Nonparametric Regression under Cluster Sampling

Yuya Shimizu

University of Wisconsin-Madison

EEA-ESEM August, 2024

Cluster sampling

- Non i.i.d. dataset
  - independent between different clusters
  - allow dependence within the same cluster
- Cluster structure is common in economics
  - school, family, hospital, firm, industry, region,...
- Researcher knows cluster $g = 1, \ldots, G$ and observes $\left\{ \{(Y_{gj}, X_{gj})\}_{j=1}^{n_g} \right\}_{g=1}^{G}$
  - each observation can be grouped into one cluster
  - n_g: cluster size of g-th cluster
  - $n = \sum_{g=1}^{G} n_g$: observations in total
  - $(Y_{gj}, X_{gj}) \perp (Y_{g'\ell}, X_{g'\ell})$ but $(Y_{gj}, X_{gj}) \not\perp (Y_{g\ell}, X_{g\ell})$ in general

Contribution

1. Derive asymptotic properties of nonparametric regression under cluster sampling
   - We allow unbounded and heterogeneous cluster sizes n_g
2. Can cover both individual and cluster-level regressors
3. Propose cluster-robust bandwidth selection methods
4. Verify cluster-robust variance estimator
   - Unbounded cluster causes dependence even in local neighborhood

- Potential applications
  - Semiparametric regression
  - Nonparametric auction estimation
  - Regression discontinuity design

Related literature

- Cluster sampling in econometrics
  - C. B. Hansen (2007): parametric regression with homogeneous cluster sizes
  - Djogbenou, MacKinnon, and Nielsen (2019); B. E. Hansen and S. Lee (2019): heterogeneous cluster sizes
- Nonparametric regression under cluster dependence
  - limited, even with homogeneous cluster sizes
  - Lin and Carroll (2000); Wang (2003); Bhattacharya (2005); P. Hu, Peng, and X. Hu (2024)
- Nonparametric regressions with other dependence
  - Robinson (1983), B. E. Hansen (2008), Vogt (2012): time series dependence
  - Robinson (2011), J. Lee and Robinson (2016): spatial dependence

Outline

1. Setup

2. Asymptotic Theory

3. Simulation

Outline

1. Setup

2. Asymptotic Theory

3. Simulation

Data generating process

- Outcome: $Y_{gj} \in \mathbb{R}$, Regressor: $X_{gj} = \left(X_{gj}^{(\text{ind})\top}, X_g^{(\text{cls})\top}\right)^\top \in \mathbb{R}^d = \mathbb{R}^{d_{\text{ind}}} \times \mathbb{R}^{d_{\text{cls}}}$
- DGP:

$$\begin{aligned} Y_{gj} &= m\left(X_{gj}\right) + e_{gj}, \\ \mathbb{E}\left[e_{gj} \mid \mathbf{X}_g\right] &= \mathbb{E}\left[e_{gj} \mid X_{gj}\right] = 0, \\ \mathbb{E}\left[e_{gj}^2 \mid \mathbf{X}_g\right] &= \mathbb{E}\left[e_{gj}^2 \mid X_{gj}\right] = \sigma^2\left(X_{gj}\right), \\ \mathbb{E}\left[e_{gj}e_{g\ell} \mid \mathbf{X}_g\right] &= \mathbb{E}\left[e_{gj}e_{g\ell} \mid X_{gj}^{(\text{ind})}, X_{g\ell}^{(\text{ind})}; X_g^{(\text{cls})}\right] \\ &= \sigma\left(X_{gj}^{(\text{ind})}, X_{g\ell}^{(\text{ind})}; X_g^{(\text{cls})}\right) \text{ for } j \neq \ell. \end{aligned} \tag{1}$$

where

$$\mathbf{X}_g = \left(X_{g1}, \ldots, X_{gn_g}\right)$$

- Goal: estimate $\mathbb{E}\left[Y_{gj} \mid \mathbf{X}_g\right] = m\left(X_{gj}\right)$
- This setup allows cluster random effects and cluster-level regressors
- Assume that $d_{\text{ind}} \geq 1$
- Assume that X_{gj} have identical marginal distribution with the density $f(X_{gj})$

