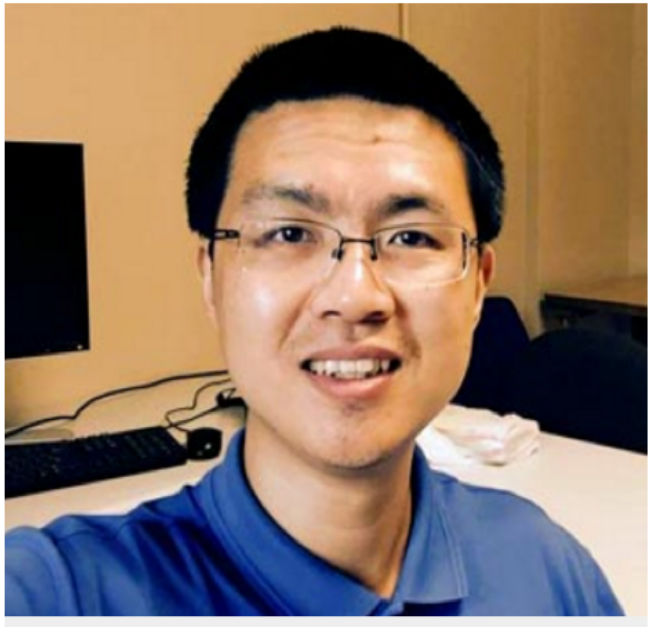
Computationally Efficient Nonparametric Testing

Guang Cheng[1]
Dept of Statistics
Purdue University
www.science.purdue.edu/bigdata/

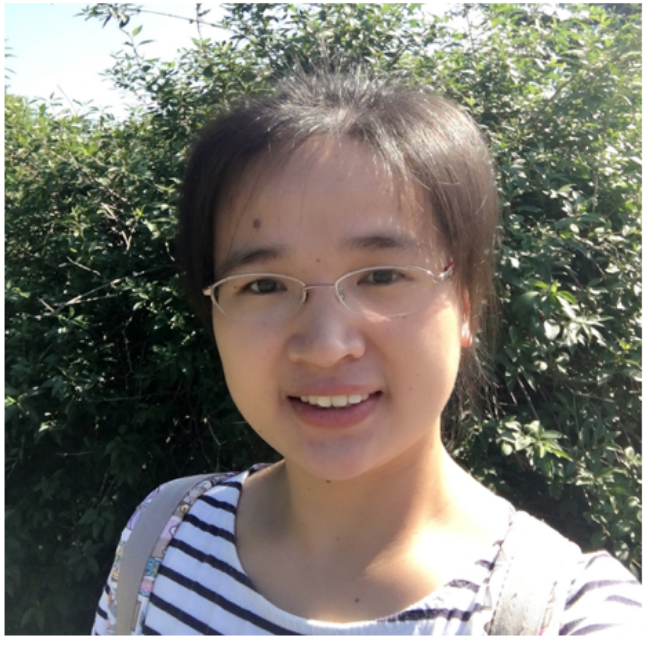
Statistics@Cambridge
May 26, 2017

[1] Acknowledge NSF, ONR and Simons Foundation.

