Semi-Nonparametric Inference for Massive Data

Guang Cheng\textsuperscript{1}

Department of Statistics
Purdue University

Statistics Seminar at Univ of California, Davis
January 26, 2015

\textsuperscript{1}Acknowledge NSF, Simons Foundation and Princeton
The massive sample size of Big Data introduces unique computational and statistical challenges summarized as 4Ds:

- 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

The massive sample size of Big Data introduces unique computational and statistical challenges summarized as 4Ds:

- 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”.
The massive sample size of Big Data introduces unique computational and statistical challenges summarized as 4Ds:

- **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”.
The massive sample size of Big Data introduces unique computational and statistical challenges summarized as 4Ds:

- 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”.
The massive sample size of Big Data introduces unique computational and statistical challenges summarized as 4Ds:

- 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”.
The massive sample size of Big Data introduces unique computational and statistical challenges summarized as *4Ds*:

- 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 does model regularity affect the computational cost?
- 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 does model regularity affect the computational cost?
- 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 does model regularity affect the computational cost?
- 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 does model regularity affect the computational cost?
- How to break the curse of heterogeneity by exploiting the commonality information?
- How to perform a large scale heterogeneity testing?
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 does model regularity affect the computational cost?
- How to break the curse of heterogeneity by exploiting the commonality information?
- How to perform a large scale heterogeneity testing?
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 does model regularity affect the computational cost?
- 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.\(^2\)

\(^2\)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:

\[ P_j = \{X_1^{(j)}, \ldots, X_n^{(j)}\} \implies \hat{f}_n^{(j)}; \]

- Aggregation such as \( \bar{f}_N = (1/s) \sum_{j=1}^{s} \hat{f}_n^{(j)}. \)
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:

\[ P_j = \{X_1^{(j)}, \ldots, X_n^{(j)}\} \implies \hat{f}_n^{(j)}; \]

Aggregation such as \( \bar{f}_N = (1/s) \sum_{j=1}^{s} \hat{f}_n^{(j)} \).
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:

\[ P_j = \{ X_1^{(j)}, \ldots, X_n^{(j)} \} \implies \hat{f}_n^{(j)}; \]

- Aggregation such as \( \bar{f}_N = (1/s) \sum_{j=1}^{s} \hat{f}_n^{(j)}. \)
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:
  \[ P_j = \{ X_1^{(j)}, \ldots, X_n^{(j)} \} \implies \hat{f}_n^{(j)}; \]
- Aggregation such as \( \bar{f}_N = (1/s) \sum_{j=1}^{s} \hat{f}_n^{(j)}. \)
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:

\[ P_j = \{X_1^{(j)}, \ldots, X_n^{(j)}\} \rightarrow \hat{f}_n^{(j)}; \]

- Aggregation such as \( \bar{f}_N = (1/s) \sum_{j=1}^{s} \hat{f}_n^{(j)} \).
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 \( (s \to \infty \text{ as } N \to \infty) \)

- In theory, we want to derive the largest possible diverging rate of \( s \) under which the following oracle rule holds:
  
  \[ \text{the nonparametric inferences constructed based on } \bar{f}_N \text{ are (asymp.) the same as those on the oracle estimator } \hat{f}_N. \]

- 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 the traditional theoretical analysis.
In theory, 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 $\hat{f}_N$.”

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 the traditional theoretical analysis.
Splitotics Theory ($s \to \infty$ as $N \to \infty$)

- In theory, 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 $\hat{f}_N$.”

- 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 the traditional theoretical analysis.
Splitotics Theory \((s \to \infty \text{ as } N \to \infty)\)

- In theory, 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 \(\hat{f}_N\).”

- 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 the traditional theoretical analysis.
Splitotics Theory \((s \rightarrow \infty \text{ as } N \rightarrow \infty)\)

- In theory, 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 \(\hat{f}_N\).”

- 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 \(s \rightarrow \infty\) significantly complicates the traditional theoretical analysis.
Kernel Ridge Regression (KRR)

- Define the KRR estimate \( \hat{f} : \mathbb{R}^1 \to \mathbb{R}^1 \) as

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

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, \( \mu_i \)'s are eigenvalues and \( \phi_i(\cdot) \)'s are eigenfunctions.

