Semi-Nonparametric Inferences for Massive Data

Guang Cheng[1]

Department of Statistics
Purdue University

Statistics Seminar at NCSU
October, 2015

[1]Acknowledge NSF, Simons Foundation and ONR. A Joint Work with Tianqi Zhao and Han Liu

The Era of Big Data

At the 2010 Google Atmosphere Convention, Google's CEO Eric Schmidt pointed out that,

"There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days."

No wonder that the era of Big Data has arrived...

Recent News on Big Data

On August 6, 2014, Nature[2] released news: "US Big-Data Health Network Launches Aspirin Study."

- In this $10-million pilot study, the use of aspirin to prevent heart disease will be investigated;
- Participants will take daily doses of aspirin that fall within the range typically prescribed for heart disease, and be monitored to determine whether one dosage works better than the others;
- The health-care data such as insurance claims, blood tests and medical histories will be collected from as many as 30 million people in the United States through PCORnet[3];

---

[2] http://www.nature.com/news/us-big-data-health-network-launches-aspirin-study-1.15675

[3] A network setup by Patient-Centered Outcomes Research (PCOR) Institute for collecting health-care data.

Recent News on Big Data

On August 6, 2014, Nature[2] released news: "US Big-Data Health Network Launches Aspirin Study."

- In this $10-million pilot study, the use of aspirin to prevent heart disease will be investigated;
- Participants will take daily doses of aspirin that fall within the range typically prescribed for heart disease, and be monitored to determine whether one dosage works better than the others;
- The health-care data such as insurance claims, blood tests and medical histories will be collected from as many as 30 million people in the United States through PCORnet[3];

[2] http://www.nature.com/news/us-big-data-health-network-launches-aspirin-study-1.15675

[3] A network setup by Patient-Centered Outcomes Research (PCOR) Institute for collecting health-care data.

Recent News on Big Data

On August 6, 2014, Nature[2] released news: "US Big-Data Health Network Launches Aspirin Study."

- In this $10-million pilot study, the use of aspirin to prevent heart disease will be investigated;
- Participants will take daily doses of aspirin that fall within the range typically prescribed for heart disease, and be monitored to determine whether one dosage works better than the others;
- The health-care data such as insurance claims, blood tests and medical histories will be collected from as many as 30 million people in the United States through PCORnet[3];

[2] http://www.nature.com/news/us-big-data-health-network-launches-aspirin-study-1.15675

[3] A network setup by Patient-Centered Outcomes Research (PCOR) Institute for collecting health-care data.

Recent News on Big Data

On August 6, 2014, Nature[2] released news: "US Big-Data Health Network Launches Aspirin Study."

- In this $10-million pilot study, the use of aspirin to prevent heart disease will be investigated;
- Participants will take daily doses of aspirin that fall within the range typically prescribed for heart disease, and be monitored to determine whether one dosage works better than the others;
- The health-care data such as insurance claims, blood tests and medical histories will be collected from as many as 30 million people in the United States through PCORnet[3];

[2] http://www.nature.com/news/us-big-data-health-network-launches-aspirin-study-1.15675

[3] A network setup by Patient-Centered Outcomes Research (PCOR) Institute for collecting health-care data.

Recent News on Big Data (cont')

- PCORnet will connect multiple smaller networks, giving researchers access to records at a large number of institutions without creating a central data repository;
- This decentralization creates one of the greatest challenges on how to merge and standardize data from different networks to enable accurate comparison;
- The many types of data – scans from medical imaging, vital-signs records and, eventually, genetic information can be messy, and record-keeping systems vary among health-care institutions.

Recent News on Big Data (cont')

- PCORnet will connect multiple smaller networks, giving researchers access to records at a large number of institutions without creating a central data repository;
- This decentralization creates one of the greatest challenges on how to merge and standardize data from different networks to enable accurate comparison;
- The many types of data – scans from medical imaging, vital-signs records and, eventually, genetic information can be messy, and record-keeping systems vary among health-care institutions.

Recent News on Big Data (cont')

- PCORnet will connect multiple smaller networks, giving researchers access to records at a large number of institutions without creating a central data repository;
- This decentralization creates one of the greatest challenges on how to merge and standardize data from different networks to enable accurate comparison;
- The many types of data – scans from medical imaging, vital-signs records and, eventually, genetic information can be messy, and record-keeping systems vary among health-care institutions.