Acknowledgment



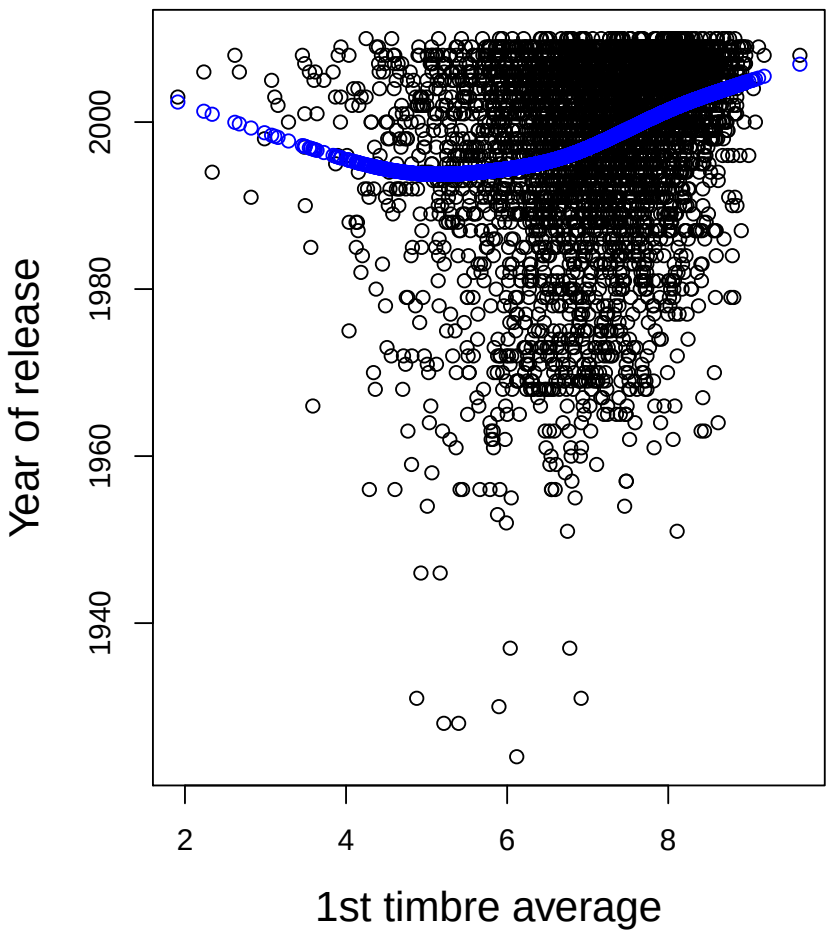
Zuofeng Shang @ Binghamton

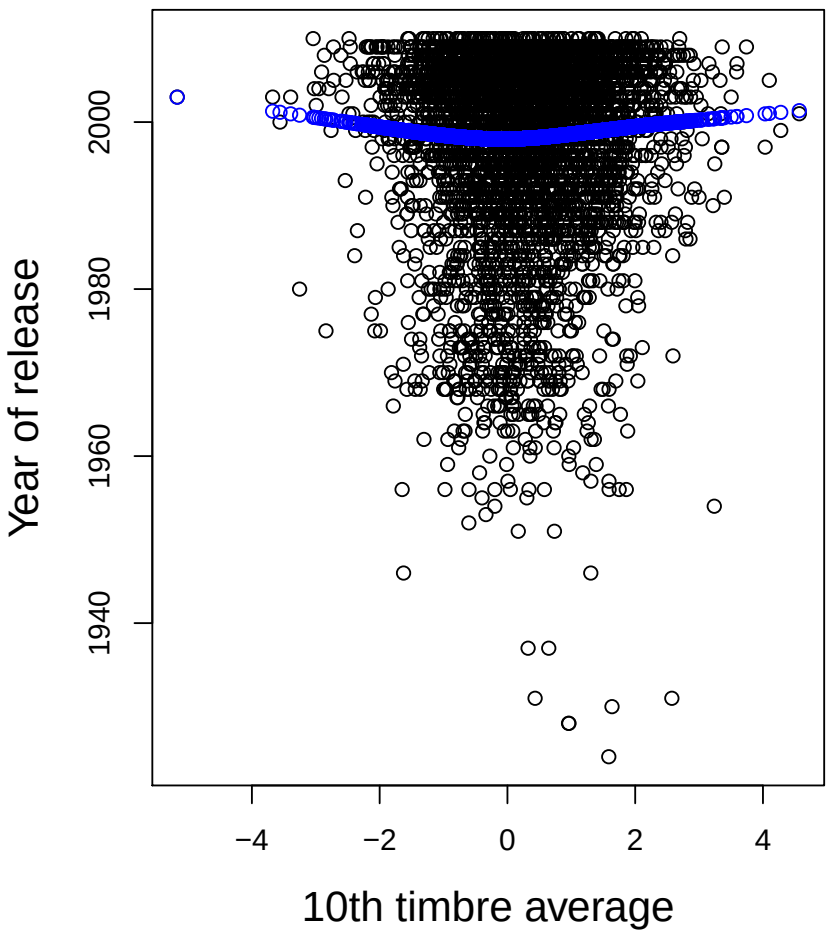


Meimei Liu @ Purdue

Million Song Dataset

- 1,000,000 music tracks released from the year 1922 to 2011
- Predictor variables: 12 timbre averages
- Goal: identify predictors associated with the year of release
- Data source: free, public, supported in part by NSF