Nadaraya-Watson estimator

- NW estimator is

$$\hat{m}_{\mathsf{nw}}(x) = \frac{\sum_{g=1}^{G} \sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right) Y_{gj}}{\sum_{g=1}^{G} \sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)}, \tag{2}$$

  where
  - $h > 0$ is bandwidth
  - $K : \mathbb{R}^d \to \mathbb{R}$ is a product kernel function $K(X) = \prod_{q=1}^{d} k\left(X^{(q)}\right)$
  - $k : \mathbb{R} \to \mathbb{R}$ is a univariate kernel function satisfying
    - boundedness
    - symmetry
    - $\int_{-\infty}^{\infty} k(u)\mathrm{d}u = 1$ (normalization)
    - $\int_{-\infty}^{\infty} u^2 k(u)\mathrm{d}u \equiv \kappa_2 < \infty$ and $\int_{-\infty}^{\infty} u^4 k(u)\mathrm{d}u < \infty$

Local linear estimator

- LL estimator is

$$\hat{m}_{\mathsf{LL}}(x) = \sum_{g=1}^{G}\sum_{j=1}^{n_g} K_{\mathsf{LL}}(X_{gj}, x)\, Y_{gj}, \tag{3}$$

  where

$$K_{\mathsf{LL}}(u, x) = \mathbf{e}_1^\top \left(\mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x\right)^{-1} \begin{bmatrix} 1 \\ u - x \end{bmatrix} K_h(u - x),$$

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \mathbf{X}_x = \begin{bmatrix} 1 & (X_{1,1} - x)^\top \\ \vdots & \vdots \\ 1 & (X_{G,n_G} - x)^\top \end{bmatrix}, \mathbf{W}_x = \begin{bmatrix} K_h(X_{1,1} - x) & & O \\ & \ddots & \\ O & & K_h(X_{G,n_G} - x) \end{bmatrix},$$

  and $K_h(\cdot) = \frac{1}{h^d} K\left(\frac{\cdot}{h}\right)$

- $\hat{m}_{\mathsf{LL}}(x)$ is also the minimizer β_0 for the localized squared error

$$\min_{\beta_0, \beta_1} \sum_{g=1}^{G}\sum_{j=1}^{n_g} K\left(\frac{X_{gj} - x}{h}\right)\left(Y_{gj} - \beta_0 - \beta_1^\top (X_{gj} - x)\right)^2$$

Outline

Main assumptions

Assumption

1. $nh^d \to \infty$
2. $h^d \to 0$ and $\left(\max_{g \leq G} n_g\right) h^{d_{\text{ind}}} = O(1)$
3. *[Smoothness conditions]: there exists some neighborhood* $\mathcal{N}$ *of* $x = \left(x^{(\text{ind})\top}, x^{(\text{cls})\top}\right)^\top$ *such that*
   - $m(x)$ *and* $f(x)$ *are twice continuously differentiable*
   - $f\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$ *is continuously differentiable*
   - *densities for* $\left(X_{gj}^{(\text{ind})}, X_{g\ell}^{(\text{ind})}, X_{gt}^{(\text{ind})}; X_g^{(\text{cls})}\right)$ *and* $\left(X_{gj}^{(\text{ind})}, X_{g\ell}^{(\text{ind})}, X_{gt}^{(\text{ind})}, X_{gs}^{(\text{ind})}; X_g^{(\text{cls})}\right)$, $\sigma^2(x)$, *and* $\sigma\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$ *are continuous*
4. $f(x) > 0$ *and (only for LL)* K *has a compact support*

- **Remark:** Condition 2 is stronger than $h^d \to 0$
- **Remark:** Condition 1 and 2 implies $\left(\max_{g \leq G} n_g\right)/n \to 0$ and $G \to \infty$

Asymptotic bias

- Let $\kappa_2 = \int_{-\infty}^{\infty} u^2 k(u)\mathrm{d}u$

Theorem

As $nh^d \to \infty$, $\left(\max_{g\leq G} n_g\right) h^{d_{\text{ind}}} = O(1)$,

$$\mathbb{E}\left[\hat{m}_{\text{nw}}(x) \mid X_1, \cdots, X_G\right] = m(x) + h^2 B_{\text{nw}}(x) + o_p\left(h^2\right) + O_p\left(\sqrt{\frac{1}{nh^{d-2}}}\right),$$

where $B_{\text{nw}}(x) = \kappa_2 \sum_{q=1}^{d} \left(\frac{1}{2} m_{qq}(x) + f(x)^{-1} f_q(x) m_q(x)\right)$. *Also,*

$$\mathbb{E}\left[\hat{m}_{\text{LL}}(x) \mid X_1, \cdots, X_G\right] = m(x) + h^2 B_{\text{LL}}(x) + o_p\left(h^2\right),$$

where $B_{\text{LL}}(x) = \frac{1}{2}\kappa_2 \sum_{q=1}^{d} m_{qq}(x)$.