Challenges of Big Data

Motivated by this US health network data, we summarize the features of big data as *4D*:

- Distributed: computation and storage bottleneck;
- Dirty: the curse of heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: non-stationary underlying distribution;
- This talk focuses on "Distributed" and "Dirty".

Challenges of Big Data

Motivated by this US health network data, we summarize the features of big data as *4D*:

- Distributed: computation and storage bottleneck;
- Dirty: the curse of heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: non-stationary underlying distribution;
- This talk focuses on “Distributed” and “Dirty”.

Challenges of Big Data

Motivated by this US health network data, we summarize the features of big data as *4D*:

- Distributed: computation and storage bottleneck;
- Dirty: the curse of heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: non-stationary underlying distribution;
- This talk focuses on "Distributed" and "Dirty".

Challenges of Big Data

Motivated by this US health network data, we summarize the features of big data as *4D*:

- Distributed: computation and storage bottleneck;
- Dirty: the curse of heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: non-stationary underlying distribution;
- This talk focuses on "Distributed" and "Dirty".

Challenges of Big Data

Motivated by this US health network data, we summarize the features of big data as *4D*:

- Distributed: computation and storage bottleneck;
- Dirty: the curse of heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: non-stationary underlying distribution;
- This talk focuses on “Distributed” and “Dirty”.

Challenges of Big Data

Motivated by this US health network data, we summarize the features of big data as *4D*:

- Distributed: computation and storage bottleneck;
- Dirty: the curse of heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: non-stationary underlying distribution;
- This talk focuses on "Distributed" and "Dirty".

General Goal

In the era of massive data, here are my questions of curiosity:

- Can we guarantee a high level of statistical inferential accuracy under a certain computation/time constraint?
- Or what is the least computational cost in obtaining the best possible statistical inferences?
- How to break the curse of heterogeneity by exploiting the commonality information?
- How to perform a large scale heterogeneity testing?

General Goal

In the era of massive data, here are my questions of curiosity:

- Can we guarantee a high level of statistical inferential accuracy under a certain computation/time constraint?
- Or what is the least computational cost in obtaining the best possible statistical inferences?
- How to break the curse of heterogeneity by exploiting the commonality information?
- How to perform a large scale heterogeneity testing?

General Goal

In the era of massive data, here are my questions of curiosity:

- Can we guarantee a high level of statistical inferential accuracy under a certain computation/time constraint?
- Or what is the least computational cost in obtaining the best possible statistical inferences?
- How to break the curse of heterogeneity by exploiting the commonality information?
- How to perform a large scale heterogeneity testing?

General Goal

In the era of massive data, here are my questions of curiosity:

- Can we guarantee a high level of statistical inferential accuracy under a certain computation/time constraint?
- Or what is the least computational cost in obtaining the best possible statistical inferences?
- How to break the curse of heterogeneity by exploiting the commonality information?
- How to perform a large scale heterogeneity testing?

General Goal

In the era of massive data, here are my questions of curiosity:

- Can we guarantee a high level of statistical inferential accuracy under a certain computation/time constraint?
- Or what is the least computational cost in obtaining the best possible statistical inferences?
- How to break the curse of heterogeneity by exploiting the commonality information?
- How to perform a large scale heterogeneity testing?

ORACLE RULE FOR MASSIVE DATA IS THE KEY[4].

[4] Simplified technical results are presented for better delivering insights.

Part I: Homogeneous Data

Outline

1. Divide-and-Conquer Strategy
2. Kernel Ridge Regression
3. Nonparametric Inference
4. Simulations

Divide-and-Conquer Approach

- Consider a univariate 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 such as $\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)}$.

Divide-and-Conquer Approach

- Consider a univariate 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 such as $\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)}$.

Divide-and-Conquer Approach

- Consider a univariate 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 such as $\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)}$.

Divide-and-Conquer Approach

- Consider a univariate 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 such as $\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)}$.

Divide-and-Conquer Approach

- Consider a univariate 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 such as $\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)}$.

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, JRSS-B) and Bayesian (Wang and Dunson, 2014, Arxiv), or nonparametric minimaxity (Zhang et al, 2014, Arxiv);
- Semi/nonparametric inferences for massive data still remain untouched (although they are crucially important in evaluating reproducibility in modern scientific studies).

