Robust Inference for Convex Pairwise Difference Estimation

Matias D. Cattaneo
Princeton University

Michael Jansson
UC Berkeley

Kenichi Nagasawa
University of Warwick

May 2025

Prepared for the Econometric Theory Lecture at the 2025 International Symposium on Econometric Theory and Applications (SETA), University of Macau (China).

Outline

Introduction

Pairwise Difference Estimation is used in economics and other disciplines.

- Identification idea:
  - Localize pair of observations with similar observable characteristics to remove unobserved confounders/heterogeneity.
- Background references:
  - Honoré & Powell (2005): "Pairwise Difference Estimators for Nonlinear Models".
  - Aradillas-Lopez, Honoré & Powell (2007): "Pairwise Difference Estimation with Nonparametric Control Variables".
  - Many, many more...

Contribution: Robust Distribution Theory and Inference Methods.

- New distributional approximation for possibly "more strict" localization.
- Two asymptotic regimes depending on smoothness of objective function.
  - *Smooth*: "Small Bandwidth" Gaussian law.
  - *Non-Smooth*: Mixture of "Small Bandwidth" Gaussian and Chernoff-type laws.

Setup: Pairwise Difference Estimation

DGP: $\mathbf{z}_i = (y_i, \mathbf{x}_i, \mathbf{w}_i)^\top$, $i = 1, \ldots, n$, random sample + regularity conditions.

Parameter of Interest:

$$\boldsymbol{\theta}_0 = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} M_0(\boldsymbol{\theta}), \qquad M_0(\boldsymbol{\theta}) = \lim_{n\to\infty} \mathbb{E}\Big[m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}) \tfrac{1}{h^d} K\big(\tfrac{\mathbf{w}_1 - \mathbf{w}_2}{h}\big)\Big].$$

- Loss function $m(\cdot)$ is example-specific.
- Kernel function K re-weights for localization.
- Bandwidth $h \to 0$ (as $n \to \infty$) key tuning parameter for localization.

Estimator:

$$\widehat{\boldsymbol{\theta}}_n = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \widehat{M}_n(\boldsymbol{\theta}), \qquad \widehat{M}_n(\boldsymbol{\theta}) = \binom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) \tfrac{1}{h^d} K\big(\tfrac{\mathbf{w}_i - \mathbf{w}_j}{h}\big).$$

- Inference sensitive to choice of h (lack of robustness).
- $m(\cdot)$ can be non-convex, non-differentiable, discontinuous, etc.

Goal: robust inference with respect to choice of h and features of $m(\cdot)$.

Example 1: Partially Linear Model

Model:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\theta}_0 + g(\mathbf{w}_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i | \mathbf{x}_i, \mathbf{w}_i] = 0.$$

Identification: $\mathbf{w}_i = \mathbf{w}_j \implies \mathbb{E}[y_i - y_j | \mathbf{x}_i, \mathbf{x}_j, \mathbf{w}_i - \mathbf{w}_j = 0] = (\mathbf{x}_i - \mathbf{x}_j)^\top \boldsymbol{\theta}_0$, and

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta} \in \Theta}{\arg\min}\, \mathbb{E}\big[m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) \big| \mathbf{w}_i - \mathbf{w}_j = 0\big]$$

with

$$m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = \big[(y_i - y_j) - (\mathbf{x}_i - \mathbf{x}_j)^\top \boldsymbol{\theta}\big]^2.$$

Estimator: localize $\mathbf{w}_i \approx \mathbf{w}_j$, and thus

$$\widehat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta} \in \Theta}{\arg\min} \binom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) \tfrac{1}{h^d} K\big(\tfrac{\mathbf{w}_i - \mathbf{w}_j}{h}\big).$$

- $m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$ is convex and differentiable.
- $\widehat{\boldsymbol{\theta}}_n$ closed-form.