- **Remark:** bias is same as the i.i.d. case, but require generalized conditions

Asymptotic variance

- $\hat{m}_*(x) \in \{\hat{m}_{\text{nw}}(x), \hat{m}_{\text{LL}}(x)\}$

Theorem

As $nh^d \to \infty$, $\left(\max_{g\leq G} n_g\right) h^{d_{\text{ind}}} = O(1)$, *and* $\left(\frac{1}{n}\sum_{g=1}^{G} n_g^2\right) h^{d_{\text{ind}}} \to \lambda \in [0, \infty)$,

$$\begin{aligned} & \operatorname{Var}\left[\hat{m}_*(x) \mid X_1, \cdots, X_G\right] \\ = \; & \frac{R_k^d \sigma^2(x)}{f(x) nh^d} + \frac{\lambda R_k^{d_{\text{cls}}} f\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right) \sigma\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)}{f(x)^2 nh^d} + o_p\left(\frac{1}{nh^d}\right), \end{aligned}$$

where

$$R_k = \int_{-\infty}^{\infty} k(u)^2 \, \mathrm{d}u.$$

- **Remark:** variance has an additional term due to cluster dependence

Consistency and asymptotic normality

Theorem

For $\hat{m}_*(x) \in \{\hat{m}_{\mathrm{nw}}(x), \hat{m}_{\mathrm{LL}}(x)\}$

- $\hat{m}_*(x) \xrightarrow{p} m(x)$
- *Under additional assumptions,* ⏩ detail

$$\sqrt{nh^d}\left(\hat{m}_*(x) - m(x) - h^2 B_*(x)\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{R_k^d \sigma^2(x)}{f(x)} + \frac{\lambda R_k^{d_{\mathrm{cls}}} f\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) \sigma\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)}{f(x)^2}\right)$$

if $\left(\frac{1}{n}\sum_{g=1}^{G} n_g^2\right) h^{d_{\mathrm{ind}}} \to \lambda \in [0, \infty)$

- **Remark:** bias $h^2 B_*(x)$ exists
  - vanishes if undersmoothing $nh^{d+4} = o(1)$ holds
- **Remark:** additional term exists
  - vanishes if $\left(\max_{g \leq G} n_g\right) h^{d_{\mathrm{ind}}} = o(1)$ holds

Uniform convergence

Theorem

- *Suppose that*
  - c_n *is a growing sequence satisfying the condition*

$$c_n = O\left(\left(\max_{g \le G} n_g\right)^{2/d} (\log n)^{1/d}\right), \tag{4}$$

and for some $s \geq 2$,

$$\frac{\left(\max_{g \le G} n_g\right)^2 \log n}{n^{1-(2/s)} h^d} = O(1) \tag{5}$$

  - *smoothness conditions hold uniformly*
  - *some regularity conditions* ⏩ *detail*
- *Then,*

$$\sup_{\|x\| \le c_n} \left|\hat{m}_*(x) - m(x)\right| = o_p(1) \tag{6}$$

⏩ *proof*

Cluster-Robust variance estimation

- CR-variance estimator

$$\hat{V} = \frac{R_k^d \hat{\sigma}^2_{\mathrm{nw}}(x)}{\hat{f}(x)} + \frac{\hat{\lambda} R_k^{d_{\mathrm{cls}}} \hat{f}\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) \hat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)}{\left(\hat{f}(x)\right)^2}$$

  where $\hat{\sigma}^2_{\mathrm{nw}}(x)$ and $\hat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ are estimated using $\hat{e}_{gj} = Y_{gj} - \hat{m}_{\mathrm{nw}}(X_{gj})$

- We can construct CI form this after normalization
  - 95% CI for $m(x) + h^2 B_{\mathsf{nw}}(x)$ is

$$\left[\hat{m}_{\mathsf{nw}}(x) - \frac{1.96 \times \sqrt{\hat{V}}}{\sqrt{nh^d}}, \hat{m}_{\mathsf{nw}}(x) + \frac{1.96 \times \sqrt{\hat{V}}}{\sqrt{nh^d}}\right]$$

