

# Robust Inference for Pairwise Estimation

Matias D. Cattaneo  
Princeton University

Michael Jansson  
UC Berkeley

Kenichi Nagasawa  
University of Warwick

December 2024

Talk based on:

- ▶ “Robust Inference for Stochastically Differentiable Pairwise Difference Estimation”.
- ▶ “Pairwise Estimation: Distribution Theory and Bootstrap Inference”.

# Outline

## 1. Introduction

## 2. Distribution Theory

## 3. Robust Bootstrap Inference

## 4. Final Remarks

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

Parameter of Interest:

$$\boldsymbol{\theta}_0 = \arg \min_{\boldsymbol{\theta} \in \Theta} M_0(\boldsymbol{\theta}), \quad M_0(\boldsymbol{\theta}) = \lim_{n \rightarrow \infty} \mathbb{E} \left[ m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}) \frac{1}{h^d} K \left( \frac{\mathbf{w}_1 - \mathbf{w}_2}{h} \right) \right].$$

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

Estimator:

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

- ▶ 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:

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

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

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

with

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

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

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

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

## Example 2: Partially Linear Logistic Model

Model:

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

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

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

with

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

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

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

- ▶  $m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$  is differentiable.
- ▶  $\hat{\boldsymbol{\theta}}_n$  **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

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

with

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

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

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

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

## Example 4: Monotone Index Model (not today)

Model:

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

with  $G$  weakly increasing in first argument, and  $\phi$  strictly increasing.

Identification: Jochmans (2013), and

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

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

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

►  $m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$  is **discontinuous**.

►  $\hat{\boldsymbol{\theta}}_n$  **not** closed-form.



# Today: Differentiable Pairwise Difference Estimation

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

Parameter of Interest:

$$\boldsymbol{\theta}_0 = \arg \min_{\boldsymbol{\theta} \in \Theta} M_0(\boldsymbol{\theta}), \quad M_0(\boldsymbol{\theta}) = \lim_{n \rightarrow \infty} \mathbb{E} \left[ m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}) \frac{1}{h^d} K \left( \frac{\mathbf{w}_1 - \mathbf{w}_2}{h} \right) \right].$$

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

Estimator:

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

- ▶ Inference sensitive to choice of  $h$  (lack of robustness).
- ▶  $\boldsymbol{\theta} \mapsto m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta})$  **stochastically differentiable**  $\implies$  Gaussian limiting distribution.

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

# Outline

1. Introduction

2. Distribution Theory

3. Robust Bootstrap Inference

4. Final Remarks

# Today: Differentiable Pairwise Difference Estimation

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

Parameter of Interest:

$$\theta_0 = \arg \min_{\theta \in \Theta} M_0(\theta), \quad M_0(\theta) = \lim_{n \rightarrow \infty} M_n(\theta).$$

Fixed- $h$  Estimands:

$$\theta_n = \arg \min_{\theta \in \Theta} M_n(\theta), \quad M_n(\theta) = \mathbb{E} \left[ m(\mathbf{z}_1, \mathbf{z}_2; \theta) \frac{1}{h^d} K \left( \frac{\mathbf{w}_1 - \mathbf{w}_2}{h} \right) \right].$$

Estimator:

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

- Inference sensitive to choice of  $h$  (lack of robustness).
- $\theta \mapsto m(\mathbf{z}_1, \mathbf{z}_2; \theta)$  **stochastically differentiable**  $\implies$  Gaussian limiting distribution.

## Distribution Theory: Basics + Bias

$$\boldsymbol{\theta}_0 = \arg \min_{\boldsymbol{\theta} \in \Theta} M_0(\boldsymbol{\theta}), \quad M_0(\boldsymbol{\theta}) = \lim_{n \rightarrow \infty} M_n(\boldsymbol{\theta}).$$

$$\boldsymbol{\theta}_n = \arg \min_{\boldsymbol{\theta} \in \Theta} M_n(\boldsymbol{\theta}), \quad M_n(\boldsymbol{\theta}) = \mathbb{E} \left[ m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}) \frac{1}{h^d} K \left( \frac{\mathbf{w}_1 - \mathbf{w}_2}{h} \right) \right].$$

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

**Basic Decomposition.** Stochastic contribution and “Localization” bias:

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

► **Bias:** not difficult to show that

$$r_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) = \arg \min_{\mathbf{s} \in \Theta_n} M_n(\boldsymbol{\theta}_0 + \mathbf{s}/r_n) - M_n(\boldsymbol{\theta}_0) = O(r_n h^P)$$

►  $P$  denotes the order of the kernel  $K$ .

► Thus, we need:  $r_n h^P \rightarrow 0$ .

