi}_e(z_i;\hat{\beta}_{ES}) = \sum_{i=1}^{n} \left(a(y_{2i})\nabla_1 \hat{h}(y_{2i}) - \hat{\beta}_{ES} - [\hat{\Gamma}(x_i) - \hat{\alpha}_e(x_i)] \left(y_{1i} - \hat{h}(y_{2i}) \right) \right) = 0,$$

- $\hat{h}$: ANN or spline sieve MD estimator of h_0;
- $\hat{\Gamma}$: any consistent (e.g., sieve least squares) plug-in estimator of Γ;
- $\hat{\alpha}_e = \alpha_e(\hat{v}_e, \hat{\Sigma})$: any consistent plug-in estimator of $\alpha_e = \frac{\mathbb{E}[v_e^{\star}(Y_2)|X]}{\Sigma(X)}$.

Optimal SMD estimation of $\beta_0 = \mathbb{E}[a(Y_2)\nabla_1 h_0(Y_2)]$

- Orthogonalized plug-in optimal sieve MD estimation (Ai and Chen (2012)) [OP-OSMD]:

$$\hat{\beta}_{OP}(\hat{h}) = \frac{1}{n}\sum_{i=1}^{n}[a(y_{2i})\nabla_1 \hat{h}(y_{2i}) - \hat{\Gamma}(x_i)(y_{1i} - \hat{h}(y_{2i}))]$$

- $\hat{h}$: ANN or spline sieve MD estimator of h_0;
- $\hat{\Gamma}$: any consistent (e.g., sieve least squares) plug-in estimator of Γ;
- Asymptotically linear, normal and efficient:

$$\sqrt{n}[\hat{\beta}_{OP}(\hat{h}_{OSMD}) - \beta_0] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_e(z_i, \beta_0) + o_p(1) \rightsquigarrow \mathcal{N}\left(0, \mathbb{E}[(\psi_e(Z, \beta_0))^2]\right),$$

Inefficient Estimators for β_0

- Simple plug-in SMD estimation (Ai and Chen (2007)) [P-ISMD]:

$$\hat{\beta}_P(\hat{h}) = \frac{1}{n}\sum_{i=1}^{n} a(y_{2i})\nabla_1 \hat{h}(y_{2i}).$$

$$\sqrt{n}[\hat{\beta}_P(\hat{h}) - \beta_0] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi_{ie}(z_i, \beta_0) + o_p(1) \rightsquigarrow \mathcal{N}\left(0, \mathbb{E}[(\psi_{ie}(Z, \beta_0))^2]\right),$$

$$\psi_{ie}(Z;\beta_0) = a(Y_2)\nabla_1 h_0(Y_2) - \beta_0 + \underbrace{\mathbb{E}[v_{ie}^{\star}(Y_2)|X]}_{\alpha_{ie}(X),\ \text{Riesz}}(Y_1 - h_0(Y_2)).$$

- Inefficient score/IF based estimation: $\mathbb{E}[\psi_{ie}(Z, \beta_0)] = 0$. [IS]

$$\sum_{i=1}^{n} \hat{\psi}_{ie}(z_i;\hat{\beta}_{IS}) = \sum_{i=1}^{n}\left(a(y_{2i})\nabla_1 \hat{h}(y_{2i}) - \hat{\beta}_{IS} + \hat{\alpha}_{ie}(x_i)\left(y_{1i} - \hat{h}(y_{2i})\right)\right) = 0,$$

  - $\hat{h}$: ANN or spline sieve MD estimator of h_0;
  - $\hat{\alpha}_{ie}$: any consistent plug-in estimator of $\alpha_{ie}(X) = \mathbb{E}[v_{ie}^{\star}(Y_2) \mid X]$.

Comparisons of Estimators for $\beta_0 = \mathbb{E}[a(Y_2)\nabla_1 h_0(Y_2)]$

- 3 (first-order) asymptotically equivalent inefficient estimators.
  - $\hat{\beta}_P = \frac{1}{n}\sum_{i=1}^n a(y_{2i})\nabla_1 \hat{h}(y_{2i})$
  - $\hat{\beta}_{IS} = \frac{1}{n}\sum_{i=1}^n \left[a(y_{2i})\nabla_1 \hat{h}(y_{2i}) + \hat{\alpha}_{ie}(x_i)(y_{1i} - \hat{h}(y_{2i}))\right]$
  - $\hat{\beta}_{IS-X}$: split-sample or cross-fit version of $\hat{\beta}_{IS}$ (inspired by Chernozhukov *et al.* (2018, 2021)).

- 3 (first-order) asymptotically equivalent efficient estimators.
  - $\hat{\beta}_{OP} = \frac{1}{n}\sum_{i=1}^n \left[a(y_{2i})\nabla_1 \hat{h}(y_{2i}) - \hat{\Gamma}(x_i)(y_{1i} - \hat{h}(y_{2i}))\right]$
  - $\hat{\beta}_{ES} = \frac{1}{n}\sum_{i=1}^n \left[a(y_{2i})\nabla_1 \hat{h}(y_{2i}) - [\hat{\Gamma}(x_i) - \hat{\alpha}_e(x_i)](y_{1i} - \hat{h}(y_{2i}))\right]$
  - $\hat{\beta}_{ES-X}$: split-sample or cross-fit version of $\hat{\beta}_{ES}$.