Example 2: Partially Linear Logistic Model

Model:

$$y_i = \mathbb{1}(\mathbf{x}_i^\top \boldsymbol{\theta}_0 + g(\mathbf{w}_i) + \varepsilon_i \geq 0), \qquad \varepsilon_i \perp\!\!\!\perp \mathbf{x}_i, \mathbf{w}_i, \qquad \varepsilon_i \sim \text{Logistic with cdf } \Lambda(\cdot).$$

Identification: $\mathbb{P}[y_i = 1 | \mathbf{x}_i, \mathbf{x}_j, y_i + y_j = 1, \mathbf{w}_i - \mathbf{w}_j = \mathbf{0}] = \Lambda\big((\mathbf{x}_i - \mathbf{x}_j)^\top \boldsymbol{\theta}_0\big)$, and

$$\boldsymbol{\theta}_0 = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\big[m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) \big| \mathbf{w}_i - \mathbf{w}_j = \mathbf{0}\big]$$

with

$$m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = -\mathbb{1}(y_1 \neq y_2)\big[y_1 \ln \Lambda(\mathbf{x}_2'\boldsymbol{\theta} - \mathbf{x}_1'\boldsymbol{\theta}) + y_2 \ln \Lambda(\mathbf{x}_1'\boldsymbol{\theta} - \mathbf{x}_2'\boldsymbol{\theta})\big].$$

Estimator: localize $\mathbf{w}_i \approx \mathbf{w}_j$, and thus

$$\widehat{\boldsymbol{\theta}}_n = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \binom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) \frac{1}{h^d} K\Big(\frac{\mathbf{w}_i - \mathbf{w}_j}{h}\Big).$$

- $m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$ is convex and differentiable.
- $\widehat{\boldsymbol{\theta}}_n$ not closed-form.

Example 3: Partially Linear Tobit Model

Model:

$$y_i = \max\{\mathbf{x}_i^\top \boldsymbol{\theta}_0 + g(\mathbf{w}_i) + \varepsilon_i, 0\}.$$

Identification: Honoré (1992), and

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, \mathbb{E}\big[m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})\big|\mathbf{w}_i - \mathbf{w}_j = \mathbf{0}\big]$$

with

$$m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = \begin{cases} |y_1| - ((\mathbf{x}_1 - \mathbf{x}_2)'\boldsymbol{\theta} + y_2)\,\mathrm{sgn}(y_1) & \text{if } (\mathbf{x}_1 - \mathbf{x}_2)'\boldsymbol{\theta} \le -y_2 \\ |y_1 - y_2 - (\mathbf{x}_1 - \mathbf{x}_2)'\boldsymbol{\theta}| & \text{if } -y_2 < (\mathbf{x}_1 - \mathbf{x}_2)'\boldsymbol{\theta} < y_1 \\ |y_2| + ((\mathbf{x}_1 - \mathbf{x}_2)'\boldsymbol{\theta} - y_1)\,\mathrm{sgn}(y_2) & \text{if } y_1 \le (\mathbf{x}_1 - \mathbf{x}_2)'\boldsymbol{\theta} \end{cases}.$$

Estimator: localize $\mathbf{w}_i \approx \mathbf{w}_j$, and thus

$$\widehat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min} \binom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) \tfrac{1}{h^d} K\big(\tfrac{\mathbf{w}_i - \mathbf{w}_j}{h}\big).$$

- $m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$ is convex and stochastically differentiable.
- $\widehat{\boldsymbol{\theta}}_n$ not closed-form.

Example 4: Monotone Index Model (not cover)

Model:

$$y_i = G\big(\phi(\mathbf{x}_i'\boldsymbol{\theta}_0, \varepsilon_i), \mathbf{w}_i\big), \qquad \mathbf{x}_i = (x_{i1}, \mathbf{x}_{i2})^\top, \qquad \boldsymbol{\theta}_0 \mapsto (1, \boldsymbol{\theta}_0).$$

with G weakly increasing in first argument, and ϕ strictly increasing.