- Explicitly, \( \hat{f}_n(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) \) with \( \alpha = (K + \lambda I)^{-1} y \).

- 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 $\hat{f}: \mathbb{R}^1 \mapsto \mathbb{R}^1$ as

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

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, $\mu_i$’s are eigenvalues and $\phi_i(\cdot)$’s are eigenfunctions.

- Explicitly, $\hat{f}_n(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ with $\alpha = (K + \lambda I)^{-1} y$.

- 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 \( \hat{f} : \mathbb{R}^1 \mapsto \mathbb{R}^1 \) as

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

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, \( \mu_i \)'s are eigenvalues and \( \phi_i(\cdot) \)'s are eigenfunctions.

- Explicitly, \( \hat{f}_n(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) \) with \( \alpha = (K + \lambda I)^{-1} y \).
- 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 $\hat{f} : \mathbb{R}^1 \mapsto \mathbb{R}^1$ as

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

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, $\mu_i$’s are eigenvalues and $\phi_i(\cdot)$’s are eigenfunctions.

- Explicitly, $\hat{f}_n(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ with $\alpha = (K + \lambda I)^{-1} y$.

- 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.
The decay rate of $\mu_k$ 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;
Commonly Used Kernels

The decay rate of $\mu_k$ 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;
Commonly Used Kernels

The decay rate of $\mu_k$ 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 ($\mu_k \approx \exp(-\alpha k^p)$ for some $\alpha, p > 0$):**
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ for $p = 2$;
- **Polynomial Decay ($\mu_k \approx k^{-2m}$ 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;
Commonly Used Kernels

The decay rate of $\mu_k$ 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** ($\mu_k \approx \exp(-\alpha k^p)$ for some $\alpha, p > 0$):
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ for $p = 2$;
- **Polynomial Decay** ($\mu_k \approx k^{-2m}$ 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;
The decay rate of $\mu_k$ 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** ($\mu_k \propto \exp(-\alpha k^p)$ for some $\alpha, p > 0$):
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'|^2/\sigma^2)$ for $p = 2$;
- **Polynomial Decay** ($\mu_k \propto k^{-2m}$ 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;
Commonly Used Kernels

The decay rate of $\mu_k$ 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** ($\mu_k \approx \exp(-\alpha k^p)$ for some $\alpha, p > 0$):
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ for $p = 2$;
- **Polynomial Decay** ($\mu_k \approx k^{-2m}$ 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;
Commonly Used Kernels

The decay rate of $\mu_k$ 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** ($\mu_k \approx \exp(-\alpha k^p)$ for some $\alpha, p > 0$):
  - Gaussian kernel $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$ for $p = 2$;
- **Polynomial Decay** ($\mu_k \approx k^{-2m}$ 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;
The decay rate of $\mu_k$ 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;
Theorem 1. Suppose regularity conditions on $\epsilon$, $K(\cdot, \cdot)$ and $\phi_j(\cdot)$ hold, e.g., tail condition on $\epsilon$ 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 $x_0 \in \mathcal{X}$,

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

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 $\sigma_{x_0}^2$ are the same as those of $\hat{f}_N$ (based on the entire dataset). Hence, the oracle property of the local confidence interval holds under the above conditions that determine $s$ and $\lambda$.

---

3Simultaneous confidence band result delivers similar theoretical insights
In Theorem 1, some under-smoothing condition is implicitly assumed (so, there is no estimation bias).