- $\hat{h}$ is ANN SMD or spline SMD for h_0 solving $\mathbb{E}[Y_1 - h_0(Y_2)|X] = 0$.
- $\hat{\Gamma}$, $\hat{\alpha}_e$, $\hat{\alpha}_{ie}$ are plug-in estimators that all depend on $\hat{h}$.

Sieve Minimum Distance (SMD) Estimation of h_0

- NPIV model: $\mathbb{E}[Y_1 - h(Y_2) \mid X] = 0$ iff $h = h_0 \in \mathcal{H}$
- In sample, let $y_1, y_2, \phi(x)$ be vector/matrices of n observations, with $\phi(x)$ an $n \times d_\phi$ matrix of linear sieve basis for $L^2(X)$.
  - For any given h, the residuals are $y_1 - h(y_2)$
  - Cond. expectation operator $\mathbb{E}[\cdot \mid X]$ can be approximated using IV sieve $\phi(x)$:

$$P_\phi = \phi(x)(\phi(x)'\phi(x))^- \phi(x)',$$

- The SMD estimator (Ai and Chen (2003)):

$$\hat{h}_{SMD} = \arg\min_{h \in \mathcal{H}_n} \| \underbrace{P_\phi[y_1 - h(y_2)]}_{\text{residuals projected onto IV sieve}} \|^2_{\hat{W}} \quad \text{for a weighting matrix } \hat{W},$$

  $\mathcal{H}_n$ is a sieve for $\mathcal{H}$, can be nonlinear (e.g., ANN) or linear (e.g., spline).
- $\hat{h}_{ISMD}$ for $\hat{W} = I$; and $\hat{h}_{OSMD}$ for $\hat{W}$ a consistent estimate of $[\Sigma(X)]^{-1}$.

Rest of the Talk

- Recall semiparametric efficiency bound for $\beta_0 = \mathbb{E}[a(Y_2)\nabla_1 h_0(Y_2)]$
- Two types of efficient estimation for β_0:
  - efficient score/influence estimation
  - optimal minimum distance (MD) estimation.
- ANN approximation error rates for function h_0.
- Monte Carlo comparisons of many inefficient and efficient estimators for β_0:
  - various ANN sieve MD estimators
  - ANN sieve MD vs spline sieve MD vs AGMM estimators
  - ANN sieve MD vs sieve score vs cross-fit sieve score estimators
  - various ways to compute standard errors.
- Empirical illustrations: averaged price elasticity in endogenous demand curves
- Conclusion and extension.

Sieve approximation error rates

- Linear sieves (polynomials, splines, orthogonal wavelets) typically have approximation rates (in $\|\cdot\|_\infty$):

$$O\left((\text{sieve terms})^{-\text{smoothness}/\text{dimension}}\right)$$

(for $h_0(Y_2) \in$ Hölder smooth class) Linear sieve details

  - Curse of dimensionality: given smoothness, approximation error rates goes worse as $\dim(Y_2)$ grows.

- For single hidden layer ANNs, Makovoz (1996); Chen and White (1999) show that the approximation rates (in L^2 norm) are

$$o\left((\text{Number of neurons})^{-1/2}\right)$$

(for $h_0(Y_2) \in$ Barron (1993) class), Nonlinear sieve details

Examples of nonlinear sieves: Multi-Layer ANNs

- Feedforward ANNs are compositions of simple functions

$$f_{\sigma_j, W_j, b_j}(x) = \sigma_j(W_j x + b_j), \quad j = 1, ..., L,$$

  activation $\sigma_j(\cdot)$ is applied component-wise; known and nonlinear; e.g.,
- Sigmoid activation $\sigma_j(t) = \frac{1}{1+e^{-t}}$. ReLU activation $\sigma_j(t) = \max(t, 0)$.
- ANN sieves: ANN $(f_1, \ldots, f_L) =$

$$\left\{ W_{L+1} f_{\sigma_L, W_L, b_L} \circ \cdots \circ f_{\sigma_1, W_1, b_1} + b_{L+1} : W_1, \ldots, W_{L+1}, b_1, \ldots, b_{L+1} \text{ conformable} \right\}$$

- Complexity/flexibility of ANN $(f_1, \ldots, f_L)$ is intuitively in terms of
  - L hidden layers, max dimension of width W_l, or growth of norm $\|(W, b)\|$.
- For multi-layer ReLU ANNs, Yarotsky (2017); Schmidt-Hieber (2019); Shen *et al.* (2021b) approximation error rates under different settings
- For other activation ANNs, all kinds of approximation error rates, see e.g., Shen *et al.* (2021a)

Rest of the Talk

- Recall semiparametric efficiency bound for $\beta_0 = \mathbb{E}[a(Y_2)\nabla_1 h_0(Y_2)]$
- Two types of efficient estimation for β_0:
    - efficient score/influence estimation
    - optimal minimum distance (MD) estimation.
- ANN approximation error rates for function h_0.
- Monte Carlo comparisons of many inefficient and efficient estimators for β_0:
    - various ANN sieve MD estimators
    - ANN sieve MD vs spline sieve MD vs AGMM estimators
    - ANN sieve MD vs sieve score vs cross-fit sieve score estimators
    - various ways to compute standard errors.
- Empirical illustrations: averaged price elasticity in endogenous demand curves
- Conclusion and extension.

