Semi-Nonparametric Inference for Massive Data

Guang Cheng

Department of Statistics
Purdue University

Statistics Seminar at University of Florida
October 30, 2014
A joint work with T. Zhao and H. Liu at Princeton

Challenges of Big Data

- The massive sample size and high dimensionality of Big Data introduce unique computational and statistical challenges such as
  - scalability and storage bottleneck;
  - dynamical underlying distributions;
  - heavy computational cost;
  - heterogeneous subpopulations.
- See more in *Challenges of Big Data Analysis* by Fan et al in National Science Review (2014).

Challenges of Big Data

- The massive sample size and high dimensionality of Big Data introduce unique computational and statistical challenges such as
  - scalability and storage bottleneck;
  - dynamical underlying distributions;
  - heavy computational cost;
  - heterogeneous subpopulations.
- See more in *Challenges of Big Data Analysis* by Fan et al in National Science Review (2014).

Challenges of Big Data

- The massive sample size and high dimensionality of Big Data introduce unique computational and statistical challenges such as
  - scalability and storage bottleneck;
  - dynamical underlying distributions;
  - heavy computational cost;
  - heterogeneous subpopulations.
- See more in *Challenges of Big Data Analysis* by Fan et al in National Science Review (2014).

Challenges of Big Data

- The massive sample size and high dimensionality of Big Data introduce unique computational and statistical challenges such as
  - scalability and storage bottleneck;
  - dynamical underlying distributions;
  - heavy computational cost;
  - heterogeneous subpopulations.
- See more in *Challenges of Big Data Analysis* by Fan et al in National Science Review (2014).

Challenges of Big Data

- The massive sample size and high dimensionality of Big Data introduce unique computational and statistical challenges such as
  - scalability and storage bottleneck;
  - dynamical underlying distributions;
  - heavy computational cost;
  - heterogeneous subpopulations.
- See more in *Challenges of Big Data Analysis* by Fan et al in National Science Review (2014).

Challenges of Big Data

- The massive sample size and high dimensionality of Big Data introduce unique computational and statistical challenges such as
  - scalability and storage bottleneck;
  - dynamical underlying distributions;
  - heavy computational cost;
  - heterogeneous subpopulations.
- See more in *Challenges of Big Data Analysis* by Fan et al in National Science Review (2014).

General Goal

- In the era of massive data, I am curious about the following questions:
  - what is the least computational cost in obtaining the best possible statistical inferences?
  - how to efficiently extract common features across all sub-populations in presence of heterogeneity?
  - how to boost the efficiency of heterogeneity estimation by taking advantage of commonality information?
- More subtle (technical) questions include:
  - the impact of model regularity on the computational cost;
  - the optimal choice of smoothing parameter;
  - heterogeneity testing.
- *Oracle rule* for massive data is the key to these questions.

General Goal

- In the era of massive data, I am curious about the following questions:
  - what is the least computational cost in obtaining the best possible statistical inferences?
  - how to efficiently extract common features across all sub-populations in presence of heterogeneity?
  - how to boost the efficiency of heterogeneity estimation by taking advantage of commonality information?
- More subtle (technical) questions include:
  - the impact of model regularity on the computational cost;
  - the optimal choice of smoothing parameter;
  - heterogeneity testing.
- *Oracle rule* for massive data is the key to these questions.

General Goal

- In the era of massive data, I am curious about the following questions:
  - what is the least computational cost in obtaining the best possible statistical inferences?
  - how to efficiently extract common features across all sub-populations in presence of heterogeneity?
  - how to boost the efficiency of heterogeneity estimation by taking advantage of commonality information?
- More subtle (technical) questions include:
  - the impact of model regularity on the computational cost;
  - the optimal choice of smoothing parameter;
  - heterogeneity testing.
- *Oracle rule* for massive data is the key to these questions.

General Goal

- In the era of massive data, I am curious about the following questions:
  - what is the least computational cost in obtaining the best possible statistical inferences?
  - how to efficiently extract common features across all sub-populations in presence of heterogeneity?
  - how to boost the efficiency of heterogeneity estimation by taking advantage of commonality information?
- More subtle (technical) questions include:
  - the impact of model regularity on the computational cost;
  - the optimal choice of smoothing parameter;
  - heterogeneity testing.
- *Oracle rule* for massive data is the key to these questions.

General Goal

- In the era of massive data, I am curious about the following questions:
  - what is the least computational cost in obtaining the best possible statistical inferences?
  - how to efficiently extract common features across all sub-populations in presence of heterogeneity?
  - how to boost the efficiency of heterogeneity estimation by taking advantage of commonality information?
- More subtle (technical) questions include:
  - the impact of model regularity on the computational cost;
  - the optimal choice of smoothing parameter;
  - heterogeneity testing.
- *Oracle rule* for massive data is the key to these questions.

General Goal

- In the era of massive data, I am curious about the following questions:
  - what is the least computational cost in obtaining the best possible statistical inferences?
  - how to efficiently extract common features across all sub-populations in presence of heterogeneity?
  - how to boost the efficiency of heterogeneity estimation by taking advantage of commonality information?
- More subtle (technical) questions include:
  - the impact of model regularity on the computational cost;
  - the optimal choice of smoothing parameter;
  - heterogeneity testing.
- *Oracle rule* for massive data is the key to these questions.

General Goal

- In the era of massive data, I am curious about the following questions:
  - what is the least computational cost in obtaining the best possible statistical inferences?
  - how to efficiently extract common features across all sub-populations in presence of heterogeneity?
  - how to boost the efficiency of heterogeneity estimation by taking advantage of commonality information?
- More subtle (technical) questions include:
  - the impact of model regularity on the computational cost;
  - the optimal choice of smoothing parameter;
  - heterogeneity testing.
- *Oracle rule* for massive data is the key to these questions.

General Goal

- In the era of massive data, I am curious about the following questions:
  - what is the least computational cost in obtaining the best possible statistical inferences?
  - how to efficiently extract common features across all sub-populations in presence of heterogeneity?
  - how to boost the efficiency of heterogeneity estimation by taking advantage of commonality information?
- More subtle (technical) questions include:
  - the impact of model regularity on the computational cost;
  - the optimal choice of smoothing parameter;
  - heterogeneity testing.
- *Oracle rule* for massive data is the key to these questions.

General Goal

- In the era of massive data, I am curious about the following questions:
  - what is the least computational cost in obtaining the best possible statistical inferences?
  - how to efficiently extract common features across all sub-populations in presence of heterogeneity?
  - how to boost the efficiency of heterogeneity estimation by taking advantage of commonality information?
- More subtle (technical) questions include:
  - the impact of model regularity on the computational cost;
  - the optimal choice of smoothing parameter;
  - heterogeneity testing.
- *Oracle rule* for massive data is the key to these questions.

Part I: Homogeneous Data

Outline

Divide-and-Conquer Approach

- Consider a nonparametric regression model:

$$Y = f(Z) + \epsilon;$$

- Entire Dataset (iid data):

$$X_1, X_2, \ldots, X_N, \text{ for } X = (Y, Z);$$

- *Randomly* split dataset into s subsamples (with equal sample size $n = N/s$): $P_1, \ldots, P_s$;
- Perform nonparametric estimating in each subsample:

$$P_j = \{X_1^{(j)}, \ldots, X_n^{(j)}\} \Longrightarrow \widehat{f}_n^{(j)};$$

- Aggregation: $\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)}$;

Divide-and-Conquer Approach

- Consider a nonparametric regression model:

$$Y = f(Z) + \epsilon;$$

- Entire Dataset (iid data):

$$X_1, X_2, \ldots, X_N, \text{ for } X = (Y, Z);$$

- *Randomly* split dataset into s subsamples (with equal sample size $n = N/s$): $P_1, \ldots, P_s$;
- Perform nonparametric estimating in each subsample:

$$P_j = \{X_1^{(j)}, \ldots, X_n^{(j)}\} \Longrightarrow \widehat{f}_n^{(j)};$$

- Aggregation: $\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)}$;

Divide-and-Conquer Approach

- Consider a nonparametric regression model:

$$Y = f(Z) + \epsilon;$$

- Entire Dataset (iid data):

$$X_1, X_2, \ldots, X_N, \text{ for } X = (Y, Z);$$

- *Randomly* split dataset into s subsamples (with equal sample size $n = N/s$): $P_1, \ldots, P_s$;
- Perform nonparametric estimating in each subsample:

$$P_j = \{X_1^{(j)}, \ldots, X_n^{(j)}\} \Longrightarrow \widehat{f}_n^{(j)};$$

- Aggregation: $\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)}$;