Technical Challenges:

- the first set of statistical inferences for KRR by generalizing the functional Bahadur representation developed for smoothing spline estimation (Shang and C., 2013, AoS);
- employ empirical process theory to study an 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:

- the first set of statistical inferences for KRR by generalizing the functional Bahadur representation developed for smoothing spline estimation (Shang and C., 2013, AoS);
- employ empirical process theory to study an 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:
- the first set of statistical inferences for KRR by generalizing the functional Bahadur representation developed for smoothing spline estimation (Shang and C., 2013, AoS);
- employ empirical process theory to study an 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:

- the first set of statistical inferences for KRR by generalizing the functional Bahadur representation developed for smoothing spline estimation (Shang and C., 2013, AoS);
- employ empirical process theory to study an average of $s$ asymptotic linear expansions as $s \to \infty$. 
Examples

The oracle property of local confidence interval holds under the following conditions on $\lambda$ 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 $\gamma < 1/2 - (8m - 1)/(8m^2)$. 
Examples

The oracle property of local confidence interval holds under the following conditions on \( \lambda \) and \( s \):

- **Finite Rank (with a rank \( r \)):**
  \[
  \lambda = o(N^{-1/2}), \quad \log(\lambda^{-1}) = o(\log^2 N) \quad \text{and} \quad s = o(N^{1/2}/\{\log^{1/2}(\lambda^{-1}) \log^3(N)\});
  \]

- **Exponential Decay (with a power \( p \)):**
  \[
  \lambda = o((\log N)^{1/(2p)}/\sqrt{N}), \quad \log(\lambda^{-1}) = o(\log^2 (N)) \quad \text{and} \quad s = o(N^{1/2} h^{3/2} /\{[\log(h/\lambda)]^{(p+1)/2p} \log^3(N)\}) \quad \text{with} \quad h = \lfloor \log(1/\lambda) \rfloor^{-1/p};
  \]

- **Polynomial Decay (with a power \( m > 1/2 \)):**
  \[
  \lambda \asymp N^{-d} \quad \text{for some} \quad 2m/(4m + 1) < d < 4m^2/(8m - 1) \quad \text{and} \quad s = N^{\gamma} \quad \text{with} \quad \gamma < 1/2 - (8m - 1)/(8m^2)d.
  \]
Examples

The oracle property of local confidence interval holds under the following conditions on $\lambda$ 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 $m > 1/2$):**
  - $\lambda \approx N^{-d}$ 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 $\lambda$ 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 $h = \lfloor \log(1/\lambda) \rfloor^{-1/p}$;

- **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 $\gamma < 1/2 - (8m - 1)/(8m^2)d$. 
The oracle property of local confidence interval holds under the following conditions on $\lambda$ 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 $\gamma < 1/2 - (8m - 1)/(8m^2)d$. 
The oracle property of local confidence interval holds under the following conditions on $\lambda$ and $s$:

- **Finite Rank (with a rank $r$):**
  
  \[ \lambda = o(N^{-1/2}), \quad \log(\lambda^{-1}) = o(\log^2 N) \quad \text{and} \quad s = o(N^{1/2}/\{\log^{1/2}(\lambda^{-1}) \log^3(N)\}); \]

- **Exponential Decay (with a power $p$):**
  
  \[ \lambda = o((\log N)^{1/(2p)}/\sqrt{N}), \quad \log(\lambda^{-1}) = o(\log^2(N)) \quad \text{and} \quad s = o(N^{1/2}h^{3/2}/\{[\log(h/\lambda)]^{(p+1)/2p}) \log^3(N)\}) \quad \text{with} \quad h = [\log(1/\lambda)]^{-1/p}; \]

- **Polynomial Decay (with a power $m > 1/2$):**
  
  \[ \lambda \asymp N^{-d} \quad \text{for some} \quad 2m/(4m + 1) < d < 4m^2/(8m - 1) \quad \text{and} \quad s = N^\gamma \quad \text{with} \quad \gamma < 1/2 - (8m - 1)/(8m^2)d. \]
Specifically, we have the following upper bounds for $s$:

- For finite rank kernel (with any finite rank $r$),
  $$s = O(N^\gamma) \text{ for any } \gamma < 1/2;$$

- 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 $m = 2$),
  \[ s = o(N^{4/27}) \approx o(N^{0.29}). \]
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 $\lambda$:
Choose $\lambda$ as if working on the entire dataset with sample size $N$ although it is sub-optimal for each sub-estimation\(^4\).

This theoretical finding leads to a modified GCV formula used in practice.