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, JRSS-B) and Bayesian (Wang and Dunson, 2014, Arxiv), or nonparametric minimaxity (Zhang et al, 2014, Arxiv);
- Semi/nonparametric inferences for massive data still remain untouched (although they are crucially important in evaluating reproducibility in modern scientific studies).

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

- In theory, we want to derive a theoretical upper bound for 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;
- Allowing $s \to \infty$ significantly complicates the traditional theoretical analysis.

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

- In theory, we want to derive a theoretical upper bound for 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;
- Allowing $s \to \infty$ significantly complicates the traditional theoretical analysis.

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

- In theory, we want to derive a theoretical upper bound for 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;
- Allowing $s \to \infty$ significantly complicates the traditional 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 $\boldsymbol{\alpha} = (K + \lambda n I)^{-1} \boldsymbol{y}$.
- 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 $\boldsymbol{\alpha} = (K + \lambda n I)^{-1} \boldsymbol{y}$.
- 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 $\boldsymbol{\alpha} = (K + \lambda n I)^{-1} \boldsymbol{y}$.
- 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 $\boldsymbol{\alpha} = (K + \lambda n I)^{-1}\boldsymbol{y}$.
- 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

The decay rate of μ_k characterizes the smoothness of f.

- 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).

Commonly Used Kernels

The decay rate of μ_k characterizes the smoothness of f.

- 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).

Commonly Used Kernels

The decay rate of μ_k characterizes the smoothness of f.

- 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).

Commonly Used Kernels

The decay rate of μ_k characterizes the smoothness of f.

- 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).

Commonly Used Kernels

The decay rate of μ_k characterizes the smoothness of f.

- 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).

Commonly Used Kernels

The decay rate of μ_k characterizes the smoothness of f.

- 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).

Commonly Used Kernels

The decay rate of μ_k characterizes the smoothness of f.

- 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).

Commonly Used Kernels

The decay rate of μ_k characterizes the smoothness of f.

- 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).

Local Confidence Interval[5]

Theorem 1. Suppose regularity conditions on ϵ, $K(\cdot,\cdot)$ and $\phi_j(\cdot)$ hold, e.g., tail condition on ϵ and $\sup_j \|\phi_j\|_\infty \le C_\phi$. 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 that determine s and λ.

[5]Simultaneous confidence band result delivers similar theoretical insights

Examples

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

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$ and $\log(\lambda^{-1}) = o(\log^2 N)$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$ and $\log(\lambda^{-1}) = o(\log^2(N))$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$.
- Choose λ as if working on the entire dataset with sample size N. Hence, the standard generalized cross validation method (applied to each subsample) fails in this case.

Examples

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

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$ and $\log(\lambda^{-1}) = o(\log^2 N)$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$ and $\log(\lambda^{-1}) = o(\log^2(N))$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$.
- Choose λ as if working on the entire dataset with sample size N. Hence, the standard generalized cross validation method (applied to each subsample) fails in this case.

Examples

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

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$ and $\log(\lambda^{-1}) = o(\log^2 N)$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$ and $\log(\lambda^{-1}) = o(\log^2(N))$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$.
- Choose λ as if working on the entire dataset with sample size N. Hence, the standard generalized cross validation method (applied to each subsample) fails in this case.

Examples

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

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$ and $\log(\lambda^{-1}) = o(\log^2 N)$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$ and $\log(\lambda^{-1}) = o(\log^2(N))$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$.
- Choose λ as if working on the entire dataset with sample size N. Hence, the standard generalized cross validation method (applied to each subsample) fails in this case.

Examples

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

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$ and $\log(\lambda^{-1}) = o(\log^2 N)$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$ and $\log(\lambda^{-1}) = o(\log^2(N))$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$.
- Choose λ as if working on the entire dataset with sample size N. Hence, the standard generalized cross validation method (applied to each subsample) fails in this case.

Examples

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

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$ and $\log(\lambda^{-1}) = o(\log^2 N)$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$ and $\log(\lambda^{-1}) = o(\log^2(N))$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$.
- Choose λ as if working on the entire dataset with sample size N. Hence, the standard generalized cross validation method (applied to each subsample) fails in this case.