Divide-and-Conquer Approach

- Consider a nonparametric regression model:

$$Y = f(Z) + \epsilon;$$

- Entire Dataset (iid data):

$$X_1, X_2, \ldots, X_N, \text{ for } X = (Y, Z);$$

- *Randomly* split dataset into s subsamples (with equal sample size $n = N/s$): $P_1, \ldots, P_s$;
- Perform nonparametric estimating in each subsample:

$$P_j = \{X_1^{(j)}, \ldots, X_n^{(j)}\} \Longrightarrow \widehat{f}_n^{(j)};$$

- Aggregation: $\bar{f}_N = (1/s)\sum_{j=1}^{s} \widehat{f}_n^{(j)}$;

Divide-and-Conquer Approach

- Consider a nonparametric regression model:

$$Y = f(Z) + \epsilon;$$

- Entire Dataset (iid data):

$$X_1, X_2, \ldots, X_N, \text{ for } X = (Y, Z);$$

- *Randomly* split dataset into s subsamples (with equal sample size $n = N/s$): $P_1, \ldots, P_s$;
- Perform nonparametric estimating in each subsample:

$$P_j = \{X_1^{(j)}, \ldots, X_n^{(j)}\} \Longrightarrow \widehat{f}_n^{(j)};$$

- Aggregation: $\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)};$

A Few Comments

- As far as we are aware, the *statistical studies* of the D&C method focus on either parametric inferences, e.g., Bootstrap (Kleiner et al, 2014) and Bayesian (Wang and Dunson, http://arxiv.org/abs/1312.4605), or nonparametric minimaxity (Zhang et al, 2014). Other relevant work includes high dimensional linear models with variable selection (Chen and Xie, 2012);
- Semi/nonparametric inference for massive data still remains untouched;
- For homogeneous data, we want to prove a *Free Lunch Theorem*: significantly reduce computational cost without sacrificing any inferential accuracy (oracle rule).

A Few Comments

- As far as we are aware, the *statistical studies* of the D&C method focus on either parametric inferences, e.g., Bootstrap (Kleiner et al, 2014) and Bayesian (Wang and Dunson, http://arxiv.org/abs/1312.4605), or nonparametric minimaxity (Zhang et al, 2014). Other relevant work includes high dimensional linear models with variable selection (Chen and Xie, 2012);
- Semi/nonparametric inference for massive data still remains untouched;
- For homogeneous data, we want to prove a *Free Lunch Theorem*: significantly reduce computational cost without sacrificing any inferential accuracy (oracle rule).

A Few Comments

- As far as we are aware, the *statistical studies* of the D&C method focus on either parametric inferences, e.g., Bootstrap (Kleiner et al, 2014) and Bayesian (Wang and Dunson, http://arxiv.org/abs/1312.4605), or nonparametric minimaxity (Zhang et al, 2014). Other relevant work includes high dimensional linear models with variable selection (Chen and Xie, 2012);
- Semi/nonparametric inference for massive data still remains untouched;
- For homogeneous data, we want to prove a *Free Lunch Theorem*: significantly reduce computational cost without sacrificing any inferential accuracy (oracle rule).

Splitotics Theory ($s \to \infty$ as $N \to \infty$)

- Specifically, we want to derive the largest possible diverging rate of s under which the following oracle rule holds: "*the nonparametric inferences constructed based on $\bar{f}_N$ are (asymp.) the same as those on the oracle estimator $\widehat{f}_N$.*"
- Meanwhile, we want to know
  - how to choose the smoothing parameter in each sub-sample;
  - how the smoothness of f_0 affects the rate of s.
- Allowing $s \to \infty$ significantly complicates our theoretical analysis.

Splitotics Theory ($s \to \infty$ as $N \to \infty$)

- Specifically, we want to derive the largest possible diverging rate of s under which the following oracle rule holds: "*the nonparametric inferences constructed based on $\bar{f}_N$ are (asymp.) the same as those on the oracle estimator $\widehat{f}_N$.*"
- Meanwhile, we want to know
  - how to choose the smoothing parameter in each sub-sample;
  - how the smoothness of f_0 affects the rate of s.
- Allowing $s \to \infty$ significantly complicates our theoretical analysis.

Splitotics Theory ($s \to \infty$ as $N \to \infty$)

- Specifically, we want to derive the largest possible diverging rate of s under which the following oracle rule holds: "*the nonparametric inferences constructed based on $\bar{f}_N$ are (asymp.) the same as those on the oracle estimator $\widehat{f}_N$.*"
- Meanwhile, we want to know
    - how to choose the smoothing parameter in each sub-sample;
    - how the smoothness of f_0 affects the rate of s.
- Allowing $s \to \infty$ significantly complicates our theoretical analysis.

Splitotics Theory ($s \to \infty$ as $N \to \infty$)

- Specifically, we want to derive the largest possible diverging rate of s under which the following oracle rule holds: "*the nonparametric inferences constructed based on $\bar{f}_N$ are (asymp.) the same as those on the oracle estimator $\widehat{f}_N$.*"
- Meanwhile, we want to know
  - how to choose the smoothing parameter in each sub-sample;
  - how the smoothness of f_0 affects the rate of s.
- Allowing $s \to \infty$ significantly complicates our theoretical analysis.

Splitotics Theory ($s \to \infty$ as $N \to \infty$)

- Specifically, we want to derive the largest possible diverging rate of s under which the following oracle rule holds: "*the nonparametric inferences constructed based on $\bar{f}_N$ are (asymp.) the same as those on the oracle estimator $\widehat{f}_N$.*"
- Meanwhile, we want to know
  - how to choose the smoothing parameter in each sub-sample;
  - how the smoothness of f_0 affects the rate of s.
- Allowing $s \to \infty$ significantly complicates our theoretical analysis.

Kernel Ridge Regression (KRR)

- Define the KRR estimate $\widehat{f}: \mathbb{R}^1 \mapsto \mathbb{R}^1$ as

$$\widehat{f}_n = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(Z_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\},$$

  where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with a kernel $K(z, z') = \sum_{i=1}^{\infty} \mu_i \phi_i(z) \phi_i(z')$. Here, μ_i's are eigenvalues and $\phi_i(\cdot)$'s are eigenfunctions.
- Explicitly, $\widehat{f}_n(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ with $\boldsymbol{\alpha} = (K + \lambda I)^{-1} \boldsymbol{y}$.
- Smoothing spline is a special case of KRR estimation.
- The early study on KRR estimation in large dataset focuses on either low rank approximation or early-stopping.

Kernel Ridge Regression (KRR)

- Define the KRR estimate $\widehat{f} : \mathbb{R}^1 \mapsto \mathbb{R}^1$ as

$$\widehat{f}_n = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(Z_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\},$$

  where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with a kernel $K(z, z') = \sum_{i=1}^{\infty} \mu_i \phi_i(z) \phi_i(z')$. Here, μ_i's are eigenvalues and $\phi_i(\cdot)$'s are eigenfunctions.
- Explicitly, $\widehat{f}_n(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ with $\boldsymbol{\alpha} = (K + \lambda I)^{-1} \boldsymbol{y}$.
- Smoothing spline is a special case of KRR estimation.
- The early study on KRR estimation in large dataset focuses on either low rank approximation or early-stopping.

Kernel Ridge Regression (KRR)

- Define the KRR estimate $\widehat{f}: \mathbb{R}^1 \mapsto \mathbb{R}^1$ as

$$\widehat{f}_n = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(Z_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\},$$

  where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with a kernel $K(z, z') = \sum_{i=1}^{\infty} \mu_i \phi_i(z) \phi_i(z')$. Here, μ_i's are eigenvalues and $\phi_i(\cdot)$'s are eigenfunctions.
- Explicitly, $\widehat{f}_n(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ with $\boldsymbol{\alpha} = (K + \lambda I)^{-1} \boldsymbol{y}$.
- Smoothing spline is a special case of KRR estimation.
- The early study on KRR estimation in large dataset focuses on either low rank approximation or early-stopping.