Identification: Jochmans (2013), and

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, \mathbb{E}\big[m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})\big|\mathbf{w}_i - \mathbf{w}_j = \mathbf{0}\big]$$

with

$$m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = \mathbb{1}(y_i > y_j)\mathbb{1}((\mathbf{x}_i - \mathbf{x}_j)^\top(1, \boldsymbol{\theta}) > 0) + \mathbb{1}(y_i < y_j)\mathbb{1}((\mathbf{x}_i - \mathbf{x}_j)^\top(1, \boldsymbol{\theta}) < 0).$$

Estimator: localize $\mathbf{w}_i \approx \mathbf{w}_j$, and thus

$$\widehat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, \binom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) \tfrac{1}{h^d} K\big(\tfrac{\mathbf{w}_i - \mathbf{w}_j}{h}\big).$$

- $m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$ is non-convex and discontinuous.
- $\widehat{\boldsymbol{\theta}}_n$ not closed-form.

Today: Convex and "Differentiable" Pairwise Difference Estimation

Parameter of Interest:

$$\boldsymbol{\theta}_0 = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} M_0(\boldsymbol{\theta}), \qquad M_0(\boldsymbol{\theta}) = \lim_{n\to\infty} \mathbb{E}\Big[m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}) \tfrac{1}{h^d} K\big(\tfrac{\mathbf{w}_1 - \mathbf{w}_2}{h}\big)\Big].$$

- $\boldsymbol{\theta} \mapsto m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta})$ convex and stochastically differentiable.
- Localization: kernel K re-weights; bandwidth $h \to 0$ (as $n \to \infty$) key tuning parameter.

Estimator:

$$\widehat{\boldsymbol{\theta}}_n = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \widehat{M}_n(\boldsymbol{\theta}), \qquad \widehat{M}_n(\boldsymbol{\theta}) = \binom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) \tfrac{1}{h^d} K\big(\tfrac{\mathbf{w}_i - \mathbf{w}_j}{h}\big).$$

- Inference sensitive to choice of h (lack of robustness).
- convexity $\implies$ fast computation, minimal reg conditions.
- stochastically differentiability $\implies$ Gaussian limiting distribution.

Goal: robust inference with respect to choice of h and features of $m(\cdot)$.

Outline

Distribution Theory: Basics + Bias

$$\boldsymbol{\theta}_0 = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} M_0(\boldsymbol{\theta}), \qquad M_0(\boldsymbol{\theta}) = \lim_{n\to\infty} M_n(\boldsymbol{\theta}).$$

$$\boldsymbol{\theta}_n = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}), \qquad M_n(\boldsymbol{\theta}) = \mathbb{E}\Big[m(\mathbf{z}_1,\mathbf{z}_2;\boldsymbol{\theta})\tfrac{1}{h^d}K\big(\tfrac{\mathbf{w}_1-\mathbf{w}_2}{h}\big)\Big].$$

$$\widehat{\boldsymbol{\theta}}_n = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \widehat{M}_n(\boldsymbol{\theta}), \qquad \widehat{M}_n(\boldsymbol{\theta}) = \binom{n}{2}^{-1}\sum_{i<j} m(\mathbf{z}_i,\mathbf{z}_j;\boldsymbol{\theta})\tfrac{1}{h^d}K\big(\tfrac{\mathbf{w}_i-\mathbf{w}_j}{h}\big).$$

Basic Decomposition. Stochastic contribution and "Localization" bias:

$$r_n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \underbrace{r_n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n)}_{\text{“Variance”}} + \underbrace{r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)}_{\text{“Bias”}}$$

- Bias: not difficult to show that

$$r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) = \operatorname*{arg\,min}_{\mathbf{s}\in\Theta_n} M_n(\boldsymbol{\theta}_0 + \mathbf{s}/r_n) - M_n(\boldsymbol{\theta}_0) = O(r_n h^2)$$

  - Second-order kernel is used to preserved convexity.
  - Higher-order debiasing is obtained via generalized jackknifing (preserves convexity).
  - Usual small-bias condition is: $r_n h^P \to 0$.