Examples

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

- Finite Rank (with a rank r):
  - $\lambda = o(N^{-1/2})$ and $\log(\lambda^{-1}) = o(\log^2 N)$;
- Exponential Decay (with a power p):
  - $\lambda = o((\log N)^{1/(2p)}/\sqrt{N})$ and $\log(\lambda^{-1}) = o(\log^2(N))$;
- Polynomial Decay (with a power $m > 1/2$):
  - $\lambda \asymp N^{-d}$ for some $2m/(4m+1) < d < 4m^2/(8m-1)$.
- Choose λ as if working on the entire dataset with sample size N. Hence, the standard generalized cross validation method (applied to each subsample) fails in this case.

Specifically, we have the following upper bounds for s:

- For finite rank kernel (with any finite rank r),

$$s = O(N^{\gamma}) \text{ for any } \gamma < 1/2;$$

- For exponential decay kernel (with any finite power p),

$$s = O(N^{\gamma'}) \text{ for any } \gamma' < \gamma < 1/2;$$

- 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),

$$s = O(N^{\gamma}) \text{ for any } \gamma < 1/2;$$

- For exponential decay kernel (with any finite power p),

$$s = O(N^{\gamma'}) \text{ for any } \gamma' < \gamma < 1/2;$$

- 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),

$$s = O(N^{\gamma}) \text{ for any } \gamma < 1/2;$$

- For exponential decay kernel (with any finite power p),

$$s = O(N^{\gamma'}) \text{ for any } \gamma' < \gamma < 1/2;$$

- For polynomial decay kernel (with $m = 2$),

$$s = o(N^{4/27}) \approx o(N^{0.29}).$$

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

- 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:
  - $\widetilde{PLRT}_{N,\lambda} = (1/s) \sum_{j=1}^{s} PLRT_{n,\lambda}^{(j)}$;
  - $P\widehat{LRT}_{N,\lambda} = \mathcal{L}_{N,\lambda}(\bar{f}_N) - \mathcal{L}_{N,\lambda}(f_0)$;

Penalized Likelihood Ratio Test

Theorem 2. 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 required for consistency.

- An additional big data insight: we have to sacrifice certain amount of computational efficiency (avoid choosing the largest possible s) for obtaining the optimality.

Penalized Likelihood Ratio Test

Theorem 2. 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 required for consistency.

- An additional big data insight: we have to sacrifice certain amount of computational efficiency (avoid choosing the largest possible s) for obtaining the optimality.

Summary

- Big Data Insights:
  - Oracle rule holds when s does not grow too fast;
  - choose the smoothing parameter as if not splitting the data;
  - sacrifice computational efficiency for obtaining optimality.
- Key technical tool: Functional Bahadur Representation in Shang and C. (2013, AoS).

Summary

- Big Data Insights:
  - Oracle rule holds when s does not grow too fast;
  - choose the smoothing parameter as if not splitting the data;
  - sacrifice computational efficiency for obtaining optimality.
- Key technical tool: Functional Bahadur Representation in Shang and C. (2013, AoS).

Summary

- Big Data Insights:
    - Oracle rule holds when s does not grow too fast;
    - choose the smoothing parameter as if not splitting the data;
    - sacrifice computational efficiency for obtaining optimality.
- Key technical tool: Functional Bahadur Representation in Shang and C. (2013, AoS).

Summary

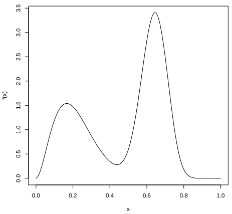
- Big Data Insights:
  - Oracle rule holds when s does not grow too fast;
  - choose the smoothing parameter as if not splitting the data;
  - sacrifice computational efficiency for obtaining optimality.
- Key technical tool: Functional Bahadur Representation in Shang and C. (2013, AoS).

Summary

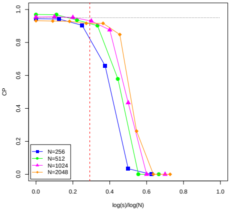
- Big Data Insights:
  - Oracle rule holds when s does not grow too fast;
  - choose the smoothing parameter as if not splitting the data;
  - sacrifice computational efficiency for obtaining optimality.
- Key technical tool: Functional Bahadur Representation in Shang and C. (2013, AoS).