Outline

1. Setup

2. Asymptotic Theory

3. Simulation

Data generating process

- Number of clusters: $G = 100$
- Cluster sizes
  - $n_g = 20$ for $g = 1, \ldots, G-1$
  - $n_G \in \{20, 100\}$
    - $\left(\max_{g \leq G} n_g\right)/n \approx \{0.02, 0.09\}$
- Generated 500 datasets from Setup 1 or 2
- **Setup 1 (homoskedastic errors):**

$$Y_{gj} = \sin\left(2X_{gj}\right) + 2\exp\left(-16X_{gj}^2\right) + 0.5e_{gj}, \tag{7}$$

  or

  Setup 2 (heteroskedastic errors):

$$Y_{gj} = X_{gj}\sin\left(2\pi X_{gj}\right) + \sigma\left(X_{gj}\right)e_{gj}, \tag{8}$$

$$\sigma\left(X_{gj}\right) = \frac{2 + \cos\left(2\pi X_{gj}\right)}{5},$$

- where $X_{gj} = \sqrt{\rho_x}\left(X_1\right)_g + \sqrt{1-\rho_x}\left(X_2\right)_{gj}$ and $e_{gj} = \sqrt{\rho_e}c_g + \sqrt{1-\rho_e}u_{gj}$
  - $\left(X_1\right)_g \sim \mathcal{N}(0,1)$, $\left(X_2\right)_{gj} \sim \mathcal{N}(0,1)$, $c_g \sim \mathcal{N}(0,1)$, and $u_{gj} \sim \mathcal{N}(0,1)$ independently

Simulation setup

- Given true bias
- Compare three 95% CIs
  1. $[CI]$ ignore additional term in variance
  2. $[CI_{\mathrm{CR}}]$ ignore additional term in variance, but use jackknife estimator for $\sigma^2(x)$
  3. $[CI_{\lambda}]$ estimate additional term in variance, and use jackknife estimator for $\sigma^2(x)$
- Evaluated by coverage
- Use analytical bias correction with bias
- Bandwidth is selected in cluster-robust way detail

Coverage and average length of 95% CI for each standard error (m_{nw}, Setup 1, $x = 0.75$)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.925	0.931	0.954	0.915	0.920	0.952
	{0.192}	{0.195}	{0.217}	{0.189}	{0.192}	{0.217}
(ρ_X, ρ_e)=(0.2,0.5)	0.879	0.886	0.960	0.861	0.869	0.951
	{0.192}	{0.195}	{0.246}	{0.188}	{0.192}	{0.250}
(ρ_X, ρ_e)=(0.5,0.2)	0.920	0.925	0.956	0.906	0.908	0.954
	{0.191}	{0.194}	{0.227}	{0.188}	{0.191}	{0.228}
(ρ_X, ρ_e)=(0.5,0.5)	0.857	0.867	0.964	0.833	0.844	0.957
	{0.191}	{0.195}	{0.261}	{0.188}	{0.191}	{0.267}

Coverage and average length of 95% CI for each standard error (m_{nw}, Setup 2, $x = 0.8$)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.893	0.897	0.931	0.882	0.886	0.923
	{0.167}	{0.170}	{0.187}	{0.165}	{0.168}	{0.187}
(ρ_X, ρ_e)=(0.2,0.5)	0.844	0.850	0.918	0.831	0.836	0.924
	{0.167}	{0.171}	{0.209}	{0.164}	{0.168}	{0.211}
(ρ_X, ρ_e)=(0.5,0.2)	0.903	0.909	0.936	0.878	0.884	0.927
	{0.166}	{0.170}	{0.191}	{0.164}	{0.167}	{0.192}
(ρ_X, ρ_e)=(0.5,0.5)	0.826	0.837	0.932	0.806	0.816	0.924
	{0.166}	{0.170}	{0.218}	{0.164}	{0.167}	{0.223}

Other results covered in the paper

- Bandwidth selection ⏩ detail
- Simulation for bandwidth selection ⏩ detail
- Empirical illustration ⏩ detail

Conclusion and future work

Conclusion

- We allow growing and bounded size clusters
- The theory can cover cluster-level regressors
- Derive asymptotic properties of nonparametric regression under cluster sampling
  - key condition $(\max_{g \leq G} n_g)\, h^{d_{\text{ind}}} = O(1)$
- We propose cluster-robust variance estimator and bandwidth selection

Future work (open questions)

- Boundary analysis
- Local polynomial regressions and series regressions
- Cluster bootstrap inference
- Uniform inference

Thank you!