Kernel Ridge Regression (KRR)

- Define the KRR estimate $\widehat{f}: \mathbb{R}^1 \mapsto \mathbb{R}^1$ as

$$\widehat{f}_n = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(Z_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\},$$

  where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with a kernel $K(z, z') = \sum_{i=1}^{\infty} \mu_i \phi_i(z) \phi_i(z')$. Here, μ_i's are eigenvalues and $\phi_i(\cdot)$'s are eigenfunctions.
- Explicitly, $\widehat{f}_n(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ with $\boldsymbol{\alpha} = (K + \lambda I)^{-1} \boldsymbol{y}$.
- Smoothing spline is a special case of KRR estimation.
- The early study on KRR estimation in large dataset focuses on either low rank approximation or early-stopping.

Commonly Used Kernels

- Finite Rank ($\mu_k = 0$ for $k > r$):
  - polynomial kernel $K(x, x') = (1 + xx')^d$ with rank $r = d + 1$;
- Exponential Decay ($\mu_k \asymp \exp(-\alpha k^p)$ for some $\alpha, p > 0$):
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ for $p = 2$;
- Polynomial Decay ($\mu_k \asymp k^{-2m}$ for some $m > 1/2$):
  - Kernels for the Sobolev spaces, e.g., $K(x, x') = 1 + min\{x, x'\}$ for the first order Sobolev space;
  - Smoothing spline estimate (Wahba, 1990).
- The decay rate of μ_k characterizes the smoothness of f.

Commonly Used Kernels

- Finite Rank ($\mu_k = 0$ for $k > r$):
  - polynomial kernel $K(x, x') = (1 + xx')^d$ with rank $r = d + 1$;
- Exponential Decay ($\mu_k \asymp \exp(-\alpha k^p)$ for some $\alpha, p > 0$):
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ for $p = 2$;
- Polynomial Decay ($\mu_k \asymp k^{-2m}$ for some $m > 1/2$):
  - Kernels for the Sobolev spaces, e.g., $K(x, x') = 1 + min\{x, x'\}$ for the first order Sobolev space;
  - Smoothing spline estimate (Wahba, 1990).
- The decay rate of μ_k characterizes the smoothness of f.

Commonly Used Kernels

- Finite Rank ($\mu_k = 0$ for $k > r$):
  - polynomial kernel $K(x, x') = (1 + xx')^d$ with rank $r = d + 1$;
- Exponential Decay ($\mu_k \asymp \exp(-\alpha k^p)$ for some $\alpha, p > 0$):
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ for $p = 2$;
- Polynomial Decay ($\mu_k \asymp k^{-2m}$ for some $m > 1/2$):
  - Kernels for the Sobolev spaces, e.g., $K(x, x') = 1 + min\{x, x'\}$ for the first order Sobolev space;
  - Smoothing spline estimate (Wahba, 1990).
- The decay rate of μ_k characterizes the smoothness of f.

Commonly Used Kernels

- Finite Rank ($\mu_k = 0$ for $k > r$):
  - polynomial kernel $K(x, x') = (1 + xx')^d$ with rank $r = d + 1$;
- Exponential Decay ($\mu_k \asymp \exp(-\alpha k^p)$ for some $\alpha, p > 0$):
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ for $p = 2$;
- Polynomial Decay ($\mu_k \asymp k^{-2m}$ for some $m > 1/2$):
  - Kernels for the Sobolev spaces, e.g., $K(x, x') = 1 + min\{x, x'\}$ for the first order Sobolev space;
  - Smoothing spline estimate (Wahba, 1990).
- The decay rate of μ_k characterizes the smoothness of f.

Commonly Used Kernels

- Finite Rank ($\mu_k = 0$ for $k > r$):
  - polynomial kernel $K(x, x') = (1 + xx')^d$ with rank $r = d + 1$;
- Exponential Decay ($\mu_k \asymp \exp(-\alpha k^p)$ for some $\alpha, p > 0$):
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ for $p = 2$;
- Polynomial Decay ($\mu_k \asymp k^{-2m}$ for some $m > 1/2$):
  - Kernels for the Sobolev spaces, e.g., $K(x, x') = 1 + min\{x, x'\}$ for the first order Sobolev space;
  - Smoothing spline estimate (Wahba, 1990).
- The decay rate of μ_k characterizes the smoothness of f.

Commonly Used Kernels

- Finite Rank ($\mu_k = 0$ for $k > r$):
  - polynomial kernel $K(x, x') = (1 + xx')^d$ with rank $r = d + 1$;
- Exponential Decay ($\mu_k \asymp \exp(-\alpha k^p)$ for some $\alpha, p > 0$):
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ for $p = 2$;
- Polynomial Decay ($\mu_k \asymp k^{-2m}$ for some $m > 1/2$):
  - Kernels for the Sobolev spaces, e.g.,
    $K(x, x') = 1 + min\{x, x'\}$ for the first order Sobolev space;
  - Smoothing spline estimate (Wahba, 1990).
- The decay rate of μ_k characterizes the smoothness of f.

Commonly Used Kernels

- Finite Rank ($\mu_k = 0$ for $k > r$):
  - polynomial kernel $K(x, x') = (1 + xx')^d$ with rank $r = d + 1$;
- Exponential Decay ($\mu_k \asymp \exp(-\alpha k^p)$ for some $\alpha, p > 0$):
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ for $p = 2$;
- Polynomial Decay ($\mu_k \asymp k^{-2m}$ for some $m > 1/2$):
  - Kernels for the Sobolev spaces, e.g., $K(x, x') = 1 + min\{x, x'\}$ for the first order Sobolev space;
  - Smoothing spline estimate (Wahba, 1990).
- The decay rate of μ_k characterizes the smoothness of f.

Commonly Used Kernels

- Finite Rank ($\mu_k = 0$ for $k > r$):
  - polynomial kernel $K(x, x') = (1 + xx')^d$ with rank $r = d + 1$;
- Exponential Decay ($\mu_k \asymp \exp(-\alpha k^p)$ for some $\alpha, p > 0$):
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ for $p = 2$;
- Polynomial Decay ($\mu_k \asymp k^{-2m}$ for some $m > 1/2$):
  - Kernels for the Sobolev spaces, e.g.,
    $K(x, x') = 1 + min\{x, x'\}$ for the first order Sobolev space;
  - Smoothing spline estimate (Wahba, 1990).
- The decay rate of μ_k characterizes the smoothness of f.

Local Confidence Interval

Theorem 1. Suppose regularity conditions on ϵ, $K(\cdot,\cdot)$ and $\phi_j(\cdot)$s hold, e.g., ϵ is sub-Gaussian and $\sup_j \|\phi_j\|_\infty \leq C_\phi$. Given that $\mathcal{H}$ is not too large (in terms of its packing entropy), we have for any fixed $x_0 \in \mathcal{X}$,

$$\sqrt{Nh}(\bar{f}_N(x_0) - f_0(x_0)) \xrightarrow{d} N(0, \sigma^2_{x_0}), \tag{1}$$

where $h = h(\lambda) = r(\lambda)^{-1}$ and $r(\lambda) \equiv \sum_{i=1}^{\infty}\{1 + \lambda/\mu_i\}^{-1}$.

An important consequence is that the rate $\sqrt{Nh}$ and variance $\sigma^2_{x_0}$ are the same as those of $\widehat{f}_N$ (based on the entire dataset). Hence, the oracle property of the local confidence interval holds under the above conditions on s and λ.