Monte Carlo Design 2

$Y_2 = [R_1, R_2, X_2, \tilde{X}]$, $X = [X_1, X_2, X_3, \tilde{X}]$, $\beta_0 = \mathbb{E}[\nabla_1 h_0(Y_2)] = 1$.

$$\boxed{Y_1 = h_0(Y_2) + U = R_1 + h_{01}(R_2) + h_{02}(X_2) + h_{03}(\tilde{X}) + U}$$

X is marginally uniform

$$h_{01} : \mathbb{R} \to \mathbb{R} \quad t \mapsto \frac{1}{1 + \exp(-t)}$$
$$h_{02} : \mathbb{R} \to \mathbb{R} \quad t \mapsto \log(1 + t)$$
$$h_{03} : \mathbb{R}^{d_{\tilde{x}}} \to \mathbb{R} \quad \tilde{x} \mapsto 5\tilde{x}_1^3 + \tilde{x}_2 \cdot \max_j (\tilde{x}_j \vee 0.5) + 0.5 \exp(-\tilde{x}_{d_{\tilde{x}}})$$

- $R_1 = X_1 + U + 0.5U_2$, $R_2 = \Phi\left[\Phi^{-1}(X_3) + 0.5U_3\right]$.
- $U = (U_1 + U_2 + U_3)/3 \times \sqrt{(X_1^2 + X_2^2 + X_3^2)/3}$. U_1, U_2, U_3 iid $\sim \mathcal{N}(0, 1)$.
- Two settings we tweak:
  - Dimension of $\tilde{X}$: $\{0, 5, 10\}$. Y_2 contains up to 13 continuous covariates.
  - Correlation between $\tilde{X}$ and $[X_1, X_2, X_3]$: {Yes, No}
- Sample sizes $n = 1000, 5000$. (1000 Monte Carlo replications, 1000 bootstrap runs)

Plug-in ANN SMD estimators (MC2)

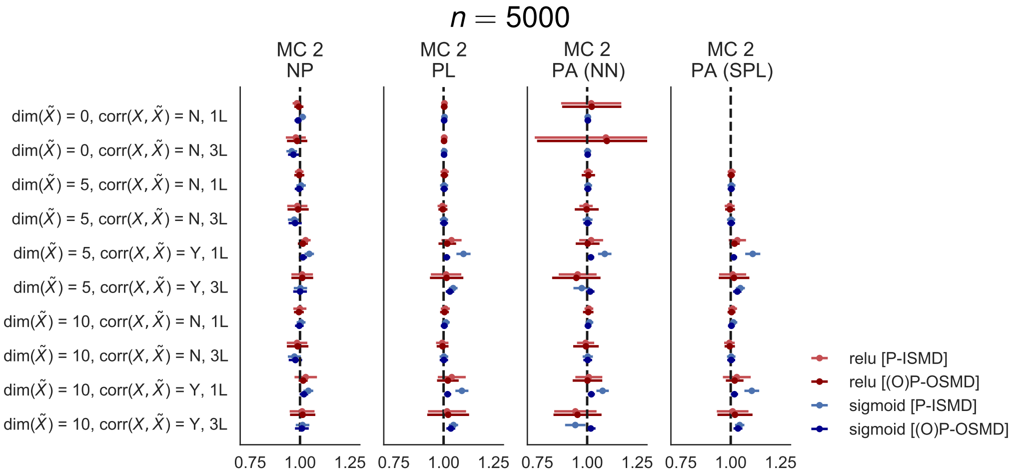


Figure: Monte Carlo Mean ± 1 Monte Carlo Stdev

NP Nonparametric; **PL** Partial Linear; **PA(NN)** Partially Additive Neural Net; **PA(SPL)** Partially Additive Spline.

Summary of MC findings for plug-in ANN SMDs so far

- Choices of ANN activation functions (ReLU vs Sigmoid) and number of layers do not matter much. Consistent with ANN approximation theory.
- ANN SMDs can perform well after several hyper-parameters tuning.
- ANN OP-OSMD ANN has smaller bias than ANN P-ISMD.
- ANN SMDs are sometimes numerically unstable (optimization doesn't converge).
- ANN SMDs are not too sensitive to choice of IV sieve, but seem less biased using larger IV sieves in complex DGPs (MC2 with correlations among $[R_1, R_2]$, $[X_1, X_2, X_3]$ and $\tilde{X}$).
- Multi-layer ANNs seem not fully "adaptive" to underlying true partially linear additive structure of the DGPs.
- Bad idea to apply ANN to estimate functions of scalar variable $h_{01}(R_2)$ and $h_{02}(X_2)$. Better to use linear sieves (such as splines) for functions of low-dim covariates.