Email: yuya.shimizu [at] wisc.edu

Homepage: https://yshimizu-econ.github.io/

References I

- Alatas, Vivi et al. (2012). "Targeting the poor: evidence from a field experiment in Indonesia". In: *American Economic Review* 102.4, pp. 1206–1240.
- Bhattacharya, Debopam (2005). "Asymptotic inference from multi-stage samples". In: *Journal of Econometrics* 126.1, pp. 145–171.
- Djogbenou, Antoine A, James G MacKinnon, and Morten Ørregaard Nielsen (2019). "Asymptotic theory and wild bootstrap inference with clustered errors". In: *Journal of Econometrics* 212.2, pp. 393–412.
- Fan, Jianqing and Irene Gijbels (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability*. Vol. 66. CRC Press.
- Hansen, Bruce E (2008). "Uniform convergence rates for kernel estimation with dependent data". In: *Econometric Theory* 24.3, pp. 726–748.
- Hansen, Bruce E and Seojeong Lee (2019). "Asymptotic theory for clustered samples". In: *Journal of Econometrics* 210.2, pp. 268–290.

References II

- Hansen, Christian B (2007). "Asymptotic properties of a robust variance matrix estimator for panel data when T is large". In: *Journal of Econometrics* 141.2, pp. 597–620.
- Hu, Pingbo, Xiuyuan Peng, and Xinglin Hu (2024). "Some new asymptotic results for series estimation under clustered dependence". In: *Statistics & Probability Letters*, p. 110156.
- Lee, Jungyoon and Peter M Robinson (2016). "Series estimation under cross-sectional dependence". In: *Journal of Econometrics* 190.1, pp. 1–17.
- Lin, Xihong and Raymond J Carroll (2000). "Nonparametric function estimation for clustered data when the predictor is measured without/with error". In: *Journal of the American Statistical Association* 95.450, pp. 520–534.
- Robinson, Peter M (1983). "Nonparametric estimators for time series". In: *Journal of Time Series Analysis* 4.3, pp. 185–207.
- – (2011). "Asymptotic theory for nonparametric regression with spatial data". In: *Journal of Econometrics* 165.1, pp. 5–19.

References III

- Vogt, Michael (2012). "Nonparametric regression for locally stationary time series". In: *The Annals of Statistics* 40.5, pp. 2601–2633.
- Wang, Naisyin (2003). "Marginal nonparametric kernel regression accounting for within-subject correlation". In: *Biometrika* 90.1, pp. 43–52.

Assumptions for asymptotic normality

Assumption

1. *There exists some* $r \geq 2$ *such that*
   1.1 *for any* $\widetilde{x} = \left(\widetilde{x}^{(\mathrm{ind})\top}, \widetilde{x}^{(\mathrm{cls})\top}\right)^\top \in \mathcal{N}$, $\mathbb{E}\left[|e|^{2r} \mid X = \widetilde{x}\right] \leq \overline{v}^2 < \infty$,
   1.2 *for some constant* $C > 0$, $\frac{\left(\sum_{g=1}^G n_g^r\right)^{1/r}}{n^{1/4}} \leq C < \infty$,
   1.3 *and* $\frac{1}{n^{r/2}h^{dr-d}} = O(1)$.
2. *We also assume* $nh^{d+4} = O(1)$,

$$R_k^d f(x)\sigma^2(x) + \lambda R_k^{d_{\mathrm{cls}}} f_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) \sigma\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) > 0,$$

and $\max_{g \leq G} \frac{n_g^4}{n} \to 0$ *as* $n \to \infty$.

back

Assumptions for uniform convergence

Assumption

For some $s \geq 2$,

$$\mathbb{E}\left|Y_i\right|^s < B_1 < \infty, \tag{9}$$

and

$$\sup_x \mathbb{E}\left[\left|Y_i\right|^s \mid X_i = x\right] f(x) < B_2 < \infty. \tag{10}$$

We also assume that

$$\frac{\left(\max_{g \leq G} n_g\right)^2 \log n}{n^{1-(2/s)} h^d} = O(1), \tag{11}$$

$\delta_n = \inf_{\|x\| \leq c_n} f(x) > 0$, *and* $\delta_n^{-1}\left(\left(\frac{\log n}{n h^d}\right)^{1/2} + h^2\right) = o(1)$.