- In Theorem 1, some under-smoothing condition is implicitly assumed (so, there is no estimation bias).
- Technical Challenges:
    - generalize the functional Bahadur representation developed for smoothing spline estimation (Shang and Cheng, 2013, AoS) to KRR estimation;
    - employ empirical process theory to study the average of s asymptotic linear expansions as $s \to \infty$;

- In Theorem 1, some under-smoothing condition is implicitly assumed (so, there is no estimation bias).
- Technical Challenges:
    - generalize the functional Bahadur representation developed for smoothing spline estimation (Shang and Cheng, 2013, AoS) to KRR estimation;
    - employ empirical process theory to study the average of s asymptotic linear expansions as $s \to \infty$;

- In Theorem 1, some under-smoothing condition is implicitly assumed (so, there is no estimation bias).
- Technical Challenges:
  - generalize the functional Bahadur representation developed for smoothing spline estimation (Shang and Cheng, 2013, AoS) to KRR estimation;
  - employ empirical process theory to study the average of s asymptotic linear expansions as $s \to \infty$;

- In Theorem 1, some under-smoothing condition is implicitly assumed (so, there is no estimation bias).
- Technical Challenges:
  - generalize the functional Bahadur representation developed for smoothing spline estimation (Shang and Cheng, 2013, AoS) to KRR estimation;
  - employ empirical process theory to study the average of s asymptotic linear expansions as $s \to \infty$;

Examples

The oracle property of local confidence interval holds under the following conditions on λ and s:

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$, $\log(\lambda^{-1}) = o(\log^2 N)$ and $s = o(N^{1/2}/\{\log^{1/2}(\lambda^{-1}) \log^3(N)\})$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$, $\log(\lambda^{-1}) = o(\log^2(N))$ and $s = o(N^{1/2} h^{3/2}/\{[\log(h/\lambda)]^{(p+1)/2p} \log^3(N)\})$ with $h = [\log(1/\lambda)]^{-1/p}$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$ and $s = N^{\gamma}$ with $\gamma < 1/2 - (8m-1)/(8m^2)d$.

Examples

The oracle property of local confidence interval holds under the following conditions on λ and s:

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$, $\log(\lambda^{-1}) = o(\log^2 N)$ and $s = o(N^{1/2}/\{\log^{1/2}(\lambda^{-1})\log^3(N)\})$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$, $\log(\lambda^{-1}) = o(\log^2(N))$ and $s = o(N^{1/2}h^{3/2}/\{[\log(h/\lambda)]^{(p+1)/2p}\log^3(N)\})$ with $h = [\log(1/\lambda)]^{-1/p}$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$ and $s = N^{\gamma}$ with $\gamma < 1/2 - (8m-1)/(8m^2)d$.

Examples

The oracle property of local confidence interval holds under the following conditions on λ and s:

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$, $\log(\lambda^{-1}) = o(\log^2 N)$ and $s = o(N^{1/2}/\{\log^{1/2}(\lambda^{-1}) \log^3(N)\})$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$, $\log(\lambda^{-1}) = o(\log^2(N))$ and $s = o(N^{1/2}h^{3/2}/\{[\log(h/\lambda)]^{(p+1)/2p} \log^3(N)\})$ with $h = [\log(1/\lambda)]^{-1/p}$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$ and $s = N^{\gamma}$ with $\gamma < 1/2 - (8m-1)/(8m^2)d$.

Examples

The oracle property of local confidence interval holds under the following conditions on λ and s:

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$, $\log(\lambda^{-1}) = o(\log^2 N)$ and $s = o(N^{1/2}/\{\log^{1/2}(\lambda^{-1}) \log^3(N)\})$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$, $\log(\lambda^{-1}) = o(\log^2(N))$ and $s = o(N^{1/2}h^{3/2}/\{[\log(h/\lambda)]^{(p+1)/2p} \log^3(N)\})$ with $h = [\log(1/\lambda)]^{-1/p}$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$ and $s = N^{\gamma}$ with $\gamma < 1/2 - (8m-1)/(8m^2)d$.

Examples

The oracle property of local confidence interval holds under the following conditions on λ and s:

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$, $\log(\lambda^{-1}) = o(\log^2 N)$ and $s = o(N^{1/2}/\{\log^{1/2}(\lambda^{-1})\log^3(N)\})$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$, $\log(\lambda^{-1}) = o(\log^2(N))$ and $s = o(N^{1/2}h^{3/2}/\{[\log(h/\lambda)]^{(p+1)/2p}\log^3(N)\})$ with $h = [\log(1/\lambda)]^{-1/p}$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$ and $s = N^{\gamma}$ with $\gamma < 1/2 - (8m-1)/(8m^2)d$.

Examples

The oracle property of local confidence interval holds under the following conditions on λ and s:

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$, $\log(\lambda^{-1}) = o(\log^2 N)$ and $s = o(N^{1/2}/\{\log^{1/2}(\lambda^{-1})\log^3(N)\})$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$, $\log(\lambda^{-1}) = o(\log^2(N))$ and $s = o(N^{1/2}h^{3/2}/\{[\log(h/\lambda)]^{(p+1)/2p}\log^3(N)\})$ with $h = [\log(1/\lambda)]^{-1/p}$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$ and $s = N^{\gamma}$ with $\gamma < 1/2 - (8m-1)/(8m^2)d$.

Specifically, we have the following upper bounds for s:

- For finite rank kernel (with any finite rank r), $s = O(N^{\gamma})$ for any $\gamma < 1/2$;
- For exponential decay kernel (with any finite power p), $s = O(N^{\gamma'})$ for any $\gamma' < \gamma < 1/2$;
- For polynomial decay kernel (with $m = 2$), $s = o(N^{4/27}) \approx o(N^{0.29})$;

Specifically, we have the following upper bounds for s:

- For finite rank kernel (with any finite rank r), $s = O(N^{\gamma})$ for any $\gamma < 1/2$;
- For exponential decay kernel (with any finite power p), $s = O(N^{\gamma'})$ for any $\gamma' < \gamma < 1/2$;
- For polynomial decay kernel (with $m = 2$), $s = o(N^{4/27}) \approx o(N^{0.29})$;

Specifically, we have the following upper bounds for s:

- For finite rank kernel (with any finite rank r), $s = O(N^{\gamma})$ for any $\gamma < 1/2$;
- For exponential decay kernel (with any finite power p), $s = O(N^{\gamma'})$ for any $\gamma' < \gamma < 1/2$;
- For polynomial decay kernel (with $m = 2$), $s = o(N^{4/27}) \approx o(N^{0.29})$;

Big Data Insights

- The number of subsets s:
  Divide-and-conquer approach prefers more smooth function in the sense that we can save more computational efforts (larger s) for achieving the oracle property in this case.
- The smoothing parameter λ:
  Choose λ as if working on the entire dataset with sample size N (although it is sub-optimal for each sub-estimating). This theoretical finding leads to a modified GCV formula used in practice. Similar phenomenon occurs for obtaining nonparametric minimaxity (Zhang et al, 2013).

Big Data Insights

- The number of subsets s:
  Divide-and-conquer approach prefers more smooth function in the sense that we can save more computational efforts (larger s) for achieving the oracle property in this case.
- The smoothing parameter λ:
  Choose λ as if working on the entire dataset with sample size N (although it is sub-optimal for each sub-estimating). This theoretical finding leads to a modified GCV formula used in practice. Similar phenomenon occurs for obtaining nonparametric minimaxity (Zhang et al, 2013).

Penalized Likelihood Ratio Test

- Consider the following test:

$$H_0 : f = f_0 \;\text{ v.s. }\; H_1 : f \neq f_0,$$

  where $f_0 \in \mathcal{H}$;
- Let $\mathcal{L}_{N,\lambda}$ be the (penalized) likelihood function based on the entire dataset.
- Let $PLRT_{n,\lambda}^{(j)}$ be the (penalized) likelihood ratio based on the j-th subsample.
- Given the Divide-and-Conquer strategy, we have two natural choices of test statistic:
  - $\widetilde{PLRT}_{N,\lambda} = (1/s)\sum_{j=1}^{s} PLRT_{n,\lambda}^{(j)}$;
  - $\widehat{PLRT}_{N,\lambda} = \mathcal{L}_{N,\lambda}(\bar{f}_N) - \mathcal{L}_{N,\lambda}(f_0)$;

Penalized Likelihood Ratio Test

- Consider the following test:

$$H_0 : f = f_0 \text{ v.s. } H_1 : f \neq f_0,$$

  where $f_0 \in \mathcal{H}$;
- Let $\mathcal{L}_{N,\lambda}$ be the (penalized) likelihood function based on the entire dataset.
- Let $PLRT^{(j)}_{n,\lambda}$ be the (penalized) likelihood ratio based on the j-th subsample.
- Given the Divide-and-Conquer strategy, we have two natural choices of test statistic:
  - $\widetilde{PLRT}_{N,\lambda} = (1/s)\sum_{j=1}^{s} PLRT^{(j)}_{n,\lambda}$;
  - $\widehat{PLRT}_{N,\lambda} = \mathcal{L}_{N,\lambda}(\bar{f}_N) - \mathcal{L}_{N,\lambda}(f_0)$;