Comparing Plug-in ANN SMDs to Other Estimators of β_0

- Simple plug-in $\hat{\beta}_P = \frac{1}{n}\sum_{i=1}^n \nabla_1 \hat{h}(y_{2i})$ [P-ISMD]
- Orthogonal plug-in $\hat{\beta}_{OP} = \frac{1}{n}\sum_{i=1}^n \left[\nabla_1 \hat{h}(y_{2i}) - \hat{\Gamma}(x_i)(y_{1i} - \hat{h}(y_{2i}))\right]$ [OP-OSMD]
- Identity-weighted score $\hat{\beta}_{IS} = \frac{1}{n}\sum_{i=1}^n \left[\nabla_1 \hat{h}(y_{2i}) + \hat{\alpha}_{ie}(x_i)(y_{1i} - \hat{h}(y_{2i}))\right]$ [IS]
- Efficient score $\hat{\beta}_{ES} = \frac{1}{n}\sum_{i=1}^n \left[\nabla_1 \hat{h}(y_{2i}) - [\hat{\Gamma}(x_i) - \hat{\alpha}_e(x_i)](y_{1i} - \hat{h}(y_{2i}))\right]$ [ES]
- Split-sample score-based estimators [IS-X, ES-X]

- $\hat{h}$ is ANN SMD or spline SMD for h_0.
- $\hat{\Gamma}$, $\hat{\alpha}_e$, $\hat{\alpha}_{ie}$ are plug-in estimators that all depend on $\hat{h}$.
- Simple plug-in using adversarial GMM $\hat{h}$ of Dikkala *et al.* (2020) [AGMM]

A horse-race among efficient estimators of β_0

MC2, Optimally-weighted estimators, Nonparametric, $n = 5000$

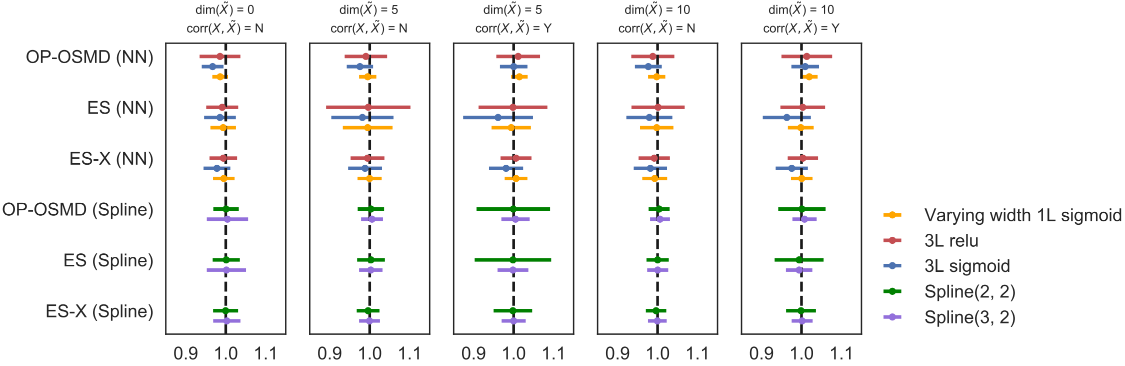


Figure: Monte Carlo Mean ± 1 Monte Carlo Stdev

OP-OSMD Optimal SMD; **ES** Efficient Score; **ES-X** Split Sample Efficient Score

Sensitivity of ES/ES-X to estimation of $\Sigma(X)^{-1}$ in $\alpha_e(X) = \mathbb{E}[v_e^*(Y_2)|X]\Sigma(X)^{-1}$

MC2, ES/ES-X estimators of β_0, Nonparametric, $n = 5000$

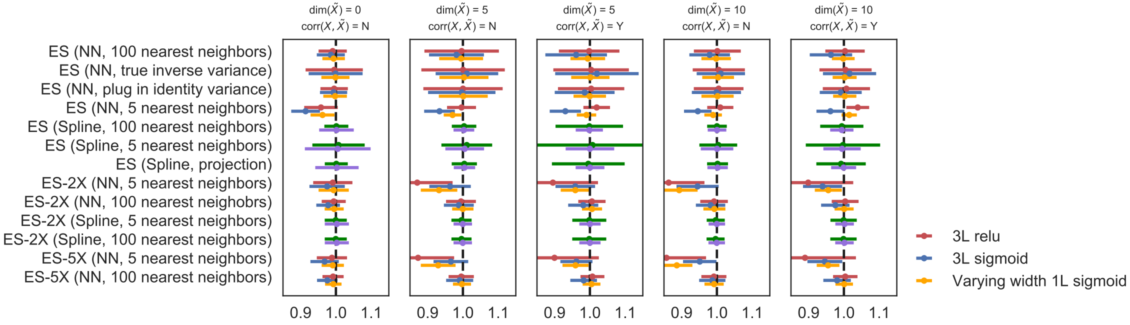


Figure: Monte Carlo Mean ± 1 SE estimates

NB: the OP-SMD estimation of h and the sieve projection estimation of Γ both involve an estimator of Σ, which are held fixed. We only vary the estimation of Σ^{-1} in α_e.