Distribution Theory: First-Order Representation

$$\boldsymbol{\theta}_n = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}), \quad M_n(\boldsymbol{\theta}) = \mathbb{E}\Big[m_{12,n}(\boldsymbol{\theta})\Big], \qquad m_{ij,n}(\boldsymbol{\theta}) = m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})\tfrac{1}{h^d}K\big(\tfrac{\mathbf{w}_i - \mathbf{w}_j}{h}\big).$$

$$\widehat{\boldsymbol{\theta}}_n = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \widehat{M}_n(\boldsymbol{\theta}), \quad \widehat{M}_n(\boldsymbol{\theta}) = \tbinom{n}{2}^{-1}\sum_{i<j} m_{ij,n}(\boldsymbol{\theta}).$$

Quadratic Approximation. For all $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|$ small:

$$\widehat{M}_n(\boldsymbol{\theta}) - \widehat{M}_n(\boldsymbol{\theta}_n) \approx (\boldsymbol{\theta} - \boldsymbol{\theta}_n)^\top \widehat{\mathbf{U}}_n + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_n)^\top \widehat{\mathbf{H}}_n (\boldsymbol{\theta} - \boldsymbol{\theta}_n)$$

with

$$\widehat{\mathbf{U}}_n = \tbinom{n}{2}^{-1}\sum_{i<j} \frac{\partial}{\partial\boldsymbol{\theta}} m_{ij,n}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_n}, \qquad \widehat{\mathbf{H}}_n = \tbinom{n}{2}^{-1}\sum_{i<j} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} m_{ij,n}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_n}.$$

First-Order Representation. For $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\| = o_{\mathbb{P}}(1)$:

$$r_n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) = \operatorname*{arg\,min}_{\mathbf{s}\in\Theta_n} \widehat{M}_n(\boldsymbol{\theta}_n + \mathbf{s}/r_n) - \widehat{M}_n(\boldsymbol{\theta}_n) = r_n \widehat{\mathbf{H}}_n^{-1}\widehat{\mathbf{U}}_n$$

Distribution Theory: Hessian

$$r_n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) = r_n \widehat{\mathbf{H}}_n^{-1} \widehat{\mathbf{U}}_n$$

Hessian Approximation. Using the Hoeffding decomposition:

$$\widehat{\mathbf{H}}_n = \binom{n}{2}^{-1} \sum_{i<j} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} m_{ij,n}(\boldsymbol{\theta}_n) = \mathbf{H}_n + \widehat{\mathcal{H}}_{1,n} + \widehat{\mathcal{H}}_{2,n}$$

where

$$\mathbf{H}_n = \mathbb{E}\Big[\tfrac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} m_{ij,n}(\boldsymbol{\theta}_n)\Big] \to \mathbf{H}_0,$$

$$\widehat{\mathcal{H}}_{1,n} = \frac{1}{n} \sum_{i=1}^{n} 2\Big(\mathbb{E}\big[\tfrac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} m_{ij,n}(\boldsymbol{\theta}_n)\big|\mathbf{z}_i\big] - \mathbf{H}_n\Big) = O_{\mathbb{P}}(\tfrac{1}{\sqrt{n}}),$$

$$\begin{aligned}\widehat{\mathcal{H}}_{2,n} &= \binom{n}{2}^{-1} \sum_{i<j} \Big(\tfrac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} m_{ij,n}(\boldsymbol{\theta}_n) - \mathbb{E}[\tfrac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} m_{ij,n}(\boldsymbol{\theta}_n)|\mathbf{z}_i] - \mathbb{E}[\tfrac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} m_{ij,n}(\boldsymbol{\theta}_n)|\mathbf{z}_j] + \mathbf{H}_n\Big) \\ &= O_{\mathbb{P}}(\tfrac{1}{\sqrt{n^2 h^d}}).\end{aligned}$$

Distribution Theory: Gaussian Approximation I

$$r_n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) = r_n \widehat{\mathbf{H}}_n^{-1} \widehat{\mathbf{U}}_n$$

Gaussian Approximation. Using the Hoeffding decomposition:

$$\widehat{\mathbf{U}}_n = \binom{n}{2}^{-1} \sum_{i<j} \frac{\partial}{\partial \boldsymbol{\theta}} m_{ij,n}(\boldsymbol{\theta}_n) = \mathbf{U}_n + \mathcal{U}_{1,n} + \mathcal{U}_{2,n}$$

where

$$\mathbf{U}_n = \mathbb{E}\Big[\tfrac{\partial}{\partial \boldsymbol{\theta}} m_{ij,n}(\boldsymbol{\theta}_n)\Big] = \mathbf{0},$$