Penalized Likelihood Ratio Test

- Consider the following test:

$$H_0 : f = f_0 \text{ v.s. } H_1 : f \neq f_0,$$

  where $f_0 \in \mathcal{H}$;
- Let $\mathcal{L}_{N,\lambda}$ be the (penalized) likelihood function based on the entire dataset.
- Let $PLRT_{n,\lambda}^{(j)}$ be the (penalized) likelihood ratio based on the j-th subsample.
- Given the Divide-and-Conquer strategy, we have two natural choices of test statistic:
  - $\widetilde{PLRT}_{N,\lambda} = (1/s)\sum_{j=1}^{s} PLRT_{n,\lambda}^{(j)}$;
  - $\widehat{PLRT}_{N,\lambda} = \mathcal{L}_{N,\lambda}(\bar{f}_N) - \mathcal{L}_{N,\lambda}(f_0)$;

Penalized Likelihood Ratio Test

- Consider the following test:

$$H_0 : f = f_0 \text{ v.s. } H_1 : f \neq f_0,$$

  where $f_0 \in \mathcal{H}$;
- Let $\mathcal{L}_{N,\lambda}$ be the (penalized) likelihood function based on the entire dataset.
- Let $PLRT^{(j)}_{n,\lambda}$ be the (penalized) likelihood ratio based on the j-th subsample.
- Given the Divide-and-Conquer strategy, we have two natural choices of test statistic:
  - $\widetilde{PLRT}_{N,\lambda} = (1/s)\sum_{j=1}^{s} PLRT^{(j)}_{n,\lambda}$;
  - $\widehat{PLRT}_{N,\lambda} = \mathcal{L}_{N,\lambda}(\bar{f}_N) - \mathcal{L}_{N,\lambda}(f_0)$;

Penalized Likelihood Ratio Test

- Consider the following test:

$$H_0 : f = f_0 \text{ v.s. } H_1 : f \neq f_0,$$

  where $f_0 \in \mathcal{H}$;
- Let $\mathcal{L}_{N,\lambda}$ be the (penalized) likelihood function based on the entire dataset.
- Let $PLRT_{n,\lambda}^{(j)}$ be the (penalized) likelihood ratio based on the j-th subsample.
- Given the Divide-and-Conquer strategy, we have two natural choices of test statistic:
  - $\widetilde{PLRT}_{N,\lambda} = (1/s) \sum_{j=1}^{s} PLRT_{n,\lambda}^{(j)}$;
  - $\widehat{PLRT}_{N,\lambda} = \mathcal{L}_{N,\lambda}(\bar{f}_N) - \mathcal{L}_{N,\lambda}(f_0)$;

Penalized Likelihood Ratio Test

- Consider the following test:

$$H_0 : f = f_0 \text{ v.s. } H_1 : f \neq f_0,$$

where $f_0 \in \mathcal{H}$;
- Let $\mathcal{L}_{N,\lambda}$ be the (penalized) likelihood function based on the entire dataset.
- Let $PLRT_{n,\lambda}^{(j)}$ be the (penalized) likelihood ratio based on the j-th subsample.
- Given the Divide-and-Conquer strategy, we have two natural choices of test statistic:
  - $\widetilde{PLRT}_{N,\lambda} = (1/s) \sum_{j=1}^{s} PLRT_{n,\lambda}^{(j)}$;
  - $\widehat{PLRT}_{N,\lambda} = \mathcal{L}_{N,\lambda}(\bar{f}_N) - \mathcal{L}_{N,\lambda}(f_0)$;

Penalized Likelihood Ratio Test

Theorem 3. We prove that $\widetilde{PLRT}_{N,\lambda}$ and $\widehat{PLRT}_{N,\lambda}$ are both consistent under some upper bound of s, but the latter is minimax optimal (Ingster, 1993) when choosing some s *strictly* smaller than the above upper bound.

- An additional big data insight: we have to sacrifice certain amount of computational efficiency (not choose the largest possible s) for obtaining the optimality.

Penalized Likelihood Ratio Test

Theorem 3. We prove that $\widetilde{PLRT}_{N,\lambda}$ and $\widehat{PLRT}_{N,\lambda}$ are both consistent under some upper bound of s, but the latter is minimax optimal (Ingster, 1993) when choosing some s *strictly* smaller than the above upper bound.

- An additional big data insight: we have to sacrifice certain amount of computational efficiency (not choose the largest possible s) for obtaining the optimality.

Summary

- Technically, this work generalizes Shang and Cheng (2013, AoS) for smoothing spline inference in two aspects:
  - KRR inference;
  - Divide and Conquer strategy $s = 1 \Longrightarrow s \to \infty$;
- Big Data Insights:
  - Oracle rule holds when s does not grow too fast;
  - D&C approach prefers more smooth regression functions;
  - choose the smoothing parameter as if not splitting the data;
  - sacrifice computational efficiency for obtaining optimality.

Summary

- Technically, this work generalizes Shang and Cheng (2013, AoS) for smoothing spline inference in two aspects:
  - KRR inference;
  - Divide and Conquer strategy $s = 1 \Longrightarrow s \to \infty$;
- Big Data Insights:
  - Oracle rule holds when s does not grow too fast;
  - D&C approach prefers more smooth regression functions;
  - choose the smoothing parameter as if not splitting the data;
  - sacrifice computational efficiency for obtaining optimality.

Summary

- Technically, this work generalizes Shang and Cheng (2013, AoS) for smoothing spline inference in two aspects:
  - KRR inference;
  - Divide and Conquer strategy $s = 1 \Longrightarrow s \to \infty$;
- Big Data Insights:
  - Oracle rule holds when s does not grow too fast;
  - D&C approach prefers more smooth regression functions;
  - choose the smoothing parameter as if not splitting the data;
  - sacrifice computational efficiency for obtaining optimality.

Summary

- Technically, this work generalizes Shang and Cheng (2013, AoS) for smoothing spline inference in two aspects:
  - KRR inference;
  - Divide and Conquer strategy $s = 1 \Longrightarrow s \to \infty$;
- Big Data Insights:
  - Oracle rule holds when s does not grow too fast;
  - D&C approach prefers more smooth regression functions;
  - choose the smoothing parameter as if not splitting the data;
  - sacrifice computational efficiency for obtaining optimality.

Summary

- Technically, this work generalizes Shang and Cheng (2013, AoS) for smoothing spline inference in two aspects:
  - KRR inference;
  - Divide and Conquer strategy $s = 1 \Longrightarrow s \to \infty$;
- Big Data Insights:
  - Oracle rule holds when s does not grow too fast;
  - D&C approach prefers more smooth regression functions;
  - choose the smoothing parameter as if not splitting the data;
  - sacrifice computational efficiency for obtaining optimality.

Summary

- Technically, this work generalizes Shang and Cheng (2013, AoS) for smoothing spline inference in two aspects:
  - KRR inference;
  - Divide and Conquer strategy $s = 1 \Longrightarrow s \to \infty$;
- Big Data Insights:
  - Oracle rule holds when s does not grow too fast;
  - D&C approach prefers more smooth regression functions;
  - choose the smoothing parameter as if not splitting the data;
  - sacrifice computational efficiency for obtaining optimality.

Summary

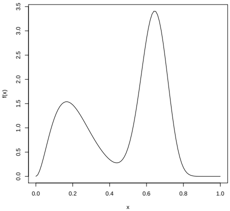
- Technically, this work generalizes Shang and Cheng (2013, AoS) for smoothing spline inference in two aspects:
  - KRR inference;
  - Divide and Conquer strategy $s = 1 \Longrightarrow s \to \infty$;
- Big Data Insights:
  - Oracle rule holds when s does not grow too fast;
  - D&C approach prefers more smooth regression functions;
  - choose the smoothing parameter as if not splitting the data;
  - sacrifice computational efficiency for obtaining optimality.

Summary

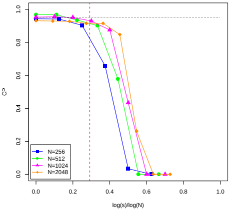
- Technically, this work generalizes Shang and Cheng (2013, AoS) for smoothing spline inference in two aspects:
  - KRR inference;
  - Divide and Conquer strategy $s = 1 \Longrightarrow s \to \infty$;
- Big Data Insights:
  - Oracle rule holds when s does not grow too fast;
  - D&C approach prefers more smooth regression functions;
  - choose the smoothing parameter as if not splitting the data;
  - sacrifice computational efficiency for obtaining optimality.