A horse-race among estimators of β_0

- Spline SMD estimators and ANN OP-OSMD estimators work very well
- ANN score estimators seem less biased than ANN SMD estimators, and also works well with the right tuning parameters
- (Two-fold) cross-fitting estimators have comparable performance in large samples and slightly poorer performance in smaller samples
- ANN ES estimators are sensitive to estimation of certain nuisance parameters in the score (Σ^{-1})
- The sensitivity is not significantly mitigated by two or five-fold cross-fitting here

Inference

- Estimation of standard errors amounts to estimating the variance of the influence function
- For SMD estimators, can also consider a multiplier bootstrap that weighs the residuals with random weights (e.g. $\omega_i \overset{\text{i.i.d.}}{\sim} \text{Expo}(1)$):

$$\tilde{U}_i = \omega_i(Y_{1i} - h(Y_{2i})), \omega_i \overset{\text{i.i.d.}}{\sim} [1, 1], \omega_i \geq 0$$

and use $\|P_\phi \tilde{u}\|$ as the objective function in SMD estimation

Inference

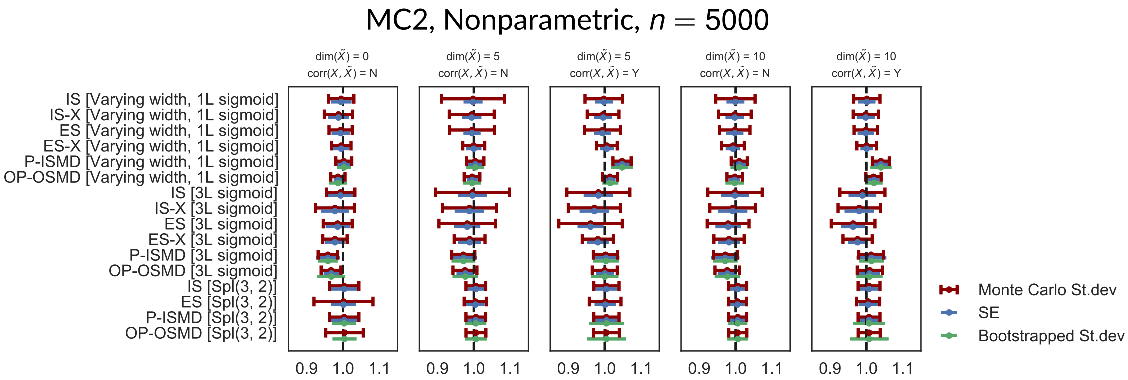


Figure: Monte Carlo Mean ± 1 SE estimates

NB: Bootstrap SE based on one realization of the data

Inference

MC2, Nonparametric, $n = 5000$

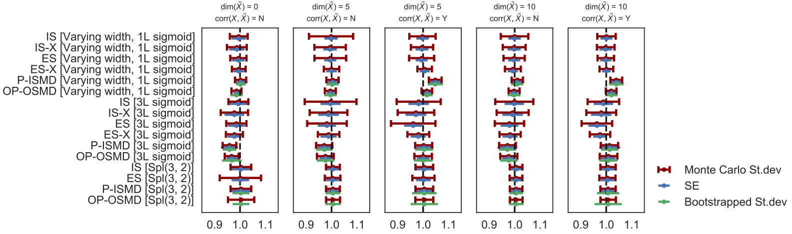


Figure: Monte Carlo Mean ± 1 SE estimates

NB: Bootstrap SE based on one realization of the data

- Estimated SEs are mostly accurate for spline-SMDs and ANN-SMDs, but less so for IS and ES
- (Criterion) bootstrap inference for SMD estimators has reasonable coverage (not shown here)

MC 2, but R_1 enters through $R_1^2/2 + R_1 f_2(X_2)$ (more sensitivity check of ANN SMD)

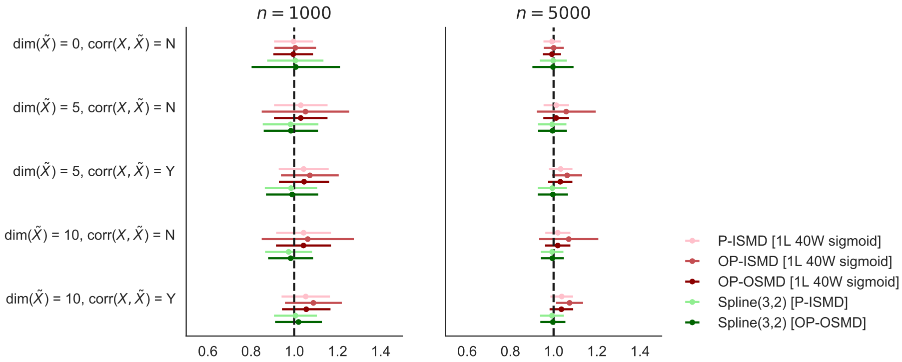


Figure: Monte Carlo Mean ± 1 Monte Carlo Stdev

Estimation of the partial derivative

In this model, the partial derivative $\nabla_1 h_0$ is of the form $f_1(R_2) + f_2(X_2)$, and we evaluate performance estimating f_1, f_2

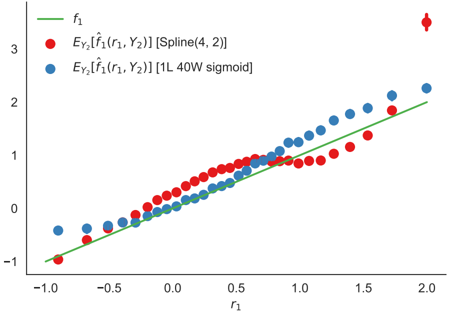


Figure: Estimated f_1 versus true f_1. Single sample for $N = 10,000$

Estimated f_1 is calculated by taking $\nabla_1 \hat{h} - f_2(x_2)$. We plot expectation marginalizing over variables other than r_1.