Kernel ridge regression (KRR)

- Consider a nonparametric model

$$y_i = f(x_i) + \epsilon_i, \text{ for } i = 1, \ldots, n$$

- The standard KRR estimate f by

$$\widehat{f}_n := \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \underbrace{\lambda \|f\|_{\mathcal{H}}^2}_{\text{roughness penalty}}$$

reproducing kernel Hilbert space (RKHS)

Computational challenges

- The KRR theory says that

$$\widehat{f}_n(\cdot) = \sum_{i=1}^{n} \widehat{w}_i K(\cdot, x_i),$$

kernel function

where

$$\widehat{\mathbf{w}} = n^{-1} \underbrace{(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}}_{n \times n} \underbrace{\mathbf{y}}_{n \times 1},$$

$\mathbf{K}$ is $n \times n$ kernel matrix and $\mathbf{y}$ is n-dimensional response

Computational challenges

- The KRR theory says that

$$\widehat{f}_n(\cdot) = \sum_{i=1}^{n} \widehat{w}_i K(\cdot, x_i),$$

kernel function

where

$$\widehat{\mathbf{w}} = n^{-1} \underbrace{(\mathbf{K} + \lambda \mathbf{I_n})^{-1}}_{n \times n} \mathbf{y}_{n \times 1},$$

$\mathbf{K}$ is $n \times n$ kernel matrix and $\mathbf{y}$ is n-dimensional response

- Time (in seconds) taken to invert an n-dimensional matrix:

n	10^3	10^4	10^5	10^6
Time	0.78	148.8	28269.5	$>$ two months

How to address "curse of sample size?"

Our contributions

- We propose two computationally efficient testing methods in nonparametric settings:
  - one is based on randomized sketches
  - another employs parallel computing

- Characterize computational limits, i.e., the minimal computational cost to preserve statistical optimality

- Existing work only focus on estimation [Zhang et al, 2015; Yang et al, 2016]

Outline

Wald type test

- Consider a hypothesis testing problem

$$H_0 : f = f_0 \text{ vs. } H_1 : f \neq f_0$$

- A Wald type test statistic is

$$T_{n,\lambda} = \|\widehat{f}_R - f_0\|_{L^2}^2$$

$\widehat{f}_R$: an estimator of f

- Reject H_0 if $T_{n,\lambda}$ is large

An algorithm based on randomized sketches

- The standard KRR is equivalent to the following

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^n}{\arg\min}\, \mathbf{w}^T \underset{\substack{\uparrow \\ n\times n}}{\mathbf{K}^2}\, \mathbf{w} - \frac{2}{n}\mathbf{y}^T \mathbf{K}\, \mathbf{w} + \lambda \mathbf{w}^T \underset{\substack{\uparrow \\ n\times n}}{\mathbf{K}}\, \mathbf{w} \qquad (1)$$

An algorithm based on randomized sketches

- The standard KRR is equivalent to the following

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n} \mathbf{w}^T \underbrace{\mathbf{K}^2}_{n\times n} \mathbf{w} - \frac{2}{n}\mathbf{y}^T \mathbf{K}\,\mathbf{w} + \lambda \mathbf{w}^T \underbrace{\mathbf{K}}_{n\times n} \mathbf{w} \qquad (1)$$

- Let $\mathbf{S}$ be an $s \times n$ random matrix. Replacing $\mathbf{w} = \mathbf{S}^T\boldsymbol{\alpha}$, (1) becomes

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^s} \boldsymbol{\alpha}^T(\underbrace{\mathbf{S}\mathbf{K}^2\mathbf{S}^{\mathrm{T}}}_{s\times s})\boldsymbol{\alpha} - \frac{2}{n}\mathbf{y}^T(\mathbf{K}\mathbf{S}^T)\boldsymbol{\alpha} + \lambda\boldsymbol{\alpha}^T(\underbrace{\mathbf{S}\mathbf{K}\mathbf{S}^{\mathrm{T}}}_{s\times s})\boldsymbol{\alpha}$$

Solution: $\widehat{\boldsymbol{\alpha}} = n^{-1}(\underbrace{\mathbf{S}\mathbf{K}^2\mathbf{S}^{\mathrm{T}} + \lambda\,\mathbf{S}\mathbf{K}\mathbf{S}^{\mathrm{T}}}_{s\times s})^{-1}(\mathbf{S}\mathbf{K})\mathbf{y}$