Phase Transition of Coverage Probability

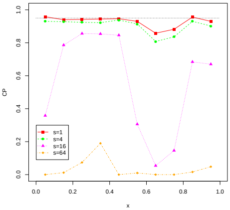
(a) True function



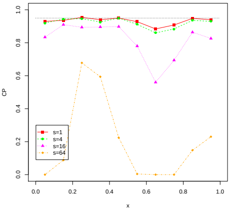
(b) CPs at $x_0 = 0.5$



(c) CPs on $[0, 1]$ for $N = 512$



(d) CPs on $[0, 1]$ for $N = 1024$



Part II: Heterogeneous Data

Outline

A Motivating Example

- Different biology labs conduct the same experiment on the relationship between a response variable Y (e.g., heart disease) and a set of predictors $Z, X_1, X_2, \ldots, X_p$;
- Biology suggests that the relation between Y and Z (e.g., blood pressure) should be homogeneous for all human;
- However, for the other covariates $X_1, X_2, \ldots, X_p$ (e.g., certain genes), we allow their relations with Y to potentially vary in different labs. For example, the genetic functionality of different races might be heterogenous.

A Motivating Example

- Different biology labs conduct the same experiment on the relationship between a response variable Y (e.g., heart disease) and a set of predictors $Z, X_1, X_2, \ldots, X_p$;
- Biology suggests that the relation between Y and Z (e.g., blood pressure) should be homogeneous for all human;
- However, for the other covariates $X_1, X_2, \ldots, X_p$ (e.g., certain genes), we allow their relations with Y to potentially vary in different labs. For example, the genetic functionality of different races might be heterogenous.

A Motivating Example

- Different biology labs conduct the same experiment on the relationship between a response variable Y (e.g., heart disease) and a set of predictors $Z, X_1, X_2, \ldots, X_p$;
- Biology suggests that the relation between Y and Z (e.g., blood pressure) should be homogeneous for all human;
- However, for the other covariates $X_1, X_2, \ldots, X_p$ (e.g., certain genes), we allow their relations with Y to potentially vary in different labs. For example, the genetic functionality of different races might be heterogenous.

A Partially Linear Model

- Assume that there exist s heterogeneous subpopulations: $P_1, \ldots, P_s$ (with equal sample size $n = N/s$);
- In the j-th subpopulation, we assume

$$Y = \mathbf{X}^T \beta_0^{(j)} + f_0(Z) + \epsilon, \tag{1}$$

  where ϵ has a sub-Gaussian tail and $Var(\epsilon) = \sigma^2$;
- We call $\beta^{(j)}$ as the heterogeneity and f as the commonality of the massive data in consideration;
- (1) is a typical semi-nonparametric model (see Cheng and Shang, 2014) since $\beta^{(j)}$ and f are both of interest.

A Partially Linear Model

- Assume that there exist s heterogeneous subpopulations: $P_1, \ldots, P_s$ (with equal sample size $n = N/s$);
- In the j-th subpopulation, we assume

$$Y = \mathbf{X}^T \boldsymbol{\beta}_0^{(j)} + f_0(Z) + \epsilon, \tag{1}$$

where ϵ has a sub-Gaussian tail and $Var(\epsilon) = \sigma^2$;
- We call $\boldsymbol{\beta}^{(j)}$ as the heterogeneity and f as the commonality of the massive data in consideration;
- (1) is a typical semi-nonparametric model (see Cheng and Shang, 2014) since $\boldsymbol{\beta}^{(j)}$ and f are both of interest.

A Partially Linear Model

- Assume that there exist s heterogeneous subpopulations: $P_1, \ldots, P_s$ (with equal sample size $n = N/s$);
- In the j-th subpopulation, we assume

$$Y = \mathbf{X}^T \boldsymbol{\beta}_0^{(j)} + f_0(Z) + \epsilon, \tag{1}$$

  where ϵ has a sub-Gaussian tail and $Var(\epsilon) = \sigma^2$;
- We call $\boldsymbol{\beta}^{(j)}$ as the heterogeneity and f as the commonality of the massive data in consideration;
- (1) is a typical semi-nonparametric model (see Cheng and Shang, 2014) since $\boldsymbol{\beta}^{(j)}$ and f are both of interest.

A Partially Linear Model

- Assume that there exist s heterogeneous subpopulations: $P_1, \ldots, P_s$ (with equal sample size $n = N/s$);
- In the j-th subpopulation, we assume

$$Y = \mathbf{X}^T \boldsymbol{\beta}_0^{(j)} + f_0(Z) + \epsilon, \tag{1}$$

  where ϵ has a sub-Gaussian tail and $Var(\epsilon) = \sigma^2$;
- We call $\boldsymbol{\beta}^{(j)}$ as the heterogeneity and f as the commonality of the massive data in consideration;
- (1) is a typical semi-nonparametric model (see Cheng and Shang, 2014) since $\boldsymbol{\beta}^{(j)}$ and f are both of interest.

Estimation Procedure for Heterogeneous Data

- Individual estimation in the j-th subpopulation:

$$(\widehat{\boldsymbol{\beta}}_n^{(j)}, \widehat{f}_n^{(j)}) = \underset{(\boldsymbol{\beta}, f) \in \mathbb{R}^p \times \mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i^{(j)} - \boldsymbol{\beta}^T \mathbf{X}_i^{(j)} - f(Z_i^{(j)}) \right)^2 + \lambda \| f \|_{\mathcal{H}}^2 \right\};$$

- Aggregation: $\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)}$;
- A plug-in estimate for the j-th heterogeneity parameter:

$$\check{\beta}_n^{(j)} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i^{(j)} - \boldsymbol{\beta}^T \mathbf{X}_i^{(j)} - \bar{f}_N(Z_i^{(j)}) \right)^2;$$

- Our final estimate is $(\check{\beta}_n^{(j)}, \bar{f}_N)$.

Estimation Procedure for Heterogeneous Data

- Individual estimation in the j-th subpopulation:

$$(\widehat{\boldsymbol{\beta}}_n^{(j)}, \widehat{f}_n^{(j)}) = \operatorname*{argmin}_{(\boldsymbol{\beta}, f) \in \mathbb{R}^p \times \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i^{(j)} - \boldsymbol{\beta}^T \mathbf{X}_i^{(j)} - f(Z_i^{(j)}) \right)^2 + \lambda \| f \|_{\mathcal{H}}^2 \right\};$$

- Aggregation: $\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)}$;
- A plug-in estimate for the j-th heterogeneity parameter:

$$\check{\beta}_n^{(j)} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i^{(j)} - \boldsymbol{\beta}^T \mathbf{X}_i^{(j)} - \bar{f}_N(Z_i^{(j)}) \right)^2;$$

- Our final estimate is $(\check{\beta}_n^{(j)}, \bar{f}_N)$.

Estimation Procedure for Heterogeneous Data

- Individual estimation in the j-th subpopulation:

$$(\widehat{\boldsymbol{\beta}}_n^{(j)}, \widehat{f}_n^{(j)}) = \underset{(\boldsymbol{\beta}, f) \in \mathbb{R}^p \times \mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i^{(j)} - \boldsymbol{\beta}^T \mathbf{X}_i^{(j)} - f(Z_i^{(j)}) \right)^2 + \lambda \| f \|_{\mathcal{H}}^2 \right\};$$

- Aggregation: $\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)}$;
- A plug-in estimate for the j-th heterogeneity parameter:

$$\check{\boldsymbol{\beta}}_n^{(j)} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i^{(j)} - \boldsymbol{\beta}^T \mathbf{X}_i^{(j)} - \bar{f}_N(Z_i^{(j)}) \right)^2;$$

- Our final estimate is $(\check{\boldsymbol{\beta}}_n^{(j)}, \bar{f}_N)$.