Estimation of f_1, f_2

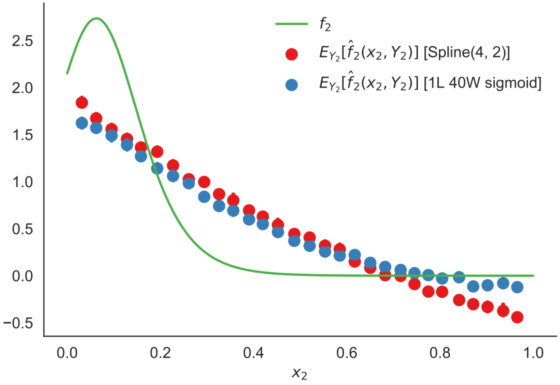
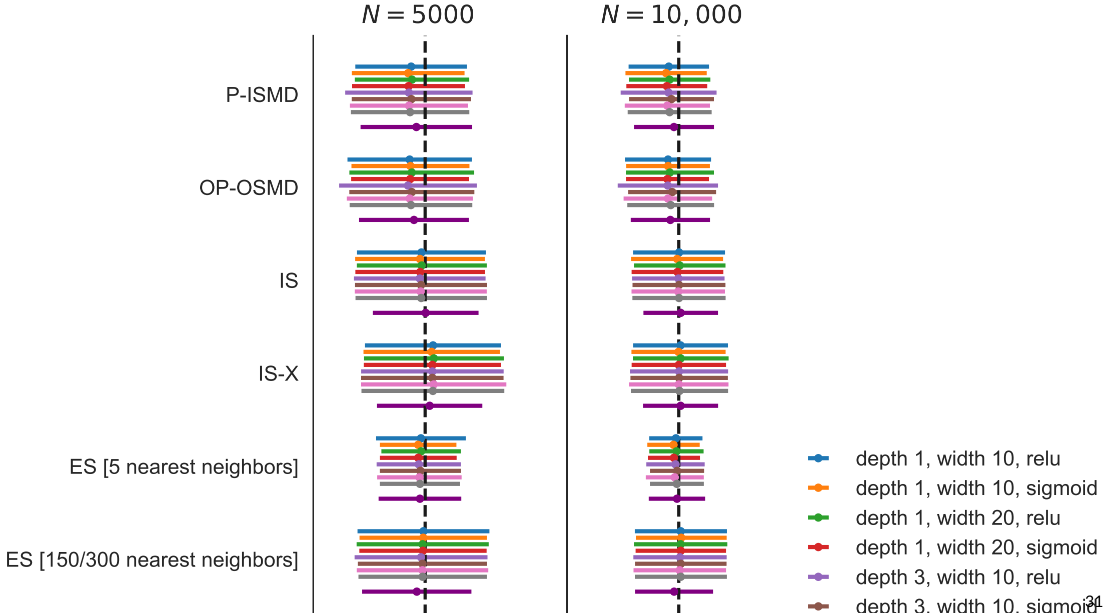


Figure: Estimated f_2 versus true f_2. Single sample for $N = 10,000$

Estimated f_2 is calculated by taking $\nabla_1 \hat{h} - f_1(r_1)$. We plot expectation marginalizing over variables other than x_2.

Calibration to Gasoline Demand



Main Takeaways from our Monte Carlo Studies

In our experience,

- ANNs are useful for approximating unknown functions of high dimensional endogenous or exogenous variables.
- Choices of ANN activation (ReLU vs Sigmoid), layers and widths do not matter much when approximating smooth functions of moderately high dimension (13).
- ANN OP-OSMD and ANN “IS” have smaller biases than ANN P-ISMD.
- Compared to ANN OP-OSMD, ANN ES/ES-X is more sensitive to estimating weighting matrix and can be more biased.
- Stable inferences are currently more difficult to achieve for ANN based estimators in NPIV models.
- Spline based estimators (P-SMD, OP-SMD, IS/IS-X, ES/ES-X) for β are less biased, more stable and accurate, even in NPIV models with high-dimensional (13) continuous covariates.
- Gap between theory and current practice.

Rest of the Talk

- Recall semiparametric efficiency bound for $\beta_0 = \mathbb{E}[a(Y_2)\nabla_1 h_0(Y_2)]$
- Two types of efficient estimation for β_0:
    - efficient score/influence estimation
    - optimal minimum distance (MD) estimation.
- ANN approximation error rates for function h_0.
- Monte Carlo comparisons of many inefficient and efficient estimators for β_0:
    - various ANN sieve MD estimators
    - ANN sieve MD vs spline sieve MD vs AGMM estimators
    - ANN sieve MD vs sieve score vs cross-fit sieve score estimators
    - various ways to compute standard errors.
- Empirical illustrations: averaged price elasticity in endogenous demand curves
- Conclusion and extension.