An algorithm based on randomized sketches

- The standard KRR is equivalent to the following

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n} \mathbf{w}^T \underbrace{\mathbf{K}^2}_{n\times n} \mathbf{w} - \frac{2}{n}\mathbf{y}^T \mathbf{K}\,\mathbf{w} + \lambda \mathbf{w}^T \underbrace{\mathbf{K}}_{n\times n} \mathbf{w} \qquad (1)$$

- Let $\mathbf{S}$ be an $s \times n$ random matrix. Replacing $\mathbf{w} = \mathbf{S}^T\boldsymbol{\alpha}$, (1) becomes

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^s} \boldsymbol{\alpha}^T(\underbrace{\mathbf{S}\mathbf{K}^2\mathbf{S}^{\mathrm{T}}}_{s\times s})\boldsymbol{\alpha} - \frac{2}{n}\mathbf{y}^T(\mathbf{K}\mathbf{S}^T)\boldsymbol{\alpha} + \lambda\boldsymbol{\alpha}^T(\underbrace{\mathbf{S}\mathbf{K}\mathbf{S}^{\mathrm{T}}}_{s\times s})\boldsymbol{\alpha}$$

Solution: $\widehat{\boldsymbol{\alpha}} = n^{-1}(\underbrace{\mathbf{S}\mathbf{K}^2\mathbf{S}^{\mathrm{T}} + \lambda\,\mathbf{S}\mathbf{K}\mathbf{S}^{\mathrm{T}}}_{s\times s})^{-1}(\mathbf{S}\mathbf{K})\mathbf{y}$

- Estimate f by the following sketched KRR (SKRR):

$$\widehat{f}_R(\cdot) = \sum_{i=1}^{n} (\mathbf{S}^T\widehat{\boldsymbol{\alpha}})_i K(\cdot, x_i)$$

Choice of random matrix

Independent sub-Gaussian entries, e.g., Gaussian or Bernoulli

Gaussian

$$\begin{pmatrix} -0.38 & -0.45 & \cdots & 1.19 \\ 1.42 & -0.94 & \cdots & 0.73 \\ \vdots & \vdots & \vdots & \vdots \\ -0.64 & 1.36 & \cdots & 0.57 \end{pmatrix}$$