# Distribution Theory: First-Order Representation

$$\boldsymbol{\theta}_n = \arg \min_{\boldsymbol{\theta} \in \Theta} M_n(\boldsymbol{\theta}), \quad M_n(\boldsymbol{\theta}) = \mathbb{E} \left[ m_{12,n}(\boldsymbol{\theta}) \right], \quad m_{ij,n}(\boldsymbol{\theta}) = m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) \frac{1}{h^d} K\left(\frac{\mathbf{w}_i - \mathbf{w}_j}{h}\right).$$

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

**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

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

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

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

## Distribution Theory: Hessian

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

**Hessian Approximation.** Using the Hoeffding decomposition:

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

where

$$\mathbf{H}_n = \mathbb{E} \left[ \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} m_{ij,n}(\boldsymbol{\theta}_n) \right] \rightarrow \mathbf{H}_0,$$

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

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

# Distribution Theory: Gaussian Approximation I

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

**Gaussian Approximation.** Using the Hoeffding decomposition:

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

where

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

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

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

## Distribution Theory: Gaussian Approximation II

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

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

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

because

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

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

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



## Distribution Theory: Smooth Case

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

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

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

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

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

where

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

- If  $nh^d \rightarrow \infty$ , the classical result: Asymptotic linear representation.
- If  $n^2 h^d \rightarrow \infty$ , 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

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

then

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

where

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

## 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

1. Introduction

2. Distribution Theory

3. Robust Bootstrap Inference

4. Final Remarks

# Nonparametric Bootstrap

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

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

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

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

where

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

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

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

## Distribution Theory: Bootstrap

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

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

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

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

where

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

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

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

where

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

# Distribution Theory: Bootstrap Validity

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

$$\widehat{\theta}_n^*(h) = \arg \min_{\theta \in \Theta} \widehat{M}_n^*(\theta; h), \quad \widehat{M}_n^*(\theta; h) = \binom{n}{2}^{-1} m(\mathbf{z}_i^*, \mathbf{z}_j^*; \theta) \frac{1}{h^d} K\left(\frac{\mathbf{w}_i^* - \mathbf{w}_j^*}{h}\right).$$

$$\mathbf{V}_n(h)^{-1/2}(\widehat{\theta}_n(h) - \theta_0) \rightsquigarrow \mathbf{N}(\mathbf{0}, \mathbf{I}_d), \quad \mathbf{V}_n(h) = \mathbf{H}_0^{-1} \left[ \frac{1}{n} \Sigma_0 + \binom{n}{2}^{-1} \frac{1}{h^d} \Delta_0 \right] \mathbf{H}_0^{-1}.$$

$$\mathbf{V}_n^*(h)^{-1/2}(\widehat{\theta}_n^*(h) - \widehat{\theta}_n(h)) \rightsquigarrow_{\mathbb{P}} \mathbf{N}(\mathbf{0}, \mathbf{I}_d), \quad \mathbf{V}_n^*(h) = \mathbf{H}_0^{-1} \left[ \frac{1}{n} \Sigma_0 + 3 \binom{n}{2}^{-1} \frac{1}{h^d} \Delta_0 \right] \mathbf{H}_0^{-1}.$$

Valid Bootstrap:

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

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

$$\mathbf{V}_n^*(3^{1/d}h)^{-1/2} = \mathbf{H}_0^{-1} \left[ \frac{1}{n} \Sigma_0 + \binom{n}{2}^{-1} \frac{1}{h^d} \Delta_0 \right] \mathbf{H}_0^{-1}.$$

# Robust Bootstrap Inference

$$\boldsymbol{\theta}_0 = \arg \min_{\boldsymbol{\theta} \in \Theta} M_0(\boldsymbol{\theta}), \quad M_0(\boldsymbol{\theta}) = \lim_{n \rightarrow \infty} M_n(\boldsymbol{\theta}).$$

$$\boldsymbol{\theta}_n = \arg \min_{\boldsymbol{\theta} \in \Theta} M_n(\boldsymbol{\theta}), \quad M_n(\boldsymbol{\theta}) = \mathbb{E} \left[ m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}) \frac{1}{h^d} K \left( \frac{\mathbf{w}_1 - \mathbf{w}_2}{h} \right) \right].$$

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

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

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

$$\text{CI}_{1-\alpha}(h) = \left[ \widehat{\boldsymbol{\theta}}_n(h) - \mathbf{c}_{1-\alpha/2}^*, \widehat{\boldsymbol{\theta}}_n(h) - \mathbf{c}_{\alpha/2}^* \right]$$

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

# Outline

1. Introduction
2. Distribution Theory
3. Robust Bootstrap Inference
4. Final Remarks



# 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: 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.