Estimation Procedure for Heterogeneous Data

- Individual estimation in the j-th subpopulation:

$$(\widehat{\boldsymbol{\beta}}_n^{(j)}, \widehat{f}_n^{(j)}) = \underset{(\boldsymbol{\beta}, f)\in\mathbb{R}^p\times\mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^n \left(Y_i^{(j)} - \boldsymbol{\beta}^T\mathbf{X}_i^{(j)} - f(Z_i^{(j)})\right)^2 + \lambda\|f\|_{\mathcal{H}}^2 \right\};$$

- Aggregation: $\bar{f}_N = (1/s)\sum_{j=1}^s \widehat{f}_n^{(j)}$;
- A plug-in estimate for the j-th heterogeneity parameter:

$$\check{\boldsymbol{\beta}}_n^{(j)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \left(Y_i^{(j)} - \boldsymbol{\beta}^T\mathbf{X}_i^{(j)} - \bar{f}_N(Z_i^{(j)})\right)^2;$$

- Our final estimate is $(\check{\boldsymbol{\beta}}_n^{(j)}, \bar{f}_N)$.

Relation to Homogeneous Data

- The major concern of homogeneous data is the extremely high computational cost. Fortunately, this can be dealt by the divide-and-conquer approach;
- When dealing with heterogeneous data, our major interest is about how to efficiently extract common features across many subpopulations while exploring heterogeneity of each subpopulation as $s \to \infty$;
- Some sub-populations in heterogeneous data may have huge sample sizes. In this case, the divide-and-conquer approach can be applied.

Relation to Homogeneous Data

- The major concern of homogeneous data is the extremely high computational cost. Fortunately, this can be dealt by the divide-and-conquer approach;
- When dealing with heterogeneous data, our major interest is about how to efficiently extract common features across many subpopulations while exploring heterogeneity of each subpopulation as $s \to \infty$;
- Some sub-populations in heterogeneous data may have huge sample sizes. In this case, the divide-and-conquer approach can be applied.

Relation to Homogeneous Data

- The major concern of homogeneous data is the extremely high computational cost. Fortunately, this can be dealt by the divide-and-conquer approach;
- When dealing with heterogeneous data, our major interest is about how to efficiently extract common features across many subpopulations while exploring heterogeneity of each subpopulation as $s \to \infty$;
- Some sub-populations in heterogeneous data may have huge sample sizes. In this case, the divide-and-conquer approach can be applied.

Theorem (Joint Normality Theorem)

Suppose regularity conditions, e.g., under-smoothing condition, and $E(\mathbf{X}_k|Z) \in \mathcal{H}$ hold. Given proper s and λ, we have

(i) *if $s \to \infty$ then*

$$\begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{(j)} - \boldsymbol{\beta}_0^{(j)}) \\ \sqrt{Nh}\big(\bar{f}_N(z_0) - f_0(z_0)\big) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^2 \begin{pmatrix} \Omega^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{pmatrix}\right),$$

where $\Omega = E(\mathbf{X} - E(\mathbf{X}|Z))^{\otimes 2}$;

(ii) *if s is fixed, then*

$$\begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{(j)} - \boldsymbol{\beta}_0^{(j)}) \\ \sqrt{Nh}\big(\bar{f}_N(z_0) - f_0(z_0)\big) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^2 \begin{pmatrix} \Omega^{-1} & \Sigma_{21}/\sqrt{s} \\ \Sigma_{12}/\sqrt{s} & \Sigma_{22} \end{pmatrix}\right).$$

Moreover, if $h \to 0$ as $N \to \infty$, then $\Sigma_{12} = \Sigma_{21} = \mathbf{0}$.

Some Consequences

- Some calculations in concrete examples indicate that an upper bound is imposed on s and λ is chosen in the order of N (as if the regularization were based on the entire data);
- Note that $\widehat{\beta}^{(j)}$ is scaled to n and $\bar{f}(z_0)$ is scaled to N. Hence, it is not surprising that as $s \to \infty$ $(n/N \to 0)$, these two estimate become asymptotically independent;
- The case that $h \nrightarrow 0$ is a trivial case. For example, $h \asymp r^{-1}$ for finite rank kernel. In this case, the semi-nonparametric estimation essentially reduces to a parametric one;

Some Consequences

- Some calculations in concrete examples indicate that an upper bound is imposed on s and λ is chosen in the order of N (as if the regularization were based on the entire data);
- Note that $\widehat{\boldsymbol{\beta}}^{(j)}$ is scaled to n and $\bar{f}(z_0)$ is scaled to N. Hence, it is not surprising that as $s \to \infty$ $(n/N \to 0)$, these two estimate become asymptotically independent;
- The case that $h \nrightarrow 0$ is a trivial case. For example, $h \asymp r^{-1}$ for finite rank kernel. In this case, the semi-nonparametric estimation essentially reduces to a parametric one;

Some Consequences

- Some calculations in concrete examples indicate that an upper bound is imposed on s and λ is chosen in the order of N (as if the regularization were based on the entire data);
- Note that $\widehat{\boldsymbol{\beta}}^{(j)}$ is scaled to n and $\bar{f}(z_0)$ is scaled to N. Hence, it is not surprising that as $s \to \infty$ $(n/N \to 0)$, these two estimate become asymptotically independent;
- The case that $h \nrightarrow 0$ is a trivial case. For example, $h \asymp r^{-1}$ for finite rank kernel. In this case, the semi-nonparametric estimation essentially reduces to a parametric one;

Oracle Rule

- The main message delivered by the above theorem is that our combined estimate enjoys the "oracle property" in the sense that $\bar{f}$ shares the same asymptotic distribution as the "oracle estimate" $\widehat{f}_{or}$ computed as if there were no heterogeneity in the data:

$$\widehat{f}_{or} = \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ \frac{1}{N} \sum_{i,j=1}^{n,s} \left(Y_i^{(j)} - (\boldsymbol{\beta}_0^{(j)})^T \mathbf{X}_i^{(j)} - f(Z_i^{(j)}) \right)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$

- In other words, our aggregation procedure can "filter out" the heterogeneity in data when s does not grow too fast with N.

Oracle Rule

- The main message delivered by the above theorem is that our combined estimate enjoys the "oracle property" in the sense that $\bar{f}$ shares the same asymptotic distribution as the "oracle estimate" $\widehat{f}_{or}$ computed as if there were no heterogeneity in the data:

$$\begin{aligned}&\widehat{f}_{or}\\ =\ &\operatorname*{argmin}_{f\in\mathcal{H}}\left\{\frac{1}{N}\sum_{i,j=1}^{n,s}\left(Y_i^{(j)}-(\boldsymbol{\beta}_0^{(j)})^T\mathbf{X}_i^{(j)}-f(Z_i^{(j)})\right)^2+\lambda\|f\|_{\mathcal{H}}^2\right\}\end{aligned}$$

- In other words, our aggregation procedure can "filter out" the heterogeneity in data when s does not grow too fast with N.

Efficiency Boosting

- The aggregation of commonality in turn boosts the estimation efficiency of $\widehat{\boldsymbol{\beta}}_n^{(j)}$ from semiparametric level to parametric level;
- Recall our final estimate for $\boldsymbol{\beta}_0^{(j)}$:

$$\check{\boldsymbol{\beta}}_n^{(j)} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{(j)} - \boldsymbol{\beta}^T\mathbf{X}_i^{(j)} - \bar{f}_N(Z_i^{(j)})\right)^2; \quad (2)$$

- By imposing some lower bound on s[1], we show that

$$\sqrt{n}(\check{\boldsymbol{\beta}}_n^{(j)} - \boldsymbol{\beta}_0^{(j)}) \rightsquigarrow N(0, \sigma^2(E[\mathbf{X}\mathbf{X}^T])^{-1})$$

as if the commonality information were available;
- This represents one important feature of massive data: strength-borrowing.

[1] This lower bound requirement slows down the convergence rate of $\check{\boldsymbol{\beta}}_n^{(j)}$ such that $\bar{f}_N$ can be treated as if it were known.

Efficiency Boosting

- The aggregation of commonality in turn boosts the estimation efficiency of $\widehat{\boldsymbol{\beta}}_n^{(j)}$ from semiparametric level to parametric level;
- Recall our final estimate for $\boldsymbol{\beta}_0^{(j)}$:

$$\check{\boldsymbol{\beta}}_n^{(j)} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{(j)} - \boldsymbol{\beta}^T\mathbf{X}_i^{(j)} - \bar{f}_N(Z_i^{(j)})\right)^2; \qquad (2)$$

- By imposing some lower bound on s[1], we show that

$$\sqrt{n}(\check{\beta}_n^{(j)} - \beta_0^{(j)}) \rightsquigarrow N(0, \sigma^2(E[\mathbf{X}\mathbf{X}^T])^{-1})$$

as if the commonality information were available;
- This represents one important feature of massive data: strength-borrowing.