Phase Transition of Coverage Probability

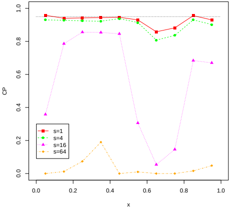
(a) True function



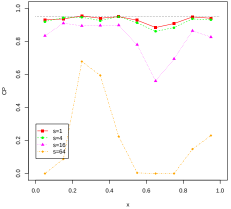
(b) CPs at $x_0 = 0.5$



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



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



Phase Transition of Mean Squared Error

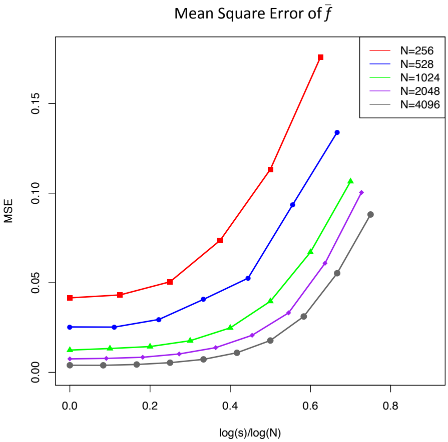


Figure: Mean-square errors of $\bar{f}_N$ under different choices of N and s

Part II: Heterogeneous Data

Outline

Revisit US Health Data

Let us revisit the news on US Big-Data Health Network.

- Different networks such as US hospitals conduct the same clinical trial on the relation between a response variable Y i.e., heart disease, and a set of predictors $Z, X_1, X_2, \ldots, X_p$ including the dosage of aspirin;
- Medical knowledge 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 (linear) relations with Y to potentially vary in different networks (located in different areas). For example, the genetic functionality of different races might be heterogenous;
- The linear relation is assumed here for simplicity, and particularly suitable when the covariates are discrete such as the dosage of aspirin, e.g., 1 or 2 tablets each day.

Revisit US Health Data

Let us revisit the news on US Big-Data Health Network.

- Different networks such as US hospitals conduct the same clinical trial on the relation between a response variable Y i.e., heart disease, and a set of predictors $Z, X_1, X_2, \ldots, X_p$ including the dosage of aspirin;
- Medical knowledge 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 (linear) relations with Y to potentially vary in different networks (located in different areas). For example, the genetic functionality of different races might be heterogenous;
- The linear relation is assumed here for simplicity, and particularly suitable when the covariates are discrete such as the dosage of aspirin, e.g., 1 or 2 tablets each day.

Revisit US Health Data

Let us revisit the news on US Big-Data Health Network.

- Different networks such as US hospitals conduct the same clinical trial on the relation between a response variable Y i.e., heart disease, and a set of predictors $Z, X_1, X_2, \ldots, X_p$ including the dosage of aspirin;
- Medical knowledge 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 (linear) relations with Y to potentially vary in different networks (located in different areas). For example, the genetic functionality of different races might be heterogenous;
- The linear relation is assumed here for simplicity, and particularly suitable when the covariates are discrete such as the dosage of aspirin, e.g., 1 or 2 tablets each day.

Revisit US Health Data

Let us revisit the news on US Big-Data Health Network.

- Different networks such as US hospitals conduct the same clinical trial on the relation between a response variable Y i.e., heart disease, and a set of predictors $Z, X_1, X_2, \ldots, X_p$ including the dosage of aspirin;
- Medical knowledge 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 (linear) relations with Y to potentially vary in different networks (located in different areas). For example, the genetic functionality of different races might be heterogenous;
- The linear relation is assumed here for simplicity, and particularly suitable when the covariates are discrete such as the dosage of aspirin, e.g., 1 or 2 tablets each day.

Revisit US Health Data

Let us revisit the news on US Big-Data Health Network.

- Different networks such as US hospitals conduct the same clinical trial on the relation between a response variable Y i.e., heart disease, and a set of predictors $Z, X_1, X_2, \ldots, X_p$ including the dosage of aspirin;
- Medical knowledge 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 (linear) relations with Y to potentially vary in different networks (located in different areas). For example, the genetic functionality of different races might be heterogenous;
- The linear relation is assumed here for simplicity, and particularly suitable when the covariates are discrete such as the dosage of aspirin, e.g., 1 or 2 tablets each day.

A Partially Linear Modelling

- 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 C. and Shang, 2015, AoS) since $\beta^{(j)}$ and f are both of interest.

A Partially Linear Modelling

- 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 C. and Shang, 2015, AoS) since $\boldsymbol{\beta}^{(j)}$ and f are both of interest.

A Partially Linear Modelling

- 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 C. and Shang, 2015, AoS) since $\boldsymbol{\beta}^{(j)}$ and f are both of interest.