\(^4\)Similar result holds for minimaxity study (Zhang et al, 2014, Arxiv)
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 $\lambda$:
Choose $\lambda$ as if working on the entire dataset with sample size $N$ although it is sub-optimal for each sub-estimation\(^4\).

This theoretical finding leads to a modified GCV formula used in practice.

\(^4\)Similar result holds for minimaxity study (Zhang et al, 2014, Arxiv)
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 $\lambda$:
Choose $\lambda$ as if working on the entire dataset with sample size $N$ although it is sub-optimal for each sub-estimation$^4$.

This theoretical finding leads to a modified GCV formula used in practice.

$^4$Similar result holds for minimaxity study (Zhang et al, 2014, Arxiv)
Penalized Likelihood Ratio Test

- Consider the following test:

\[ H_0 : f = f_0 \quad \text{v.s.} \quad H_1 : f \neq f_0, \]

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

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

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

- \( \hat{PLRT}_{N,\lambda} = (1/s) \sum_{j=1}^{s} PLRT_{n,\lambda}^{(j)} \);
- \( PLRT_{N,\lambda} = \mathcal{L}_{N,\lambda}(\bar{f}_N) - \mathcal{L}_{N,\lambda}(f_0) ; \)
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:
  - \( \hat{PLRT}_{N,\lambda} = (1/s) \sum_{j=1}^{s} PLRT_{n,\lambda}^{(j)}; \)
  - \( \widetilde{PLRT}_{N,\lambda} = \mathcal{L}_{N,\lambda}(\tilde{f}_N) - \mathcal{L}_{N,\lambda}(f_0); \)
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 \( L_{N,\lambda} \) 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:
  - \( \overline{PLRT}_{N,\lambda} = (1/s) \sum_{j=1}^{s} PLRT_{n,\lambda}^{(j)}; \)
  - \( PLRT_{N,\lambda} = L_{N,\lambda}(\bar{f}_N) - L_{N,\lambda}(f_0); \)
Penalized Likelihood Ratio Test

**Theorem 2.** We prove that $\widehat{PLRT}_{N,\lambda}$ and $\widetilde{PLRT}_{N,\lambda}$ 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.
Theorem 2. We prove that $\widehat{PLRT}_{N,\lambda}$ and $\widehat{\tilde{PLRT}}_{N,\lambda}$ 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.
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

- 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

- 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

- **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

- **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

- **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

1. A Partially Linear Modelling
2. Non-Asymptotic Bound
3. Efficiency Boosting
4. Heterogeneity Testing
A Motivating Example

- It is very common that different biology labs (around the world) sometimes conduct the same experiment for verifying the reproducibility of some scientific conclusions;
- For example, they want to understand 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

- It is very common that different biology labs (around the world) sometimes conduct the same experiment for verifying the reproducibility of some scientific conclusions;
- For example, they want to understand 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

- It is very common that different biology labs (around the world) sometimes conduct the same experiment for verifying the reproducibility of some scientific conclusions;
- For example, they want to understand 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

- It is very common that different biology labs (around the world) sometimes conduct the same experiment for verifying the reproducibility of some scientific conclusions;
- For example, they want to understand 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.
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

$$Y = X^T \beta_0^{(j)} + f_0(Z) + \epsilon,$$  \hspace{2cm} (1)

where $\epsilon$ 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

$$Y = X^T \beta_{0}^{(j)} + f_0(Z) + \epsilon,$$  

where $\epsilon$ 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

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

where $\epsilon$ 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.
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

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

where $\epsilon$ has a sub-Gaussian tail and $\text{Var}(\epsilon) = \sigma^2$;

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.
Estimation Procedure

- Individual estimation in the $j$-th subpopulation:

\[
(\hat{\beta}_n^{(j)}, \hat{f}_n^{(j)}) = \arg\min_{(\beta, f) \in \mathbb{R}^p \times \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i^{(j)} - \beta^T X_i^{(j)} - f(Z_i^{(j)}))^2 + \lambda \| f \|_{\mathcal{H}}^2 \right\};
\]

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

- A plug-in estimate for the $j$-th heterogeneity parameter:

\[
\tilde{\beta}_n^{(j)} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (Y_i^{(j)} - \beta^T X_i^{(j)} - \bar{f}_N(Z_i^{(j)}))^2;
\]