$$\widehat{\mathcal{U}}_{1,n} = \frac{r_n}{n} \sum_{i=1}^{n} \frac{2}{h^d} \mathbb{E}\big[\tfrac{\partial}{\partial \boldsymbol{\theta}} m_{ij,n}(\boldsymbol{\theta}_n) \big| \mathbf{z}_i\big] = O_{\mathbb{P}}(\tfrac{r_n}{\sqrt{n}}),$$

$$\begin{aligned}
\widehat{\mathcal{U}}_{2,n} &= r_n \binom{n}{2}^{-1} \sum_{i<j} \Big(\tfrac{\partial}{\partial \boldsymbol{\theta}} m_{ij,n}(\boldsymbol{\theta}_n) - \mathbb{E}[\tfrac{\partial}{\partial \boldsymbol{\theta}} m_{ij,n}(\boldsymbol{\theta}_n) | \mathbf{z}_i] - \mathbb{E}[\tfrac{\partial}{\partial \boldsymbol{\theta}} m_{ij,n}(\boldsymbol{\theta}_n) | \mathbf{z}_j] \Big) \\
&= O_{\mathbb{P}}(\tfrac{r_n}{\sqrt{n^2 h^d}}).
\end{aligned}$$

Distribution Theory: Gaussian Approximation II

$$r_n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) = r_n \widehat{\mathbf{H}}_n^{-1} \widehat{\mathbf{U}}_n, \qquad \mathbb{V}[\widehat{\mathbf{U}}_n] = \mathbb{V}[\widehat{\mathcal{U}}_{1,n}] + \mathbb{V}[\widehat{\mathcal{U}}_{2,n}] \approx \frac{1}{n}\boldsymbol{\Sigma}_0 + \frac{1}{n}\frac{2}{nh^d}\boldsymbol{\Delta}_0$$

Gaussian Approximation. For $nh^d \to \kappa \in (0, \infty] \implies r_n = \sqrt{n}$,

$$\sqrt{n}\, \widehat{\mathbf{U}}_n = \sqrt{n}\mathcal{U}_{1,n} + \sqrt{n}\mathcal{U}_{2,n} \rightsquigarrow \mathsf{N}\Big(\mathbf{0}, \boldsymbol{\Sigma}_0 + \frac{2}{\kappa}\boldsymbol{\Delta}_0\Big)$$

because

$$\sqrt{n}\, \widehat{\mathcal{U}}_{1,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{2}{h^d} \mathbb{E}\big[\tfrac{\partial}{\partial \boldsymbol{\theta}} m_{ij,n}(\boldsymbol{\theta}_n)\big|\mathbf{z}_i\big] \rightsquigarrow \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}_0),$$

$$\begin{aligned}\sqrt{n}\, \widehat{\mathcal{U}}_{2,n} &= \sqrt{n}\tbinom{n}{2}^{-1} \sum_{i<j} \Big(\tfrac{\partial}{\partial \boldsymbol{\theta}} m_{ij,n}(\boldsymbol{\theta}_n) - \mathbb{E}[\tfrac{\partial}{\partial \boldsymbol{\theta}} m_{ij,n}(\boldsymbol{\theta}_n)|\mathbf{z}_i] - \mathbb{E}[\tfrac{\partial}{\partial \boldsymbol{\theta}} m_{ij,n}(\boldsymbol{\theta}_n)|\mathbf{z}_j]\Big) \\ &\rightsquigarrow \mathsf{N}\Big(\mathbf{0}, \frac{2}{\kappa}\boldsymbol{\Delta}_0\Big)\end{aligned}$$

- Distribution theory for $(\widehat{\mathcal{U}}_{1,n}, \widehat{\mathcal{U}}_{2,n})$ follows from martingale CLT.
- We need: $n^2 h^d \to \infty$.