[1] This lower bound requirement slows down the convergence rate of $\check{\beta}_n^{(j)}$ such that $\bar{f}_N$ can be treated as if it were known.

Efficiency Boosting

- The aggregation of commonality in turn boosts the estimation efficiency of $\widehat{\boldsymbol{\beta}}_n^{(j)}$ from semiparametric level to parametric level;
- Recall our final estimate for $\boldsymbol{\beta}_0^{(j)}$:

$$\check{\boldsymbol{\beta}}_n^{(j)} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i^{(j)} - \boldsymbol{\beta}^T \mathbf{X}_i^{(j)} - \bar{f}_N(Z_i^{(j)}) \right)^2; \qquad (2)$$

- By imposing some lower bound on s[1], we show that

$$\sqrt{n}(\check{\boldsymbol{\beta}}_n^{(j)} - \boldsymbol{\beta}_0^{(j)}) \rightsquigarrow N(0, \sigma^2 (E[\mathbf{X}\mathbf{X}^T])^{-1})$$

as if the commonality information were available;

- This represents one important feature of massive data: strength-borrowing.

---

[1] This lower bound requirement slows down the convergence rate of $\check{\boldsymbol{\beta}}_n^{(j)}$ such that $\bar{f}_N$ can be treated as if it were known.

Efficiency Boosting

- The aggregation of commonality in turn boosts the estimation efficiency of $\widehat{\boldsymbol{\beta}}_n^{(j)}$ from semiparametric level to parametric level;
- Recall our final estimate for $\boldsymbol{\beta}_0^{(j)}$:

$$\check{\boldsymbol{\beta}}_n^{(j)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{(j)} - \boldsymbol{\beta}^T\mathbf{X}_i^{(j)} - \bar{f}_N(Z_i^{(j)})\right)^2; \qquad (2)$$

- By imposing some lower bound on s[1], we show that

$$\sqrt{n}(\check{\boldsymbol{\beta}}_n^{(j)} - \boldsymbol{\beta}_0^{(j)}) \rightsquigarrow N(0, \sigma^2(E[\mathbf{X}\mathbf{X}^T])^{-1})$$

  as if the commonality information were available;
- This represents one important feature of massive data: strength-borrowing.

[1] This lower bound requirement slows down the convergence rate of $\check{\boldsymbol{\beta}}_n^{(j)}$ such that $\bar{f}_N$ can be treated as if it were known.

- Consider a *high dimensional* simultaneous testing:

$$H_0 : \boldsymbol{\beta}^{(j)} = \widetilde{\boldsymbol{\beta}}^{(j)} \text{ for all } j \in J, \tag{3}$$

where $J \subset \{1, 2, \ldots, s\}$, versus the alternative:

$$H_1 : \boldsymbol{\beta}^{(j)} \neq \widetilde{\boldsymbol{\beta}}^{(j)} \text{ for some } j \in J; \tag{4}$$

- Test statistic:

$$T_0 = \sup_{j \in J} \sup_{k \in [p]} \sqrt{n} |\check{\beta}_k^{(j)} - \widetilde{\beta}_k|;$$

- By employing a recent Gaussian approximation theory, we can consistently approximate the quantile of the null distribution via bootstrap even when $|J|$ diverges at an exponential rate of n.

- Consider a *high dimensional* simultaneous testing:

$$H_0 : \boldsymbol{\beta}^{(j)} = \widetilde{\boldsymbol{\beta}}^{(j)} \text{ for all } j \in J, \tag{3}$$

where $J \subset \{1, 2, \ldots, s\}$, versus the alternative:

$$H_1 : \boldsymbol{\beta}^{(j)} \neq \widetilde{\boldsymbol{\beta}}^{(j)} \text{ for some } j \in J; \tag{4}$$

- Test statistic:

$$T_0 = \sup_{j \in J} \sup_{k \in [p]} \sqrt{n} |\check{\beta}_k^{(j)} - \widetilde{\beta}_k|;$$

- By employing a recent Gaussian approximation theory, we can consistently approximate the quantile of the null distribution via bootstrap even when $|J|$ diverges at an exponential rate of n.

- Consider a *high dimensional* simultaneous testing:

$$H_0 : \boldsymbol{\beta}^{(j)} = \widetilde{\boldsymbol{\beta}}^{(j)} \text{ for all } j \in J, \tag{3}$$

where $J \subset \{1, 2, \ldots, s\}$, versus the alternative:

$$H_1 : \boldsymbol{\beta}^{(j)} \neq \widetilde{\boldsymbol{\beta}}^{(j)} \text{ for some } j \in J; \tag{4}$$

- Test statistic:

$$T_0 = \sup_{j \in J} \sup_{k \in [p]} \sqrt{n} |\check{\beta}_k^{(j)} - \widetilde{\beta}_k|;$$

- By employing a recent Gaussian approximation theory, we can consistently approximate the quantile of the null distribution via bootstrap even when $|J|$ diverges at an exponential rate of n.

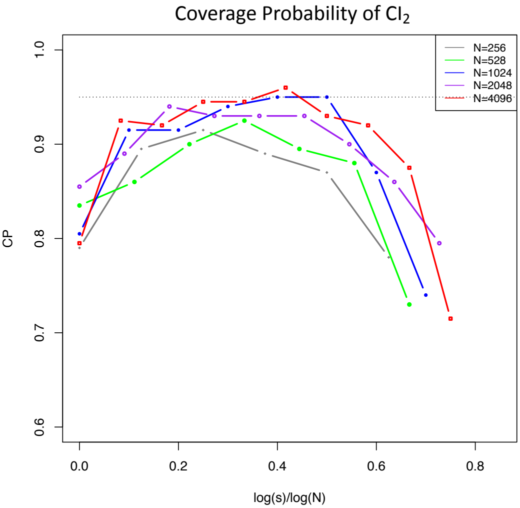


Figure: Coverage probability of 95% confidence interval based on $\check{\beta}$

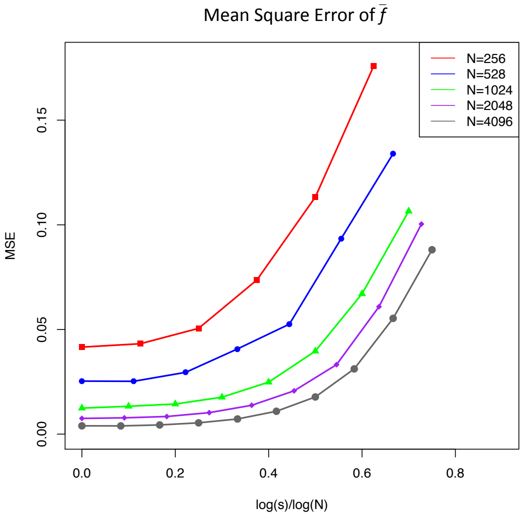


Figure: Mean-square errors of $\bar{f}$ under different choices of N and s

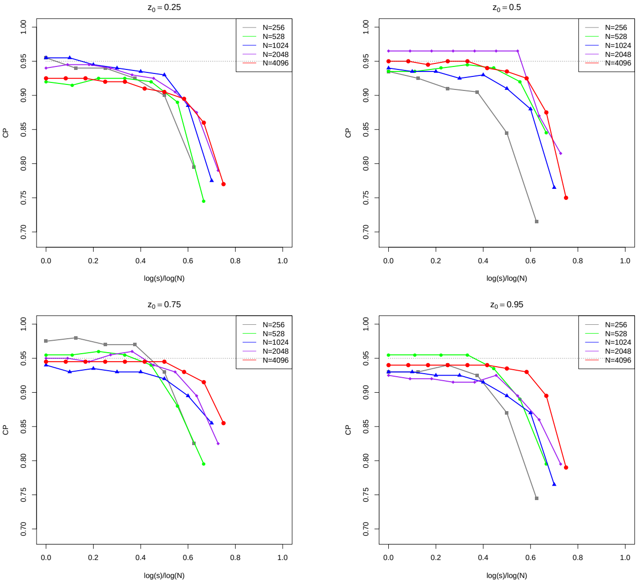


Figure: Coverage probability of 95% predictive interval with different choices of s and N