Assumption

For some $0 < L < \infty$, K *has compact support, that is,* $K(u) = 0$ *for* $\|u\| > L$. *Furthermore,* K *is Lipschitz, i.e., for some* $\Lambda < \infty$ *and for all* $u, u' \in \mathbb{R}$, $|K(u) - K(u')| \leq \Lambda \|u - u'\|$.

back

Sketch of proof for uniform convergence

1. Divide $\frac{1}{nh^d}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right) Y_{gj}$ into the tail $|Y_{gj}| > \tau_n$ and the other part
2. The tail can be bounded by inequalities
3. Control the other part by the following lemma

Lemma (Bernstein's inequality for cluster sampling)

For random variables under cluster sampling $\left\{\{Y_{gj}\}_{j=1}^{n_g}\right\}_{g=1}^{G}$ *with bounded ranges* $[-B, B]$ *and zero means,*

$$\mathbb{P}\left[\left|\widetilde{Y}_1 + \cdots + \widetilde{Y}_G\right| > \varepsilon\right] \leq 2\exp\left\{-\frac{1}{2}\frac{\varepsilon^2}{v + (\max_{g\leq G} n_g)B\varepsilon/3}\right\}$$

for every $\varepsilon > 0$ *and* $v \geq \mathrm{Var}\left(\widetilde{Y}_1 + \cdots + \widetilde{Y}_G\right)$*, where* $\widetilde{Y}_g = \sum_{j=1}^{n_g} Y_{gj}$

back

Optimal bandwidth

- Minimizes asymptotic integrated MSE (IMSE) with some weight $w(x)$

$$\text{IMSE}(h) = h^4 \bar{B} + \frac{R_k^d \bar{\sigma}^2}{nh^d} + \text{constant} + \text{(negligible term)}$$

where

$$\bar{B} = \int_{\mathbb{R}^d} B_{\text{nw}}(x)^2 f(x) w(x) \mathrm{d}x \qquad \text{and} \qquad \bar{\sigma}^2 = \int_{\mathbb{R}^d} \sigma^2(x) w(x) \mathrm{d}x$$

- When $(\max_{g \leq G} n_g)\, n^{-d_{\text{ind}}/(d+4)} \to 0$, optimal bandwidth is standard

$$h_0 = \left(\frac{dR_K \bar{\sigma}^2}{4\bar{B}} \right)^{1/(d+4)} n^{-1/(d+4)} \tag{12}$$

- h_0 does not satisfy $(\max_{g \leq G} n_g)\, h^{d_{\text{ind}}} = O(1)$
  - recommend using Cross-Validation in this case

Cross-Validation

- Issue for standard leave-one-out cross-validation: dependence within clusters
- Leave-one-cluster-out cross-validation is

$$\mathrm{CV}(h) \equiv \frac{1}{n} \sum_{g=1}^{G} \sum_{j=1}^{n_g} \tilde{e}_{gj}(h)^2 w(X_{gj}), \qquad (13)$$

where $\widetilde{e}_{gj}(h) = Y_{gj} - \widetilde{m}_{-g}(X_{gj}, h)$ and $\widetilde{m}_{-g}(x, h)$ is estimated without cluster g

- We show that

$$\mathbb{E}[\mathrm{CV}(h)] = \overline{\sigma}_w^2 + \mathrm{IMSE}_{G-1}(h)$$

where

$$\overline{\sigma}_w^2 = \mathbb{E}\left[e_{gj}^2 w(X_{gj})\right]$$

and

$$\mathrm{IMSE}_{G-1}(h) \equiv \sum_{g=1}^{G} \frac{n_g}{n} \mathbb{E}_{-g}\left[\int_{\mathbb{R}^d} \{m(x) - \widetilde{m}_{-g}(x, h)\}^2 f(x) w(x) \, \mathrm{d}x\right]$$

- We can choose $h_{\text{CR-ROT}} = \arg\min \mathrm{CV}(h)$

Rule of thumb

- **For i.i.d. data**, easy to implement method is proposed by Fan and Gijbels (1996)

$$h_{\mathsf{ROT}} = \left(\frac{d R_K \check{\sigma}^2}{4 \check{B}} \right)^{1/(d+4)} n^{-1/(d+4)},$$

  where $\check{B}$ and $\check{\sigma}^2$ are computed
  - by the 4th order *global* polynomial regression
  - under homoskedastic standard error assumption
- **For cluster sampling**, we propose $h_{\mathsf{CR\text{-}ROT}}$
  - replace $\check{B}$ and $\check{\sigma}^2$ by leave-one-cluster-out estimator