Average Elasticity of Nonparametric Gasoline Demand

	P-ISMD	OP-OSMD	IS
Sigmoid [1L]	-1.28	-1.24	-1.12
	[-1.69, -0.9]	[-1.64, -0.87]	(0.22)
Sigmoid [3L]	-1.24	-1.28	-1.11
	[-1.65, -0.9]	[-1.64, -0.87]	(0.22)
ReLU [3L]	-1.27	-1.25	-1.14
	[-1.65, -0.9]	[-1.64, -0.87]	(0.22)
Spline(3, 2)	-1.17	-1.2	
	[-1.57, -0.8]	[-1.6,-0.8]	
	Blundell *et al.* (2012) OLS	OLS	TSLS
	-0.83	-0.85	-1.24
	(0.148)	(0.15)	(0.2)

- Data: National Household Travel Survey (Blundell, Horowitz and Parey, 2012)
- 7 Covariates: log gasoline price, log income, household size, driver, household age, number working, public transit distance
- Instrumenting gasoline price with distance to Gulf of Mexico

Average Price Derivative of Nonparamatric Multi-Product Demand

Non-organic

	IS	P-ISMD	OP-OSMD
Sigm [1L]	-1.649 (0.04)	-1.530 (0.04)	-1.747 (0.03)
Relu [1L]	-1.648 (0.04)	-1.590 (0.04)	-1.706 (0.04)
Relu [3L]	-1.648 (0.04)	-1.634 (0.04)	-1.659 (0.06)
Spline(3,2)	-1.611 (0.04)	-1.648 (0.04)	-1.676 (0.04)

Organic

	IS	P-ISMD	OP-OSMD
Sigm [1L]	-3.235 (0.07)	-2.409 (0.09)	-3.382 (0.06)
Relu [1L]	-3.236 (0.07)	-2.197 (0.06)	-2.129 (0.08)
Relu [3L]	-3.232 (0.07)	-2.206 (0.07)	-2.122 (0.14)
Spline(3,2)	-3.194 (0.06)	-3.232 (0.07)	-3.124 (0.06)

- Data: Nielsen strawberry demand data (Compiani, 2019)*
- 6 Covariates: Strawberry Prices (non-organic, organic), Income, Lettuce demand (Taste for organic proxy), State-level sale of non-strawberry fresh fruits, Average outside good price
- 5 excluded Instruments: 3 Hausman IV (Prices in neighbouring markets), 2 Strawberry spot prices (marginal cost measures)

*These results do not necessarily represent the views of the Nielsen Company

Concluding Remarks

- ANNs are useful for approximating unknown functions of high dimensional endogenous or exogenous variables.
- ANN OP-OSMD and ANN IS/IS-X have smaller biases than ANN P-ISMD.
- In our experience so far, stable and accurate inferences are currently more challenging to achieve for ANN based estimators in NPIV models.
- Be aware of many free tuning parameters.

Extensions in Theory: Chen, Liao and Wang (2021)

- Multi-layer ANN optimally weighted quasi likelihood ratio inference for possibly slower than root-n functionals in general conditional moment restrictions time series setting:

$$\mathbb{E}[\rho_1(Z_t;\beta_{01},\beta_{02},h_0(Y_{2,t}))] = 0\ ,\ \mathbb{E}[\rho_2(Z_t;\beta_{02},h_0(\cdot))|X_t] = 0.$$

Leading Examples:

- weighted average derivative of nonparametric quantile instrumental variables model, (endogenous default, conditional value-at-risk, etc))

$$\beta_{01} = \mathbb{E}[a(Y_{2,t})\nabla_1 h_0(Y_{2,t})],\ \ \mathbb{E}[1(Y_{1,t} \leq h_0(Y_{2,t})) - \tau|X_t] = 0.$$

- Off-policy evaluation in reinforcement learning, Bellman equation.
- more difficult to implement accurate inference in practice.

THANK YOU for ATTENDING the TALK!

COMMENTS ARE WELCOME!

Recall semiparametric efficiency bound for β_0

Back

- Let $\sigma_0^2 \equiv \mathrm{Var}(a(Y_2)\nabla_1 h_0(Y_2) - \beta_0 - \Gamma(X)[Y_1 - h_0(Y_2)])$, with

$$\Gamma(X) \equiv \frac{\mathrm{Cov}(a(Y_2)\nabla_1 h_0(Y_2) - \beta_0, Y_1 - h_0(Y_2) \mid X)}{\Sigma(X)}, \quad \Sigma(X) \equiv \mathrm{Var}(Y_1 - h_0(Y_2) \mid X)$$

$$J_0 \equiv \inf_{r \in \overline{\mathcal{W}}} \mathbb{E}\left\{ \frac{(1 + \mathbb{E}[a(Y_2)\nabla_1 r(Y_2) + \Gamma(X) r(Y_2)])^2}{\sigma_0^2} + \frac{(\mathbb{E}[r(Y_2) \mid X])^2}{\Sigma(X)} \right\} \tag{2}$$

$$\overline{\mathcal{W}} = \{r : \mathbb{E}[\Sigma(X)^{-1}(\mathbb{E}\{r(Y_2)|X\})^2] + (\mathbb{E}\{a(Y_2)\nabla_1 r(Y_2) + \Gamma(X) r(Y_2)\})^2 < \infty\}.$$

