Statistical inference for large-scale multi-source heterogeneous data

Jiuzhou Miao

School of Statistics and Mathematics, Zhejiang Gongshang University

ICSA China 2024

June 29, 2024

Background

- In the era of big data, people are confronted with data that may not only be large-scale, but also heterogeneous.
- Large-scale data: The sample size or dimension of a set of data is at least one particularly large.
- Heterogeneous data: A set of data can come from multiple different sources.
- This is undoubtedly a huge challenge for traditional statistical methods and computer performance.

Example: Beijing multi-site air-quality data

- This dataset includes air pollutant data from K different air quality monitoring stations in Beijing. For jth station, the data observation size is denoted by n_j, and the whole data observation size is $N = n_1 + \cdots + n_K$.
- X_{ij}: Climatic state of interest from jth station data.
- Y_{ij}: the concentrations of air pollutants from jth station data.
- $(X_{ij}, Y_{ij})_{i,j=1}^{n_j,K}$ satifies

$$Y_{ij} = m_j\left(X_{ij}\right) + \sigma_j\left(X_{ij}\right)\varepsilon_{ij} \tag{1}$$

- $m_j(x)$ and $\sigma_j(x)$ are mean function and standard error function of jth station data.

Background

- We will investigate whether X_{ij} has effects on Y_{ij} and make statistical inferences on the overall trend of the influence relationship.
- Solving these two problems is equivalent to examine the form of the population mean function (say $m(x)$) between X_{ij} and Y_{ij} and make statistical inference on their overall trends.
- Drawing on the methods on estimating population mean function of stratified sampling, assume that $m\left(x\right)=\sum_{j=1}^{K}\omega_j m_j\left(x\right)$.
- Here, ω_j are weights given in advance, satifies $\sum_{j=1}^{K}\omega_j=1$.
- e.g., $\omega_j=n_j/N$ or $\omega_j=1/K$.

Background

- Especially, when $\omega_j = 1/K$ and data (X_{ij}, Y_{ij}) are homogeneous, i.e. $m_1(x) = \cdots = m_K(x) = m(x)$ and $\sigma_1^2(x) = \cdots = \sigma_K^2(x) = \sigma^2(x)$.
- For the mean function of large-scale homogeneous data, in order to solve the problems of limited computing and storage capacity and long computing time of computers, divide and conquer technology is usually used to deal with it.
- Specifically,
  1. Divide the dataset randomly into several blocks so that the single block data can be quickly calculated on one computer.
  2. The mean function $m_j(x)$ is calculated by jth computer and uploaded to the central computer.
  3. The population mean function $m(x)$ is estimated on central computer by simple linear averaging.

Our contributions

- Drawing on the idea of divide and conquer techniques, a weighted local linear estimation method is proposed for the overall mean function of multi-source heterogeneity data.
- The asymptotic pointwise confidence interval and simultaneous confidence band (SCB) are constructed for the mean function by studying the pointwise convergence properties and the extreme value distribution properties.
- By using the proposed SCB, one can make statistical inference on the overall trend of the mean function $m(x)$.

Review on divide and conquer techniques

- Chang et al. (2017) proposed kernel estimation and KNN estimation for mean function, and study the convergence rate and bandwidth selection.
- Zhang et al. (2015) studied the MSE bound of mean function estimation by using kernel-ridge Regression.
- Wang et al. (2019) studied estimation and testing problems on additive partially linear models.
- Chen and Lin (2022) studied composite quantile Regression.

Review on SCBs

- Density functions: Bickel and Rosenblatt (1973).
- Mean functions: Härdle (1989), Xia (1998), Eubank and Speckman (1993), Wang and Yang (2009).
- Variance functions: Song and Yang (2009), Cai and Yang (2015), Cai et al. (2019).
- Semi-parametric models: Gu et al. (2014), Cao et al. (2018), Gu and Yang (2015).
- Functional data: Degras (2011), Ma et al. (2012), Cao et al. (2012), Zheng et al. (2014), Cao et al. (2016).
- Functional time series: Li and Yang (2023), Zhong and Yang (2023).