- Our final estimate is \((\tilde{\beta}_n^{(j)}, \bar{f}_N)\).
Estimation Procedure

- Individual estimation in the j-th subpopulation:
  \[
  (\hat{\beta}_n^{(j)}, \hat{f}_n^{(j)}) = \arg\min_{(\beta,f) \in \mathbb{R}^p \times \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( Y_i^{(j)} - \beta^T X_i^{(j)} - f(Z_i^{(j)}) \right)^2 + \lambda \| f \|_{\mathcal{H}}^2 \right\};
  \]

- Aggregation: \( \bar{f}_N = \frac{1}{s} \sum_{j=1}^{s} \hat{f}_n^{(j)} \);

- A plug-in estimate for the j-th heterogeneity parameter:
  \[
  \tilde{\beta}_n^{(j)} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left( Y_i^{(j)} - \beta^T X_i^{(j)} - \bar{f}_N(Z_i^{(j)}) \right)^2;
  \]

- Our final estimate is \( (\tilde{\beta}_n^{(j)}, \bar{f}_N) \).
**Estimation Procedure**

- Individual estimation in the $j$-th subpopulation:

\[
(\hat{\beta}_n^{(j)}, \hat{f}_n^{(j)}) = \arg\min_{(\beta, f) \in \mathbb{R}^p \times \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( Y_i^{(j)} - \beta^T X_i^{(j)} - f(Z_i^{(j)}) \right)^2 + \lambda \| f \|_H^2 \right\};
\]

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

- A plug-in estimate for the $j$-th heterogeneity parameter:

\[
\tilde{\beta}_n^{(j)} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left( Y_i^{(j)} - \beta^T X_i^{(j)} - \bar{f}_N(Z_i^{(j)}) \right)^2;
\]

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

- Individual estimation in the $j$-th subpopulation:

$$
(\hat{\beta}_n^{(j)}, \hat{f}_n^{(j)}) = \arg\min_{(\beta,f) \in \mathbb{R}^p \times \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( Y_i^{(j)} - \beta^T X_i^{(j)} - f(Z_i^{(j)}) \right)^2 + \lambda \| f \|_{\mathcal{H}}^2 \right\};
$$

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

- A plug-in estimate for the $j$-th heterogeneity parameter:

$$
\hat{\beta}_n^{(j)} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left( Y_i^{(j)} - \beta^T X_i^{(j)} - \bar{f}_N(Z_i^{(j)}) \right)^2;
$$

- Our final estimate is $(\hat{\beta}_n^{(j)}, \bar{f}_N)$. 
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, some comparisons between \((\hat{\beta}_n^{(j)}, \bar{f}_N)\) and oracle estimate (in terms of risk and limit distribution) would be needed.

\(^1\)D&C can be applied to the sub-population with large sample size.
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, some comparisons between \((\tilde{\beta}_n^{(j)}, \tilde{f}_N)\) and oracle estimate (in terms of 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, some comparisons between \((\hat{\beta}_n^{(j)}, \bar{f}_N)\) and oracle estimate (in terms of risk and limit distribution) would be needed.

\(^1\)D&C can be applied to the sub-population with large sample size.
We define the oracle estimate for $f$ as if the heterogeneity information $\beta_j$ were known:

$$\hat{f}_{or} = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{N} \sum_{i,j=1}^{n,s} (Y^{(j)}_i - (\beta_0^{(j)})^T X^{(j)}_i - f(Z^{(j)}_i))^2 + \lambda \| f \|_\mathcal{H}^2 \right\}.$$ 

The oracle estimate for $\beta_j$ can be defined similarly:

$$\hat{\beta}_{or}^{(j)} = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y^{(j)}_i - (\beta^{(j)})^T X^{(j)}_i - f_0(Z^{(j)}_i))^2 + \lambda \| f \|_\mathcal{H}^2 \right\}.$$
Develop a finite sample valid upper bound for

\[ \text{MSE}(\bar{f}_N) := \mathbb{E}[\|\bar{f} - f_0\|^2_2]. \]

**Theorem 3.** Suppose regularity conditions, e.g., under-smoothing condition, and \( E(X_k|Z) \in \mathcal{H} \) hold\(^2\). When \( s \) does not grow too fast, then

\[ \text{MSE}(\bar{f}) \leq C_{N,K,\lambda}((Nh)^{-1} + \lambda). \]  \hspace{1cm} (2)

Furthermore, by choosing \( \lambda \asymp (Nh)^{-1} \), \( \bar{f}_N \) possesses the same minimax optimal bound as the oracle estimate \( \hat{f}_{or} \)\(^3\).