A Partially Linear Modelling

- 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 C. and Shang, 2015, AoS) since $\boldsymbol{\beta}^{(j)}$ and f are both of interest.

Estimation Procedure

- 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:

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

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

Estimation Procedure

- 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:

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

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

Estimation Procedure

- 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

- Individual estimation in the j-th subpopulation:

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

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

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

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

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;
- However, when analyzing heterogeneous data, our major interest[1] is about how to efficiently extract common features across many subpopulations while exploring heterogeneity of each subpopulation as $s \to \infty$;
- Therefore, comparisons between $(\check{\beta}_n^{(j)}, \bar{f}_N)$ and oracle estimate (in terms of both risk and limit distribution) would be needed.

[1] D&C can be applied to the sub-population with large sample size.

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;
- However, when analyzing heterogeneous data, our major interest[1] is about how to efficiently extract common features across many subpopulations while exploring heterogeneity of each subpopulation as $s \to \infty$;
- Therefore, comparisons between $(\check{\beta}_n^{(j)}, \bar{f}_N)$ and oracle estimate (in terms of both risk and limit distribution) would be needed.

[1] D&C can be applied to the sub-population with large sample size.

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;
- However, when analyzing heterogeneous data, our major interest[1] is about how to efficiently extract common features across many subpopulations while exploring heterogeneity of each subpopulation as $s \to \infty$;
- Therefore, comparisons between $(\check{\beta}_n^{(j)}, \bar{f}_N)$ and oracle estimate (in terms of both risk and limit distribution) would be needed.

[1] D&C can be applied to the sub-population with large sample size.

Oracle Estimate

We define the oracle estimate for f as if the heterogeneity information $\boldsymbol{\beta}_j$ were known:

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

The oracle estimate for $\boldsymbol{\beta}_j$ can be defined similarly:

$$\widehat{\boldsymbol{\beta}}_{or}^{(j)} = \operatorname*{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i^{(j)} - (\boldsymbol{\beta}^{(j)})^T \mathbf{X}_i^{(j)} - f_0(Z_i^{(j)}) \right)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}.$$

A Preliminary Result: Joint Asymptotics

Theorem 3. Given proper $s \to \infty$[2] and $\lambda \to 0$, we have[3]

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

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

[2] The asymptotic independence between $\widehat{\boldsymbol{\beta}}_n^{(j)}$ and $\bar{f}_N(z_0)$ is mainly due to the fact that $n/N = s^{-1} \to 0$.

[3] The asymptotic variance Σ_{22} of $\bar{f}_N$ is the same as that of $\widehat{f}_{or}$.

Efficiency Boosting

- Theorem 4 implies that $\widehat{\boldsymbol{\beta}}_n^{(j)}$ is semiparametric efficient:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{(j)} - \boldsymbol{\beta}_0) \rightsquigarrow N(0, \sigma^2(E(\mathbf{X} - E(\mathbf{X}|Z))^{\otimes 2})^{-1}).$$

- We next illustrate an important feature of massive data: strength-borrowing. That is, the aggregation of commonality in turn boosts the estimation efficiency of $\widehat{\boldsymbol{\beta}}_n^{(j)}$ from semiparametric level to parametric level.
- By imposing a lower bound on s (such that strength are borrowed from a sufficient number of sub-populations), we show that[4]

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

as if the commonality information were available.

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

Efficiency Boosting

- Theorem 4 implies that $\widehat{\boldsymbol{\beta}}_n^{(j)}$ is semiparametric efficient:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{(j)} - \boldsymbol{\beta}_0) \rightsquigarrow N(0, \sigma^2(E(\mathbf{X} - E(\mathbf{X}|Z))^{\otimes 2})^{-1}).$$

- We next illustrate an important feature of massive data: strength-borrowing. That is, the aggregation of commonality in turn boosts the estimation efficiency of $\widehat{\boldsymbol{\beta}}_n^{(j)}$ from semiparametric level to parametric level.

- By imposing a lower bound on s (such that strength are borrowed from a sufficient number of sub-populations), we show that[4]

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

as if the commonality information were available.

[4] Recall that $\check{\boldsymbol{\beta}}_n^{(j)} = \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \left(Y_i^{(j)} - \boldsymbol{\beta}^T\mathbf{X}_i^{(j)} - \bar{f}_N(Z_i^{(j)})\right)^2$.