Main results

- Consider model (1), the numbers of sub-populations K can tend to infinity. Without loss of generality, assume that each sub-population can be stored and computed on one computer.
- Local linear regression is applied to smooth the jth block data, i.e. solving

$$\arg\min_{\alpha_{0,j},\alpha_{1,j}\in\mathbf{R}} n_j^{-1}\sum_{i=1}^{n_j}\left\{Y_{ij}-\alpha_{0,j}-\alpha_{1,j}\left(X_{ij}-x\right)\right\}^2 G_h\left(X_{ij}-x\right).$$

  and setting $\hat{m}_j\left(x\right)=\hat{\alpha}_{0,j}$.
- Then the mean function $m(\cdot)$ can be estimated by $\hat{m}_{\mathrm{DC}}\left(x\right)=\sum_{j=1}^{K}\omega_j\hat{m}_j\left(x\right)$.

Assumptions

(A1) For $j = 1, \ldots, K$, $m_j \in C^2[a,b]$, $\sigma_j \in C^1[a,b]$, $f_j \in C^1[a,b]$. There exists constants $c_f, C_f, c_\sigma, C_\sigma > 0$ such that $c_f < f_j < C_f$ and $c_\sigma < \sigma_j < C_\sigma$.

(A2) Kernel function $G \in C^1[-1,1]$ is a symmetrical p.d.f. whose support is $[-1,1]$.

(A3) For $j = 1, \ldots, K$, $\omega_j \sim K^{-1}$ and $n_j \sim N/K$. There exists $0 \le \theta < 2/5$ such that $K \sim N^\theta$.

(A4) There exists $r > 2/\beta, 0 < \beta < \min\{2/5 - \theta, (1-\theta)/2\}$ such that $\mathrm{E}|\varepsilon_{11}|^r < \infty$.

(A5) As $N \to \infty$, the bandwidth h satisfies
$$\max\left\{N^{2\beta+2\theta-1}, N^{(\theta-1)/3}\log^{1/3} N\right\} \ll h \ll N^{-1/5}.$$

Asymptotic properties

- Denote $V_N\left(x,x'\right)=\sum_{j=1}^{K}\omega_j^2 n_j^{-1}\sigma_j^2\left(x\right)f_j^{-1}\left(x'\right)$,

$$A_N\left(x,x'\right)=V_N\left(x,x'\right)V_N^{-1/2}\left(x,x\right)V_N^{-1/2}\left(x',x'\right),$$
$$\Gamma_N\left(x,x'\right)=\left(2\tau_0\right)^{-1}\left\{\tau_h\left(x-x'\right)A_N\left(x,x'\right)+\tau_h\left(x'-x\right)A_N\left(x',x\right)\right\}.$$

- Let $\zeta_N(x)$ be a Gaussian process with mean function 0 and covariance function $\Gamma_N(x,x')$ and $Q_{1-\alpha}$ be the $100(1-\alpha)\%$ quantile of the distribution of $\sup_{x\in[a_0,b_0]}\left|\zeta_N(x)\right|$.

Theorem 1

Under Assumptions (A1)–(A5), as $N\to\infty$, one has that

$$P\left\{h^{1/2}\tau_0^{-1/2}\left|\left\{\hat{m}_{DC}\left(x\right)-m\left(x\right)\right\}V_N^{-1/2}\left(x,x\right)\right|\le z_{1-\alpha/2}\right\}\to 1-\alpha,$$
$$P\left\{h^{1/2}\tau_0^{-1/2}\sup_{x\in[a_0,b_0]}\left|\left\{\hat{m}_{DC}\left(x\right)-m\left(x\right)\right\}V_N^{-1/2}\left(x,x\right)\right|\le Q_{1-\alpha}\right\}\to 1-\alpha.$$