Bernoulli

$$\begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 \end{pmatrix}$$

Computing time vs. projection dimension

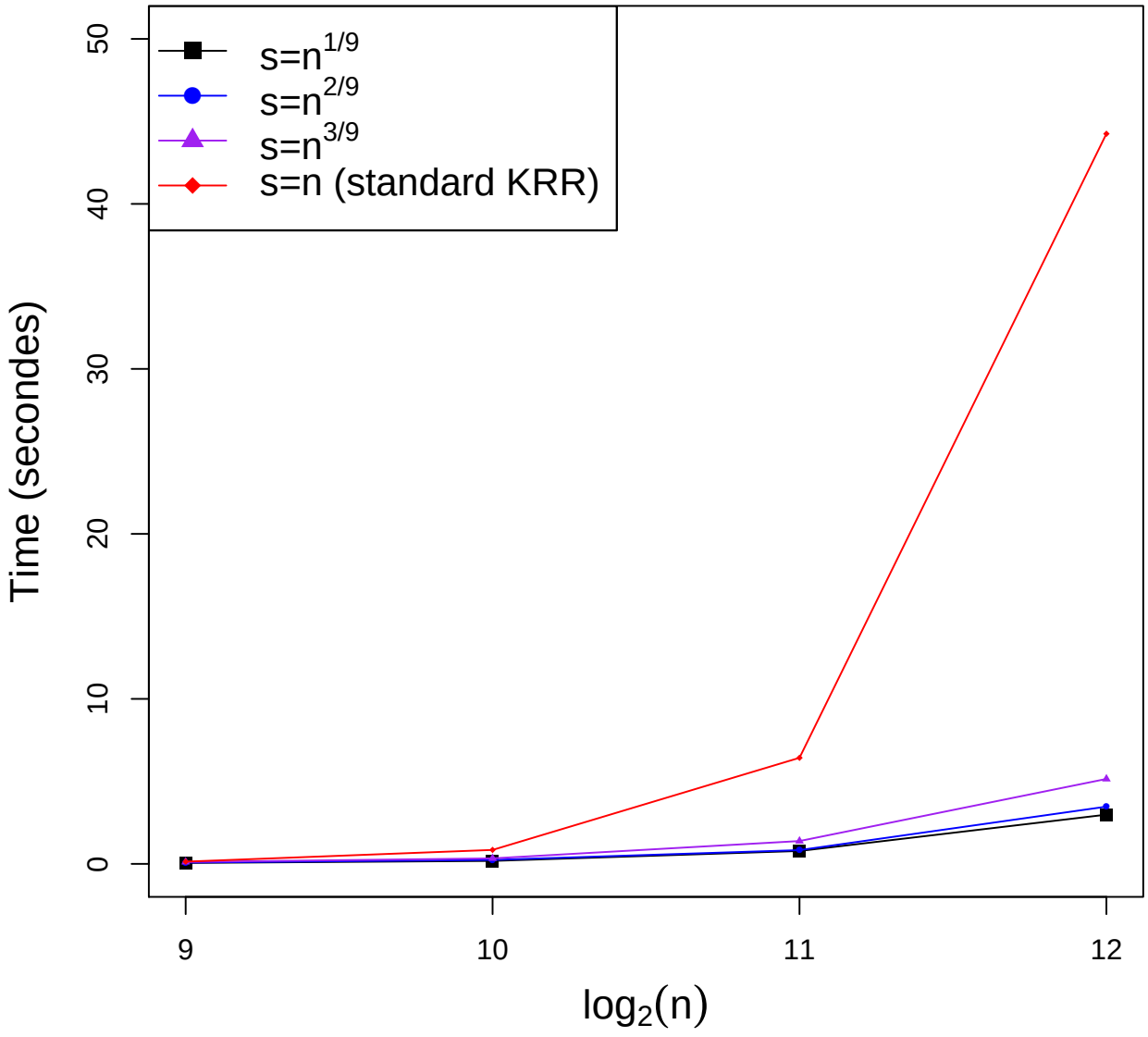


Figure 1: Computing time of $T_{n,\lambda}$

Existence of minimal projection dimension

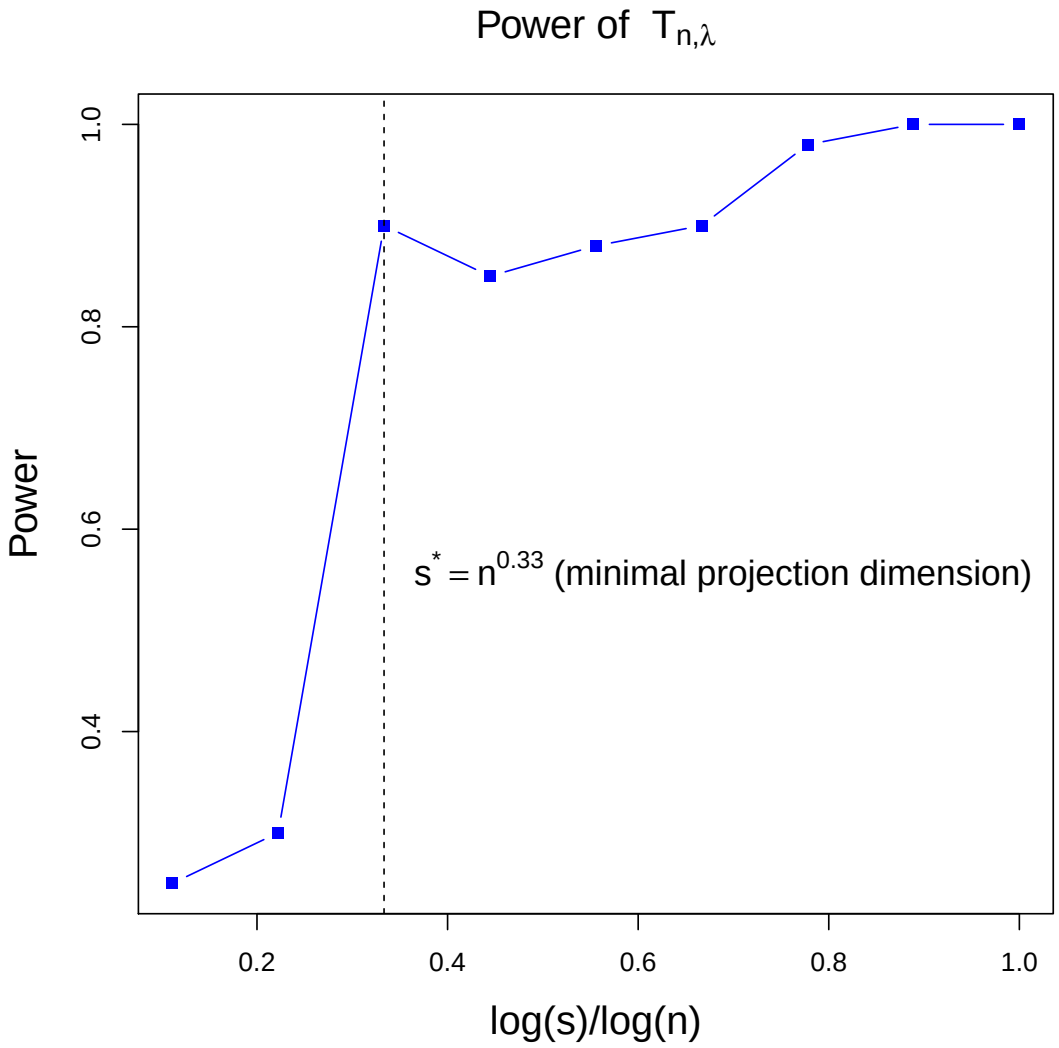


Figure 2: $n = 2^{12}$. Significance level 0.05. Run 1,000 replications

How to characterize minimal projection dimension?

Commonly used kernels

Mercer's Theorem:

$$K(x, x') = \sum_{i=1}^{\infty} \mu_i \phi_i(x) \phi_i(x'),$$

μ_i: eigenvalue; ϕ_i: eigenfunction

- Polynomial kernel:
  $\mu_k \asymp k^{-2m}$ for $m > 1/2$, e.g., $m = 2$ (cubic spline)
- Exponential kernel:
  $\mu_k \asymp \exp(-\alpha k^p)$ for $\alpha, p > 0$
- Finite rank kernel:
  $\mu_k = 0$ for $k > r$

Minimal projection dimension

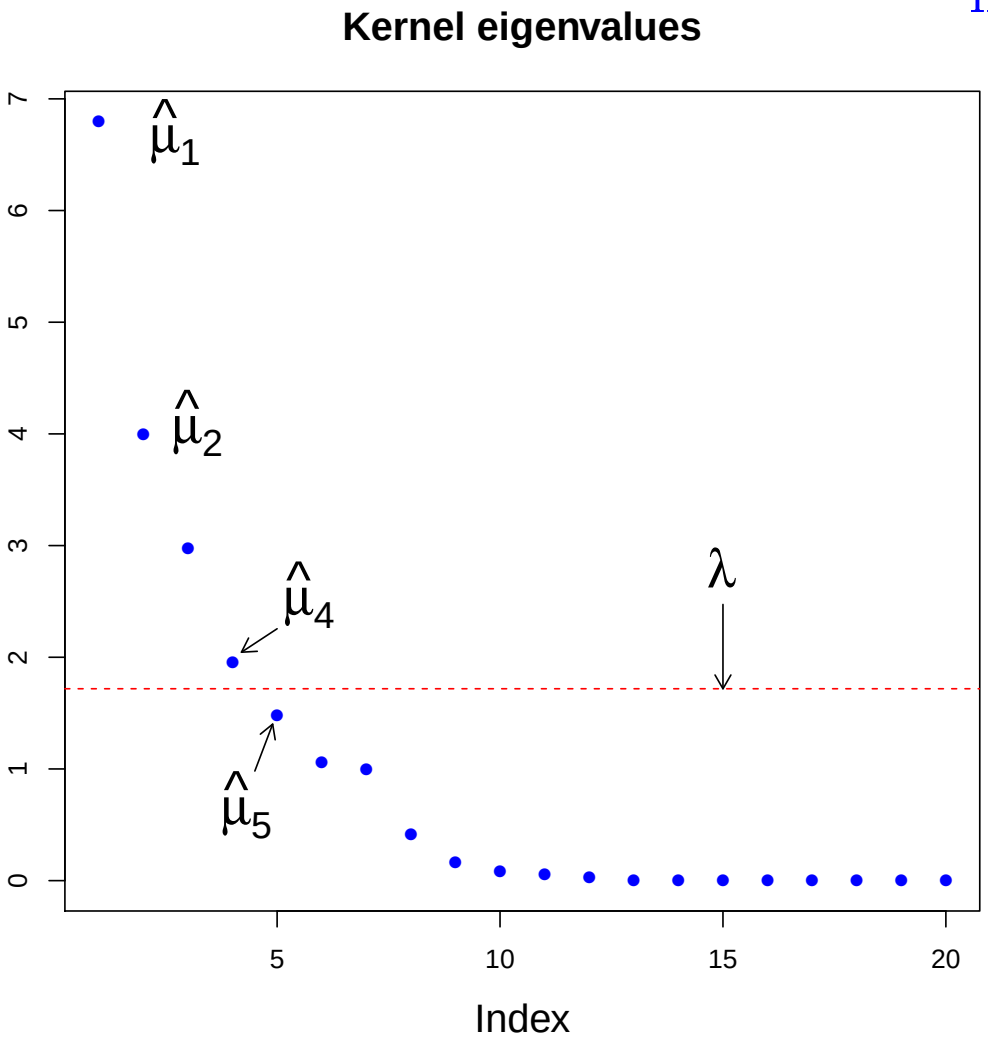
	Polynomial kernel	Exponential kernel
s^*	$n^{\frac{2}{4m+1}}$	$(\log n)^{\frac{1}{2p}}$