Distribution Theory: Smooth Case

$$r_n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) = r_n \widehat{\mathbf{H}}_n^{-1} \widehat{\mathbf{U}}_n.$$

$$\widehat{\mathbf{H}}_n = \mathbf{H}_n + O_{\mathbb{P}}(n^{-1/2} + n^{-1}h^{d/2}) \to_{\mathbb{P}} \mathbf{H}_0.$$

$$r_n \widehat{\mathbf{U}}_n \approx_d \mathsf{N}\Big(\mathbf{0}, \frac{r_n^2}{n}\boldsymbol{\Sigma}_0 + \frac{2}{n}\frac{r_n^2}{nh^d}\boldsymbol{\Delta}_0\Big).$$

Gaussian Approximation. If $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\| = o_{\mathbb{P}}(1)$ and $n^2h^d \to \infty$, then

$$\mathbf{V}_n^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) \rightsquigarrow \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$$

where

$$\mathbf{V}_n = \mathbf{H}_0^{-1}\Big[\frac{1}{n}\boldsymbol{\Sigma}_0 + \binom{n}{2}^{-1}\frac{1}{h^d}\boldsymbol{\Delta}_0\Big]\mathbf{H}_0^{-1}.$$

- If $nh^d \to \infty$, the classical result: Asymptotic linear representation.
- If $n^2h^d \to \infty$, the small bandwidth asymptotics: robust (to h) distribution theory.

Recap and Outstanding Issues

Gaussian Approximation. If $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\| = o_{\mathbb{P}}(1)$, and

$$n^2 h^d \to \infty \qquad \text{and} \qquad n \min\{1, nh^d\} h^{2P} \to 0,$$

then

$$\mathbf{V}_n^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$$

where

$$\mathbf{V}_n = \mathbf{H}_0^{-1}\Big[\frac{1}{n}\boldsymbol{\Sigma}_0 + \binom{n}{2}^{-1}\frac{1}{h^d}\boldsymbol{\Delta}_0\Big]\mathbf{H}_0^{-1}.$$

Outstanding Issues.

- How to handle stochastically differentiable $m_{ij}(\cdot)$?
  - Answer: empirical process theory for U-processes. (Technically tedious.)
  - Important: $\mathbf{H}_0$, $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Delta}_0$ can be function of nuisance functions. ($\approx$ quantile regression.)
  - Thus, plug-in estimates are possible but not advisable if avoidable...
- How to conduct robust inference in general?
  - Answer: careful application of the bootstrap.

Outline

Nonparametric Bootstrap

Estimator: $\mathbf{z}_i = (y_i, \mathbf{x}_i, \mathbf{w}_i)^\top$, $i = 1, \dots, n$, random sample + regularity conditions,

$$\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{M}_n(\boldsymbol{\theta}), \qquad \widehat{M}_n(\boldsymbol{\theta}) = \tbinom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) \tfrac{1}{h^d} K\big(\tfrac{\mathbf{w}_i - \mathbf{w}_j}{h}\big).$$

Gaussian Approximation: if $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\| = o_{\mathbb{P}}(1)$, $n^2 h^d \to \infty$, and $n \min\{1, nh^d\} h^{2P} \to 0$,

$$\mathbf{V}_n^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$$

where

$$\mathbf{V}_n = \mathbf{H}_0^{-1}\Big[\frac{1}{n}\boldsymbol{\Sigma}_0 + \binom{n}{2}^{-1}\frac{1}{h^d}\boldsymbol{\Delta}_0\Big]\mathbf{H}_0^{-1}.$$

Bootstrap: $\mathbf{z}_i^* = (y_i^*, \mathbf{x}_i^*, \mathbf{w}_i^*)^\top$, $i = 1, \dots, n$, random sample + regularity conditions.

$$\widehat{\boldsymbol{\theta}}_n^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{M}_n^*(\boldsymbol{\theta}), \qquad \widehat{M}_n^*(\boldsymbol{\theta}) = \tbinom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_i^*, \mathbf{z}_j^*; \boldsymbol{\theta}) \tfrac{1}{h^d} K\big(\tfrac{\mathbf{w}_i^* - \mathbf{w}_j^*}{h}\big).$$