Asymptotic properties

- To construct feasible pointwise confidence intervals and SCBs, one needs to estimate the unknown $V_N(x,x)$ and $Q_{1-\alpha}$.
- The pilot kernel density estimator

$$\hat{f}_j(x) = n_j^{-1} \sum_{i=1}^{n_j} G_{\bar{h}_j}(X_{ij} - x)$$

will be used to estimate the probability function $f_j(x)$ in jth block, in which $\bar{h}_j \sim n_1^{-1/5}$. By Theorem 3.1 in Bickel and Rosenblatt (1973), one has that

$$\sup_{x \in [a_0, b_0]} \left| \hat{f}_j(x) - f_j(x) \right| = o_p\left(\bar{h}_j\right) = o_p\left(n_j^{-1/5}\right).$$

Asymptotic properties

- For unknown $\sigma_j^2(x)$, use the Spline–Kernel two step estimation proposed in Cai and Yang (2015), say the estimator for $\sigma_j^2(x)$ as $\hat{\sigma}_j^2(x)$. One has that

$$\sup_{x\in[a_0,b_0]}\left|\hat{\sigma}_j^2(x)-\sigma_j^2(x)\right|=\mathcal{O}_p\left(n_j^{-1/2}\tilde{h}_j^{-1/2}\log^{1/2}n_j\right),$$

  in which $\tilde{h}_j\sim n_1^{-1/5}\log^{-1/2}n_1$.

- Denote

$$\hat{V}_N(x,x')=\sum_{j=1}^{K}\omega_j^2 n_j^{-1}\hat{\sigma}_j^2(x)\hat{f}_j^{-1}(x'),$$
$$\hat{A}_N(x,x')=\hat{V}_N(x,x')\hat{V}_N^{-1/2}(x,x)\hat{V}_N^{-1/2}(x',x'),$$
$$\hat{\Gamma}_N(x,x')=(2\tau_0)^{-1}\left\{\tau_h(x-x')\hat{A}_N(x,x')+\tau_h(x'-x)\hat{A}_N(x',x)\right\}.$$

Asymptotic properties

Proposition 1

Under Assumptions (A1)–(A5), as $N \to \infty$, for $j = 1, \ldots, K$, $\bar{h}_j \sim n_1^{-1/5}$ and $\tilde{h}_j \sim n_1^{-1/5} \log^{-1/2} n_1$, one has that

$$\sup_{x,x' \in [a_0, b_0]} \left| \hat{A}_N(x, x') - A_N(x, x') \right| + \sup_{x \in [a_0, b_0]} \left| \hat{V}_N(x, x) / V_N(x, x) - 1 \right| = o_p(1).$$

- To estimate $Q_{1-\alpha}$, use the parametric Bootstrap methods in Cai and Wang (2021). Denote $\hat{\zeta}_N(x)$ as a Gaussian process with zero mean function and covariance function $\hat{\Gamma}_N(x, x')$.
- Devide $[a_0, b_0]$ equally with $a_0 = x_1 \le x_2 \le \cdots \le x_{400} \le x_{401} = b_0$, and compute covariance matrix $(\hat{\Gamma}_N(x_i, x_j))_{401 \times 401}$.
- With 2000 replications, generate muti-Gaussian random vectors with mean vector $\mathbf{0}_{401 \times 1}$ and covariance matrix $(\hat{\Gamma}_N(x_i, x_j))_{401 \times 401}$.
- Then one can obtain the estimation of $Q_{1-\alpha}$, say $\hat{Q}_{1-\alpha}$.

Asymptotic properties

Theorem 2

Under assumptions of Proposition 1, as $N \to \infty$, for all $x \in [a_0, b_0]$, an asymptotic $100\,(1-\alpha)\,\%$ pointwise confidence interval for $m\,(x)$ is

$$\hat{m}_{DC}\,(x) \pm h^{-1/2} \tau_0^{1/2} \hat{V}_N^{1/2}\,(x, x)\, z_{1-\alpha/2},$$

and an asymptotic $100\,(1-\alpha)\,\%$ SCB for $m\,(x)$ is

$$\hat{m}_{DC}\,(x) \pm h^{-1/2} \tau_0^{1/2} \hat{V}_N^{1/2}\,(x, x)\, \hat{Q}_{1-\alpha}.$$