Practical choice of projection dimension

Eigenvalues of kernel matrix: $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \cdots$

Choose $s = \min\{j : \widehat{\mu}_j \leq \lambda\}$

how to choose λ?



$s = 5$

Testing consistency

Theorem 1

Suppose $\lambda \to 0$. Then we have under H_0,

$$\frac{T_{n,\lambda} - \mu_{n,\lambda}}{\sigma_{n,\lambda}} \xrightarrow{d} N(0,1),$$

where $\mu_{n,\lambda} = E_{H_0}\{T_{n,\lambda}\}$ and $\sigma^2_{n,\lambda} = Var_{H_0}\{T_{n,\lambda}\}$.

Testing consistency

Theorem 1

Suppose $\lambda \to 0$. Then we have under H_0,

$$\frac{T_{n,\lambda} - \mu_{n,\lambda}}{\sigma_{n,\lambda}} \xrightarrow{d} N(0,1),$$

where $\mu_{n,\lambda} = E_{H_0}\{T_{n,\lambda}\}$ and $\sigma^2_{n,\lambda} = Var_{H_0}\{T_{n,\lambda}\}$.

- Testing rule is

$$\phi_{n,\lambda} = I(|T_{n,\lambda} - \mu_{n,\lambda}| > z_{1-\alpha/2}\sigma_{n,\lambda}),$$

$$\phi_{n,\lambda} = 1 \Longleftrightarrow \text{reject } H_0$$

Power

Let

$$s^* = \begin{cases} n^{\frac{2}{4m+1}}, & \text{polynomial kernel} \\ (\log n)^{\frac{1}{p}}, & \text{exponential kernel} \end{cases}$$

Theorem 2

Suppose $s \gtrsim s^*$ and $\lambda \to 0$. Then for any $\epsilon > 0$, there exist $c_\epsilon, n_\epsilon > 0$, s.t., for any $n \geq n_\epsilon$,