Distribution Theory: Bootstrap

$$\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{M}_n(\boldsymbol{\theta}), \qquad \widehat{M}_n(\boldsymbol{\theta}) = \binom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) \frac{1}{h^d} K\Big(\frac{\mathbf{w}_i - \mathbf{w}_j}{h}\Big).$$

$$\widehat{\boldsymbol{\theta}}^*_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{M}^*_n(\boldsymbol{\theta}), \qquad \widehat{M}^*_n(\boldsymbol{\theta}) = \binom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}^*_i, \mathbf{z}^*_j; \boldsymbol{\theta}) \frac{1}{h^d} K\Big(\frac{\mathbf{w}^*_i - \mathbf{w}^*_j}{h}\Big).$$

Gaussian Approximation: if $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\| = o_{\mathbb{P}}(1)$, $n^2h^d \to \infty$, and $n \min\{1, nh^d\} h^{2P} \to 0$,

$$\mathbf{V}_n^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$$

where

$$\mathbf{V}_n = \mathbf{H}_0^{-1}\Big[\frac{1}{n}\boldsymbol{\Sigma}_0 + \binom{n}{2}^{-1}\frac{1}{h^d}\boldsymbol{\Delta}_0\Big]\mathbf{H}_0^{-1}.$$

Bootstrap Approximation: if $\|\widehat{\boldsymbol{\theta}}^*_n - \widehat{\boldsymbol{\theta}}_n\| = o_{\mathbb{P}}(1)$ and $n^2h^d \to \infty$,

$$\mathbf{V}_n^{*\,-1/2}(\widehat{\boldsymbol{\theta}}^*_n - \widehat{\boldsymbol{\theta}}_n) \rightsquigarrow_{\mathbb{P}} \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$$

where

$$\mathbf{V}^*_n = \mathbf{H}_0^{-1}\Big[\frac{1}{n}\boldsymbol{\Sigma}_0 + 3\binom{n}{2}^{-1}\frac{1}{h^d}\boldsymbol{\Delta}_0\Big]\mathbf{H}_0^{-1}.$$

Distribution Theory: Bootstrap Validity

$$\widehat{\boldsymbol{\theta}}_n(h) = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, \widehat{M}_n(\boldsymbol{\theta};h), \qquad \widehat{M}_n(\boldsymbol{\theta};h) = \binom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_i,\mathbf{z}_j;\boldsymbol{\theta})\,\frac{1}{h^d} K\Big(\frac{\mathbf{w}_i-\mathbf{w}_j}{h}\Big).$$

$$\widehat{\boldsymbol{\theta}}^*_n(h) = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, \widehat{M}^*_n(\boldsymbol{\theta};h), \qquad \widehat{M}^*_n(\boldsymbol{\theta};h) = \binom{n}{2}^{-1} m(\mathbf{z}^*_i,\mathbf{z}7^*_j;\boldsymbol{\theta})\,\frac{1}{h^d} K\Big(\frac{\mathbf{w}^*_i-\mathbf{w}^*_j}{h}\Big).$$

$$\mathbf{V}_n(h)^{-1/2}(\widehat{\boldsymbol{\theta}}_n(h)-\boldsymbol{\theta}_0) \rightsquigarrow \mathsf{N}(\mathbf{0},\mathbf{I}_d), \qquad \mathbf{V}_n(h) = \mathbf{H}_0^{-1}\Big[\frac{1}{n}\boldsymbol{\Sigma}_0 + \binom{n}{2}^{-1}\frac{1}{h^d}\boldsymbol{\Delta}_0\Big]\mathbf{H}_0^{-1}.$$

$$\mathbf{V}^*_n(h)^{-1/2}(\widehat{\boldsymbol{\theta}}^*_n(h)-\widehat{\boldsymbol{\theta}}_n(h)) \rightsquigarrow_{\mathbb{P}} \mathsf{N}(\mathbf{0},\mathbf{I}_d), \qquad \mathbf{V}^*_n(h) = \mathbf{H}_0^{-1}\Big[\frac{1}{n}\boldsymbol{\Sigma}_0 + 3\binom{n}{2}^{-1}\frac{1}{h^d}\boldsymbol{\Delta}_0\Big]\mathbf{H}_0^{-1}.$$