back

Simulation: bandwidth choice

- Compare four methods of bandwidth choice:
  1. [ROT] rule of thumb by Fan and Gijbels (1996)
  2. [CR-ROT] cluster robust rule of thumb
  3. [CV] leave-one-out cross-validation
  4. [CR-CV] leave-one-cluster-out cross-validation
- Evaluated by the average squared error (ASE):

$$\mathrm{ASE}(h) = \frac{1}{n_{\text{grid}}} \sum_{k=1}^{n_{\text{grid}}} \left\{\hat{m}_{\mathrm{nw}}\left(u_k, h\right) - m\left(u_k\right)\right\}^2,$$

  where the grid points $\left\{u_1, \ldots, u_{n_{\text{grid}}}\right\}$ are evenly distributed
  - We set $n_{grid} = 50$

Baseline data generating process

- Number of clusters: $G = 100$
- Cluster sizes
  - $n_g = 20$ for $g = 1, \ldots, G-1$
  - $n_G \in \{20, 100\}$
    - $\left(\max_{g \leq G} n_g\right)/n \approx \{0.02, 0.09\}$
- Generated 500 datasets from Setup 1 or 2
- **Setup 1 (homoskedastic errors):**

$$Y_{gj} = \sin\left(2X_{gj}\right) + 2\exp\left(-16X_{gj}^2\right) + 0.5e_{gj}, \tag{14}$$

  or

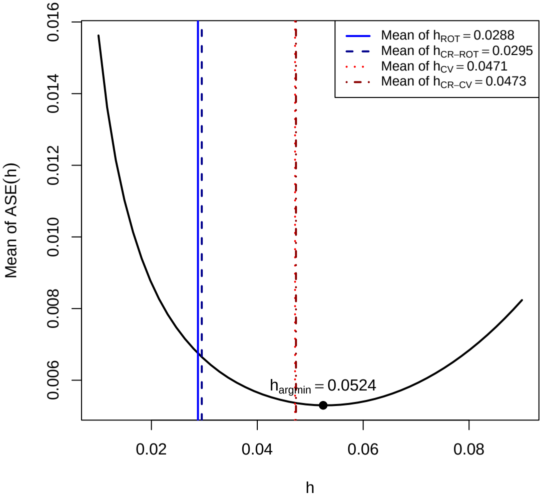
  Setup 2 (heteroskedastic errors):

$$Y_{gj} = X_{gj}\sin\left(2\pi X_{gj}\right) + \sigma\left(X_{gj}\right)e_{gj}, \tag{15}$$

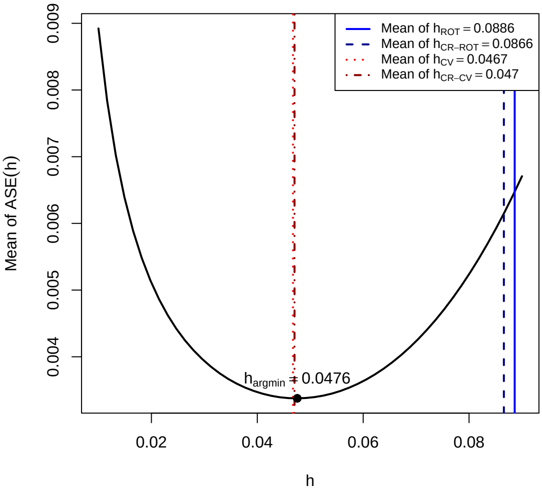
$$\sigma\left(X_{gj}\right) = \frac{2 + \cos\left(2\pi X_{gj}\right)}{5},$$

- where $X_{gj} = \sqrt{\rho_x}\left(X_1\right)_g + \sqrt{1-\rho_x}\left(X_2\right)_{gj}$ and $e_{gj} = \sqrt{\rho_e}c_g + \sqrt{1-\rho_e}u_{gj}$
  - $\left(X_1\right)_g \sim \mathcal{N}(0,1)$, $\left(X_2\right)_{gj} \sim \mathcal{N}(0,1)$, $c_g \sim \mathcal{N}(0,1)$, and $u_{gj} \sim \mathcal{N}(0,1)$ independently

Mean of $\mathrm{ASE}(h)$ for m_{nw} in Setup 1 with $\max_{g \leq G} n_g = 100$ and $\rho_X = \rho_e = 0.5$



Mean of $\mathrm{ASE}(h)$ for m_{nw} in Setup 2 with $\max_{g \leq G} n_g = 100$ and $\rho_X = \rho_e = 0.5$