\(^2\)This condition is needed for controlling the variance term \((Nh)^{-1}\) in (2).

\(^3\)E.g., \( s = o(N^{9/20} \log^{-4} N) \) and \( \lambda \asymp N^{-4/5} \) for cubic spline.
Figure: Mean-square errors of $\bar{f}_N$ under different choices of $N$ and $s$
Some Comments

- The above theorem presents a non-asymptotic version of "oracle rule" that $\bar{f}_N$ shares the same (un-improvable) minimax optimal bound as the $\hat{f}_{or}$;
- Our next result further shows that $\bar{f}_N$ possesses the same (point-wise) asymptotic distribution as the $\hat{f}_{or}$;
- Therefore, we can conclude that our aggregation procedure is able to "filter out" the heterogeneity in data when $s$ does not grow too fast and $\lambda$ is chosen in the order of $N$. 
Some Comments

- The above theorem presents a non-asymptotic version of “oracle rule” that \( \bar{f}_N \) shares the same (un-improvable) minimax optimal bound as the \( \hat{f}_{or} \);
- Our next result further shows that \( \bar{f}_N \) possesses the same (point-wise) asymptotic distribution as the \( \hat{f}_{or} \);
- Therefore, we can conclude that our aggregation procedure is able to “filter out” the heterogeneity in data when \( s \) does not grow too fast and \( \lambda \) is chosen in the order of \( N \).
Some Comments

- The above theorem presents a non-asymptotic version of “oracle rule” that $\bar{f}_N$ shares the same (un-improvable) minimax optimal bound as the $\hat{f}_{or}$;
- Our next result further shows that $\bar{f}_N$ possesses the same (point-wise) asymptotic distribution as the $\hat{f}_{or}$;
- Therefore, we can conclude that our aggregation procedure is able to “filter out” the heterogeneity in data when $s$ does not grow too fast and $\lambda$ is chosen in the order of $N$. 
A Preliminary Result: Joint Asymptotics

**Theorem 4.** Assume similar conditions as in Theorem 3. Given proper $s \to \infty^4$ and $\lambda \to 0$, we have$^5$

$$
\left( \frac{\sqrt{n}(\hat{\beta}_n^{(j)} - \beta_0^{(j)})}{\sqrt{Nh(f_N(z_0) - f_0(z_0))}} \right) \rightsquigarrow N \left( 0, \sigma^2 \begin{pmatrix} \Omega^{-1} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \right),
$$

where $\Omega = E(X - E(X|Z)) \otimes 2$.

$^4$The asymptotic independence between $\hat{\beta}_n^{(j)}$ and $f_N(z_0)$ is mainly due to the fact that $n/N = s^{-1} \to 0$.

$^5$The asymptotic variance $\Sigma_{22}$ of $f_N$ is the same as that of $f_{or}$. 
Efficiency Boosting

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

$$\sqrt{n}(\hat{\beta}_n^{(j)} - \beta_0) \rightsquigarrow N(0, \sigma^2(E(X - E(X|Z))^\otimes 2)^{-1}).$$

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

- By imposing some lower bound on $s^6$, we show that $^7$

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

as if the commonality information were available.

---

$^6$This lower bound requirement slows down the convergence rate of $\tilde{\beta}_n^{(j)}$ such that $\tilde{f}_N$ can be treated as if it were known.