Implementation

- The domain of the mean function $m_j(x)$ is set as $\left[\hat{a}, \hat{b}\right]$, in which $\hat{a}$ and $\hat{b}$ are the mininmal value and maximal value of $(X_{ij})_{i,j=1}^{n_j,K}$.
- To avoid boundary effect, take the 3%quantile and 97% quantile of $(X_{ij})_{i,j=1}^{n_j,K}$ as the endpoint value of $[a_0, b_0]$.
- Biweight kernel function $G(v) = 15\left(1 - v^2\right)^2 \mathrm{I}_{[-1,1]}(v)/16$.
- $\omega_j = n_j/N$, $K \sim \lfloor N^{1/5} \rfloor$.
- The bandwidth for local linear regression is set by $h = N^{-1/5}\log^{-1/2} N$.
- The bandwidth for density estimation is recommended to use equation (3.31) in Silverman (1986).
- The bandwidth for $\hat{\sigma}_j^2(x)$ is set by $\tilde{h}_j = 4h_{\mathrm{rot},j}\log^{-1/2} n_j$, in which $h_{\mathrm{rot},j}$ is dicided by equation (4.3) in Fan and Gijbels (1996). The knot number is decided by BIC rules, see Cai and Yang (2015) for more details.

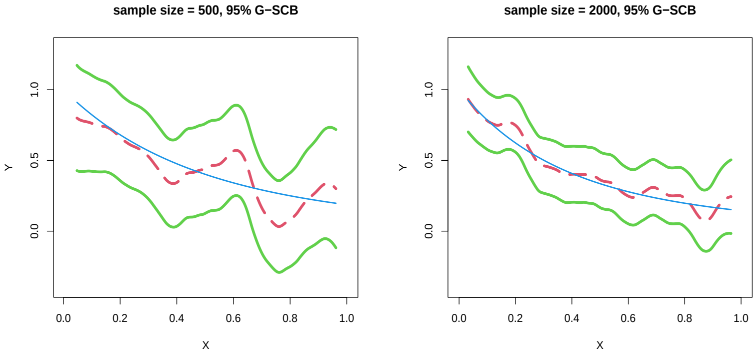
Simulation

- $X_{ij} \overset{i.i.d.}{\sim} U[0,1]$
- $m_j(x) = \exp(-jx)$, $\sigma_j(x) = \{\exp(jx) - 0.1\} / \{\exp(jx) + 0.1\}$
- $\varepsilon_{ij} \overset{i.i.d.}{\sim} N(0,1)$ or $\varepsilon_{ij} \overset{i.i.d.}{\sim} \sqrt{0.8}t_{10}$
- $N = 500, 1000, 2000, 6000, 10000$
- The jth block sample size: $n_j = \lfloor N/K \rfloor + n_j^{\text{rest}}$.
- $\left\{n_j^{\text{rest}}\right\}_{j=1}^{K-1}$samples from $\left\{-\lfloor N^{1/3} \rfloor - 1, \ldots, \lfloor N^{1/3} \rfloor + 1\right\}$.
- Set $n_K^{\text{rest}} = N - K \times \lfloor N/K \rfloor - \sum_{j=1}^{K-1} n_j^{\text{rest}}$.