Coverage and average length of 95% CI for each standard error (m_{nw}, Setup 1, $x = 0.75$, with bias)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.918	0.925	0.953	0.907	0.914	0.948
	{0.192}	{0.195}	{0.217}	{0.189}	{0.192}	{0.217}
(ρ_X, ρ_e)=(0.2,0.5)	0.880	0.888	0.956	0.861	0.868	0.949
	{0.192}	{0.195}	{0.246}	{0.188}	{0.192}	{0.250}
(ρ_X, ρ_e)=(0.5,0.2)	0.916	0.921	0.956	0.906	0.910	0.951
	{0.191}	{0.194}	{0.227}	{0.188}	{0.191}	{0.228}
(ρ_X, ρ_e)=(0.5,0.5)	0.859	0.865	0.960	0.837	0.845	0.955
	{0.191}	{0.195}	{0.261}	{0.188}	{0.191}	{0.267}

Coverage and average length of 95% CI for each standard error (m_{nw}, Setup 2, $x = 0.8$, with bias)

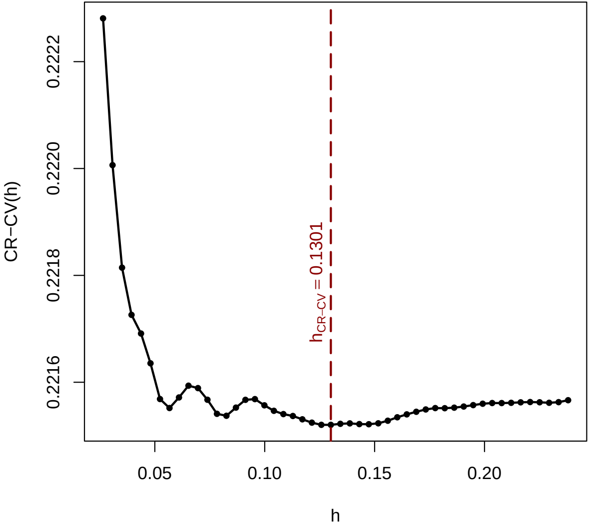
	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.772	0.783	0.821	0.782	0.791	0.840
	{0.167}	{0.170}	{0.187}	{0.165}	{0.168}	{0.187}
(ρ_X, ρ_e)=(0.2,0.5)	0.737	0.745	0.842	0.734	0.743	0.852
	{0.167}	{0.171}	{0.209}	{0.164}	{0.168}	{0.211}
(ρ_X, ρ_e)=(0.5,0.2)	0.748	0.756	0.819	0.752	0.758	0.829
	{0.166}	{0.170}	{0.191}	{0.164}	{0.167}	{0.192}
(ρ_X, ρ_e)=(0.5,0.5)	0.701	0.714	0.846	0.707	0.718	0.853
	{0.166}	{0.170}	{0.218}	{0.164}	{0.167}	{0.223}

back

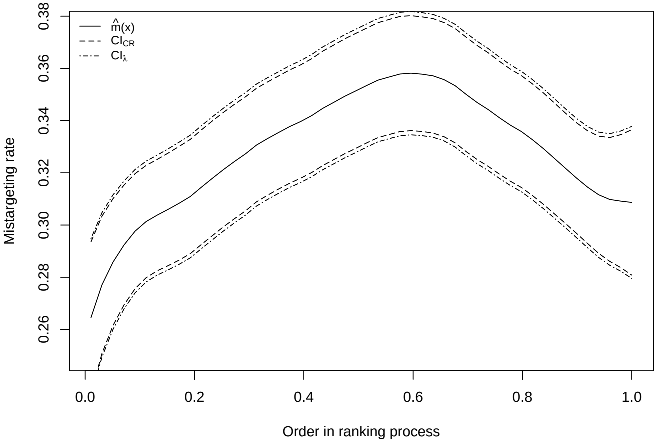
Empirical illustration

- Poverty targeting dataset from Alatas et al. (2012)
- Human errors could happen during poverty ranking process in villages
- Alatas et al. (2012) investigated this concern by running a nonparametric regression
  - the mistarget rate (Y_{gj}) on the order in the ranking process (X_{gj})
- $n = 3784$ observations, $G = 431$ villages, and each village has $n_g \in [4, 9]$

Cluster-robust cross-validation function $\mathrm{CV}(h)$



Local linear estimation and 95% CIs on Alatas et al. (2012)'s dataset



back