$^7$Recall that $\tilde{\beta}_n^{(j)} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (Y_i^{(j)} - \beta^T X_i^{(j)} - \tilde{f}_N(Z_i^{(j)}))^2$. 
Efficiency Boosting

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

$$\sqrt{n}(\hat{\beta}_n^{(j)} - \beta_0) \sim N(0, \sigma^2(E(X - E(X|Z)) \otimes 2)^{-1}).$$

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

- By imposing some lower bound on $s^6$, we show that $^7$

$$\sqrt{n}(\tilde{\beta}_n^{(j)} - \beta_0^{(j)}) \sim N(0, \sigma^2(E[XX^T])^{-1})$$

as if the commonality information were available.

---

6 This lower bound requirement slows down the convergence rate of $\tilde{\beta}_n^{(j)}$ such that $f_N$ can be treated as if it were known.

7 Recall that $\tilde{\beta}_n^{(j)} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (Y_i^{(j)} - \beta^T X_i^{(j)} - \tilde{f}_N(Z_i^{(j)}))^2.$
Theorem 4 implies that $\hat{\beta}_n^{(j)}$ is semiparametric efficient:

$$\sqrt{n}(\hat{\beta}_n^{(j)} - \beta_0) \rightsquigarrow N(0, \sigma^2(E(X - E(X|Z)) \otimes^2)^{-1}).$$

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

By imposing some lower bound on $s$\(^6\), we show that\(^7\)

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

as if the commonality information were available.

\(^6\)This lower bound requirement slows down the convergence rate of $\tilde{\beta}_n^{(j)}$ such that $\tilde{f}_N$ can be treated as if it were known.

\(^7\)Recall that $\tilde{\beta}_n^{(j)} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (Y_i^{(j)} - \beta^T X_i^{(j)} - \tilde{f}_N(Z_i^{(j)}))^2$. 

---

**Efficiency Boosting**
Figure: Coverage probability of 95% confidence interval based on $\hat{\beta}_n^{(j)}$
A Partially Linear Modelling Non-Asymptotic Bound Efficiency Boosting Heterogeneity Testing

Coverage Prob/Ave Length when $N = 512$

Coverage Prob/Ave Length when $N = 1024$

Coverage Prob/Ave Length when $N = 2048$

Coverage Prob/Ave Length when $N = 4096$

Figure: Coverage probabilities and average lengths of 95% confidence intervals constructed based on $\hat{\beta}$ and $\check{\beta}$. In the above figures, dashed lines represent CI$_1$, which is constructed based on $\check{\beta}$, and solid lines represent CI$_2$, which is constructed based on $\hat{\beta}$. 
Consider a \textit{high dimensional} simultaneous testing:

\[ H_0 : \beta^{(j)} = \tilde{\beta}^{(j)} \text{ for all } j \in J, \]  

(3)

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

\[ H_1 : \beta^{(j)} \neq \tilde{\beta}^{(j)} \text{ for some } j \in J; \]  

(4)

Test statistic:

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

We can consistently approximate the quantile of the null distribution via bootstrap even when \( |J| \) diverges at an exponential rate of \( n^8 \).

\[ \text{By a nontrivial application of a recent Gaussian approximation theory.} \]
Consider a high dimensional simultaneous testing:

\[ H_0 : \beta^{(j)} = \tilde{\beta}^{(j)} \text{ for all } j \in J, \] (3)

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

\[ H_1 : \beta^{(j)} \neq \tilde{\beta}^{(j)} \text{ for some } j \in J; \] (4)

Test statistic:

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

We can consistently approximate the quantile of the null distribution via bootstrap even when \( |J| \) diverges at an exponential rate of \( n^8 \). 

---

By a nontrivial application of a recent Gaussian approximation theory.
Consider a high dimensional simultaneous testing:

\[ H_0 : \beta^{(j)} = \tilde{\beta}^{(j)} \text{ for all } j \in J, \]

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

\[ H_1 : \beta^{(j)} \neq \tilde{\beta}^{(j)} \text{ for some } j \in J; \]

Test statistic:

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

We can consistently approximate the quantile of the null distribution via bootstrap even when \( |J| \) diverges at an exponential rate of \( n^8 \).

---

\textsuperscript{8}By a nontrivial application of a recent Gaussian approximation theory.