Efficiency Boosting

- Theorem 4 implies that $\widehat{\boldsymbol{\beta}}_n^{(j)}$ is semiparametric efficient:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{(j)} - \boldsymbol{\beta}_0) \rightsquigarrow N(0, \sigma^2 (E(\mathbf{X} - E(\mathbf{X}|Z))^{\otimes 2})^{-1}).$$

- We next illustrate an important feature of massive data: strength-borrowing. That is, the aggregation of commonality in turn boosts the estimation efficiency of $\widehat{\boldsymbol{\beta}}_n^{(j)}$ from semiparametric level to parametric level.
- By imposing a lower bound on s (such that strength are borrowed from a sufficient number of sub-populations), we show that[4]

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

as if the commonality information were available.

[4] Recall that $\check{\boldsymbol{\beta}}_n^{(j)} = \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \left(Y_i^{(j)} - \boldsymbol{\beta}^T \mathbf{X}_i^{(j)} - \bar{f}_N(Z_i^{(j)})\right)^2$.

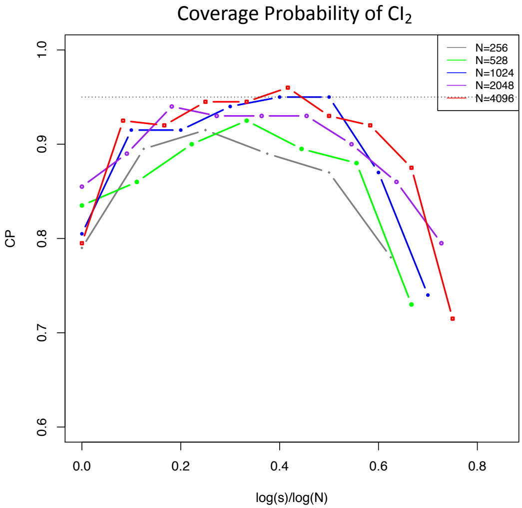


Figure: Coverage probability of 95% confidence interval based on $\check{\boldsymbol{\beta}}_n^{(j)}$

Large Scale Heterogeneity Testing

- Consider a *high dimensional* simultaneous testing:

$$H_0 : \boldsymbol{\beta}^{(j)} = \widetilde{\boldsymbol{\beta}}^{(j)} \text{ for all } j \in J, \tag{2}$$

where $J \subset \{1, 2, \ldots, s\}$ and $|J| \to \infty$, versus

$$H_1 : \boldsymbol{\beta}^{(j)} \neq \widetilde{\boldsymbol{\beta}}^{(j)} \text{ for some } j \in J; \tag{3}$$

- Test statistic:

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

- We can consistently approximate the quantile of the null distribution via bootstrap even when $|J|$ diverges at an exponential rate of n by a nontrivial application of a recent Gaussian approximation theory.

Large Scale Heterogeneity Testing

- Consider a *high dimensional* simultaneous testing:

$$H_0 : \boldsymbol{\beta}^{(j)} = \widetilde{\boldsymbol{\beta}}^{(j)} \text{ for all } j \in J, \tag{2}$$

where $J \subset \{1, 2, \ldots, s\}$ and $|J| \to \infty$, versus

$$H_1 : \boldsymbol{\beta}^{(j)} \neq \widetilde{\boldsymbol{\beta}}^{(j)} \text{ for some } j \in J; \tag{3}$$

- Test statistic:

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

- We can consistently approximate the quantile of the null distribution via bootstrap even when $|J|$ diverges at an exponential rate of n by a nontrivial application of a recent Gaussian approximation theory.

Large Scale Heterogeneity Testing

- Consider a *high dimensional* simultaneous testing:

$$H_0 : \boldsymbol{\beta}^{(j)} = \widetilde{\boldsymbol{\beta}}^{(j)} \text{ for all } j \in J, \tag{2}$$

where $J \subset \{1, 2, \ldots, s\}$ and $|J| \to \infty$, versus

$$H_1 : \boldsymbol{\beta}^{(j)} \neq \widetilde{\boldsymbol{\beta}}^{(j)} \text{ for some } j \in J; \tag{3}$$

- Test statistic:

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

- We can consistently approximate the quantile of the null distribution via bootstrap even when $|J|$ diverges at an exponential rate of n by a nontrivial application of a recent Gaussian approximation theory.

Thank You!