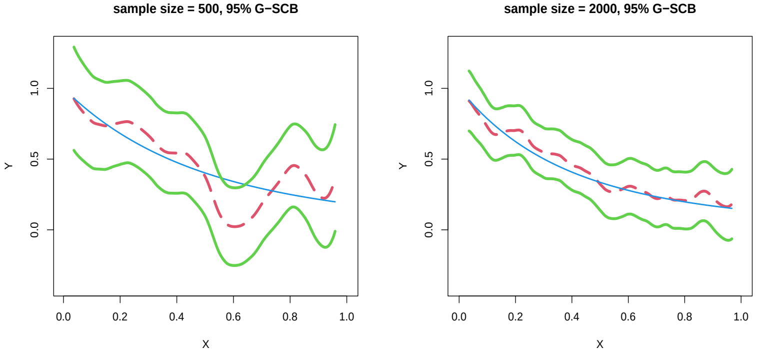
Simulation

n	$1-\alpha$	(i): $\varepsilon_{ij} \sim N(0,1)$		(ii): $\varepsilon_{ij} \sim \sqrt{0.8}t_{10}$	
		G-SCB	N-SCB	G-SCB	N-SCB
500	0.95	0.901 (0.624)	0.983 (0.795)	0.898 (0.625)	0.989 (0.796)
	0.99	0.969 (0.726)	0.994 (0.874)	0.966 (0.728)	0.996 (0.874)
1000	0.95	0.927 (0.500)	0.990 (0.624)	0.942 (0.500)	0.994 (0.624)
	0.99	0.975 (0.578)	0.998 (0.686)	0.985 (0.578)	0.999 (0.686)
2000	0.95	0.943 (0.400)	0.997 (0.489)	0.944 (0.400)	0.994 (0.490)
	0.99	0.993 (0.461)	0.999 (0.538)	0.984 (0.461)	0.997 (0.539)
6000	0.95	0.954 (0.276)	0.994 (0.329)	0.945 (0.275)	0.993 (0.329)
	0.99	0.987 (0.315)	0.999 (0.361)	0.989 (0.315)	0.997 (0.361)
10000	0.95	0.950 (0.232)	0.993 (0.273)	0.951 (0.232)	0.997 (0.273)
	0.99	0.989 (0.265)	0.999 (0.300)	0.993 (0.265)	0.999 (0.300)

The figures of 95% G-SCBs ($\varepsilon_{ij} \sim N(0,1)$)



The figures of 95% G-SCBs ($\varepsilon_{ij} \sim \sqrt{0.8}t_{10}$)

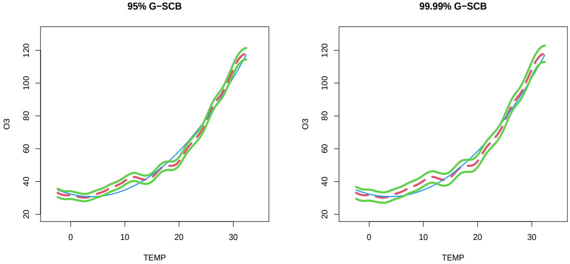


Beijing multi-site air-quality data

- The proposed methods will be applied on Beijing multi-site air-quality data. ($K = 12$)
- X_{ij}^{origin}: The temperature at 11 hour per day from jth monitoring station data.
- Y_{ij}: The ozone concentration at 11 hour per day from jth monitoring station data.
- In order to reduce data variance fluctuations, make a linear transformation $X_{ij} = \left(X_{ij}^{\text{origin}} - X_{\min}\right) / \left(X_{\max} - X_{\min}\right)$, in which $X_{\min} = -15.8$, $X_{\max} = 39.8$.
- For the data from jth station, use model (1) to smooth data (X_{ij}, Y_{ij}), that is $Y_{ij} = m_j\left(X_{ij}\right) + \sigma_j\left(X_{ij}\right)\varepsilon_{ij}, j = 1, \ldots, 12$.
- Consider the following hypothesis testing problem

$$H_0 : m(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \quad \text{vs.} \quad H_1 : m(x) \neq \beta_0 + \beta_1 x + \beta_2 x^2.$$

Beijing multi-site air-quality data



- The p-value is much less than 0.0001 and H_0 is rejected.
- Air temperature has a positive effect on ozone concentration, which increases with the increase of air temperature, but the relationship between the two cannot be described by quadratic functions.

Conclusions

- In this study, a nonparametric local linear estimation method for the mean function of large-scale multi-source heterogeneous data is established, which overcomes the limitations of the existing researches mainly for homogeneous data and parametric regression methods.
- The proposed methods solve the limitation that the existing research mainly examines the estimation convergence rate and cannot make statistical inference on the unknown information of the data.
- The proposed methods are also applicable to the divide and conquer local linear estimation of mean functions under homogeneous big data, which is a further extension and expansion of the methods on analyzing homogeneous big data.

Thank You