- $\mathbb{E}[v_e^\star(Y_2)|X] = (J_0)^{-1}\mathbb{E}[r_0(Y_2)|X]$, where r_0 is one solution (not necessarily unique) to the optimization (2).
- Under completeness condition,

$$v_e^\star(Y_2) = (J_0)^{-1} r_0(Y_2) = \frac{r_0(Y_2)\sigma_0^2}{\mathbb{E}[1 + a(Y_2)\nabla_1 r_0(Y_2) + \Gamma(X) r_0(Y_2)]}$$

Examples of Linear Sieves (Series)

Back

- Let $\rho_{\infty n} \equiv \inf_{g \in \mathcal{H}_n} ||g - h_0||_\infty$ be the sieve approximation errors to $h_0 \in \mathcal{H} = \Lambda^p([0,1]^d)$ (Hölder class) in $L_\infty([0,1]^d, leb)$-norm.
- Let $\mathcal{H}_n$ be a tensor product linear sieve for $\mathcal{H}$, with $\dim(\mathcal{H}_n) = k_n$.
- The linear sieve approximation error rates for $h_0 \in \mathcal{H} = \Lambda^p([0,1]^d)$ are:
  Polynomials. $\rho_{\infty n} = O(k_n^{-p/d})$. (see Timan 63)
  Trigonometric polynomials. $\rho_{\infty n} = O(k_n^{-p/d})$. (see Timan 63)
  $r-$**th order Splines** (with $r > p$). $\rho_{\infty n} = O(k_n^{-p/d})$ (see Schumaker 81).
  $m-$**th order Orthogonal wavelets** (with $m > p$). $\rho_{\infty n} = O(k_n^{-p/d})$ (see Meyer, 92).

- "Curse of Dimensionality": for fixed smoothness p, the approximation error rate $\rho_{\infty n} = O(1)$ as $d = \dim(X)$ goes to infinity

Examples of Nonlinear Sieves: Single-hidden Layer ANN

Back

Barron class: $\mathcal{H} = \{h \in L_2(\mathcal{X}, leb) : \int_{\mathcal{R}^d} |w||\widetilde{h}(w)|dw < \infty\}$, $\widetilde{h}(w) \equiv \int \exp(-iwx)h(x)dx$ is the Fourier transform of h.

Sigmoid ANN. $\mathrm{sANN}(k_n) = \left\{\sum_{j=1}^{k_n} \alpha_j S(\gamma_j' x + \gamma_{0,j}) : \gamma_j \in \mathcal{R}^d, \alpha_j, \gamma_{0,j} \in \mathcal{R}\right\}$, where

$S : \mathcal{R} \to \mathcal{R}$ is a sigmoid activation function, i.e., a bounded non-decreasing function such that $\lim_{u \to -\infty} S(u) = 0$ and $\lim_{u \to \infty} S(u) = 1$. Examples of $S()$:

- heaviside $S(u) = 1\{u \geq 0\}$;
- logistic $S(u) = 1/(1 + \exp\{-u\})$;
- Gaussian sigmoid $S(u) = (2\pi)^{-1/2} \int_{-\infty}^{u} \exp(-y^2/2)dy$;

Barron (1993): $\mathrm{sANN}(k_n)$ sieve approximation error rate in $L_2(\mathcal{X}, leb)$-norm is no slower than $O([k_n]^{-1/2})$. Makovoz (1996) improved it to $O([k_n]^{-1/2-1/(2d)})$ for the heaviside S; Chen and White (1999) improved it to $O([k_n]^{-1/2-1/(d+1)})$ for general S. (For other nonlinear sieves see, e.g. Chen (2007))

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71** (6), 1795–1843.

— and — (2007). Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, **141** (1), 5–43.

— and — (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, **170** (2), 442–457.

Athey, S., Imbens, G. W., Metzger, J. and Munro, E. (2021). Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, **39** (3), 930–945.

Blundell, R., Horowitz, J. L. and Parey, M. (2012). Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics*, **3** (1), 29–51.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, **6**, 5549–5632.

— and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, **45** (2), 682–691.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

—, Escanciano, J. C., Ichimura, H., Newey, W. K. and Robins, J. M. (2021). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Compiani, G. (2019). Market counterfactuals and the specification of multi-product demand: A nonparametric approach. *Available at SSRN*.

Dikkala, N., Lewis, G., Mackey, L. and Syrgkanis, V. (2020). Minimax estimation of conditional moment models. *arXiv preprint arXiv:2006.07201*.

Farrell, M. H., Liang, T. and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, **89** (1), 181–213.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, **2** (5), 359–366.

Makovoz, Y. (1996). Random approximants and neural networks. *Journal of Approximation Theory*, **85** (1), 98–109.

Schmidt-Hieber, J. (2019). Deep relu network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*.

Shen, Z., Yang, H. and Zhang, S. (2021a). Neural network approximation: Three hidden layers are enough. *Neural Networks*, **141**, 160–173.

—, — and — (2021b). Optimal approximation rate of relu networks in terms of width and depth. *arXiv preprint arXiv:2103.00502*.

Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, **94**, 103–114.