Valid Bootstrap:

$$\widehat{\boldsymbol{\theta}}^*_n(3^{1/d}h) = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, \widehat{M}^*_n(\boldsymbol{\theta};3^{1/d}h), \qquad \widehat{M}^*_n(\boldsymbol{\theta};3^{1/d}h) = \binom{n}{2}^{-1} m(\mathbf{z}^*_i,\mathbf{z}^*_j;\boldsymbol{\theta})\,\frac{1}{h^d} K\Big(\frac{\mathbf{w}^*_i-\mathbf{w}^*_j}{3^{1/d}h}\Big).$$

$$\mathbf{V}^*_n(3^{1/d}h)^{-1/2}(\widehat{\boldsymbol{\theta}}^*_n(3^{1/d}h)-\widehat{\boldsymbol{\theta}}_n(3^{1/d}h)) \rightsquigarrow_{\mathbb{P}} \mathsf{N}(\mathbf{0},\mathbf{I}_d),$$

$$\mathbf{V}^*_n(3^{1/d}h)^{-1/2} = \mathbf{H}_0^{-1}\Big[\frac{1}{n}\boldsymbol{\Sigma}_0 + \binom{n}{2}^{-1}\frac{1}{h^d}\boldsymbol{\Delta}_0\Big]\mathbf{H}_0^{-1}.$$

Robust Bootstrap Inference

$$
\begin{aligned}
\boldsymbol{\theta}_0 &= \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} M_0(\boldsymbol{\theta}), & M_0(\boldsymbol{\theta}) &= \lim_{n\to\infty} M_n(\boldsymbol{\theta}). \\
\boldsymbol{\theta}_n &= \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}), & M_n(\boldsymbol{\theta}) &= \mathbb{E}\Big[m(\mathbf{z}_1,\mathbf{z}_2;\boldsymbol{\theta})\tfrac{1}{h^d}K\big(\tfrac{\mathbf{w}_1-\mathbf{w}_2}{h}\big)\Big]. \\
\widehat{\boldsymbol{\theta}}_n(h) &= \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \widehat{M}_n(\boldsymbol{\theta};h), & \widehat{M}_n(\boldsymbol{\theta};h) &= \tbinom{n}{2}^{-1}\sum_{i<j} m(\mathbf{z}_i,\mathbf{z}_j;\boldsymbol{\theta})\tfrac{1}{h^d}K\big(\tfrac{\mathbf{w}_i-\mathbf{w}_j}{h}\big). \\
\widehat{\boldsymbol{\theta}}_n^*(3^{1/d}h) &= \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \widehat{M}_n^*(\boldsymbol{\theta};3^{1/d}h), & \widehat{M}_n^*(\boldsymbol{\theta};3^{1/d}h) &= \tbinom{n}{2}^{-1} m(\mathbf{z}_i^*,\mathbf{z}_j^*;\boldsymbol{\theta})\tfrac{1}{h^d}K\big(\tfrac{\mathbf{w}_i^*-\mathbf{w}_j^*}{3^{1/d}h}\big).
\end{aligned}
$$

Robust ("Small Bandwidth Asymptotics") Confidence Intervals: $\alpha \in (0,1)$,

$$\mathrm{CI}_{1-\alpha}(h) = \Big[\widehat{\boldsymbol{\theta}}_n(h) - \mathfrak{c}^*_{1-\alpha/2}\ ,\ \widehat{\boldsymbol{\theta}}_n(h) - \mathfrak{c}^*_{\alpha/2}\Big]$$

$$\mathfrak{c}^*_\alpha = \inf\Big\{c \in \mathbb{R} : \mathbb{P}^*\big[\widehat{\boldsymbol{\theta}}^*_n(3^{1/d}h) - \widehat{\boldsymbol{\theta}}_n(3^{1/d}h) \leq c\big] \geq \alpha\Big\}$$

Outline

Overview

- Pairwise Difference Estimation is used in economics and other disciplines.
- Rely on “localization” as determined by bandwidth h.
- Classical distributional approximations are sensitive to h.
- New distribution theory and bootstrap-based inference more robust to h.
    - Small bandwidth asymptotics.
- Results require stochastic differentiability of objective function.
- Upcoming research: non-convex and/or discontinuous objective function.
    - Distribution theory: Mixture of “Small Bandwidth” Gaussian and Chernoff-type laws.
    - New way to conduct robust inference due to the non-Gaussian limit.