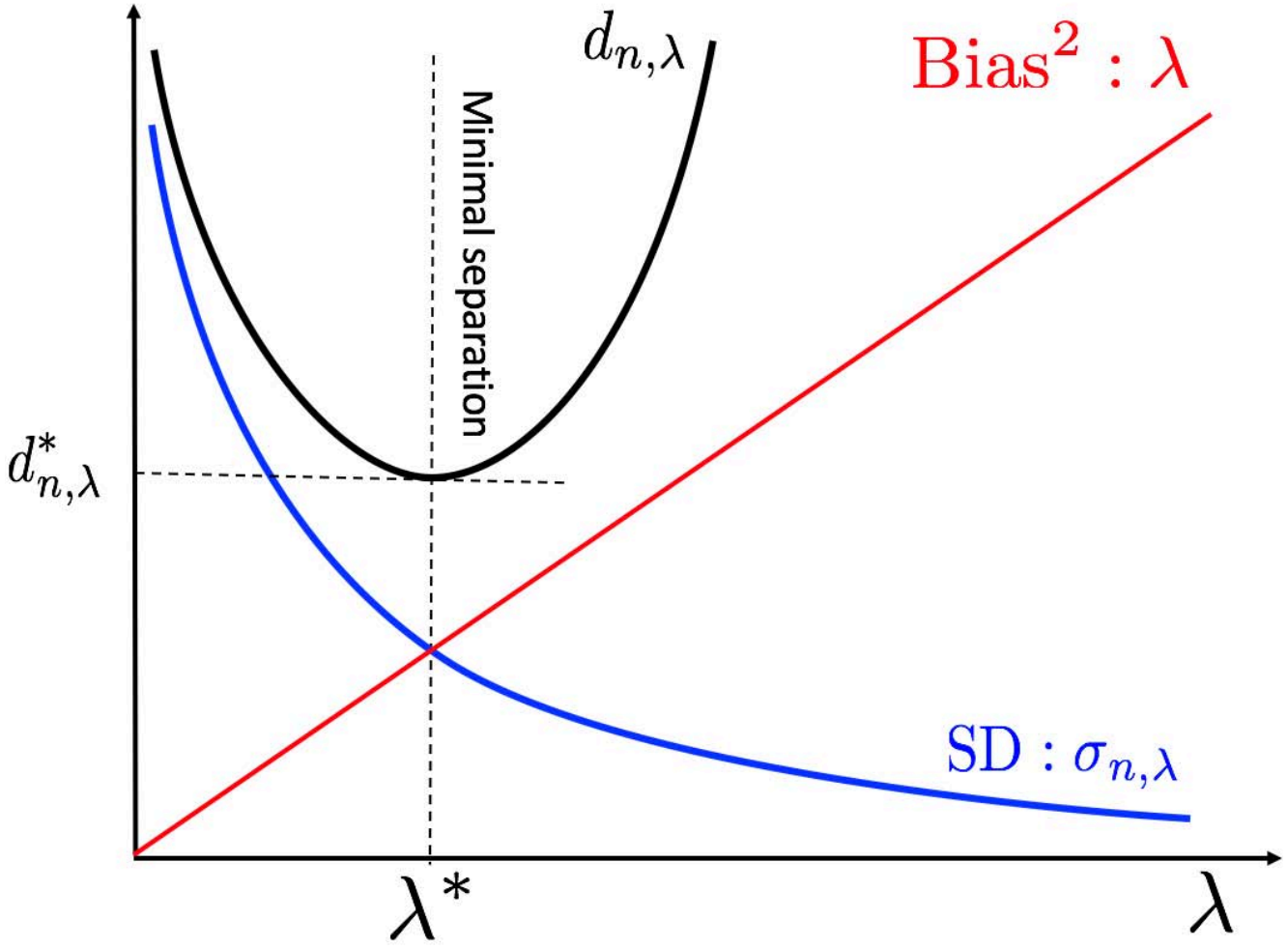
$$\inf_{f \in \mathcal{B}, \|f - f_0\|_{L^2} \geq c_\epsilon d_{n,\lambda}} P_f(\phi_{n,\lambda} = 1) \geq 1 - \epsilon, \quad \text{(high power)}$$

where $\mathcal{B} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq C\}$ for a positive constant C and the separation rate $d_{n,\lambda} = \sqrt{\lambda + \sigma_{n,\lambda}}$.

λ: Bias^2; $\sigma_{n,\lambda}$: SD

- Small separation rate $\Longleftrightarrow$ powerful test

Minimal separation rate: Bias^2 vs. SD tradeoff



$d_{n,\lambda}$ attains $d^*_{n,\lambda}$ at $\lambda = \lambda^*$. Rate of λ^*: Rademacher complexity

Minimal projection dimension

Theorem 3

Suppose $s \ll s^*$ and $\lambda \to 0$. Then there exists a positive sequence $\beta_{n,\lambda}$ with $\lim_{n\to\infty} \beta_{n,\lambda} = \infty$, s.t.

$$\limsup_{n\to\infty} \inf_{f\in\mathcal{B}, \|f-f_0\|_{L^2} \geq \beta_{n,\lambda} d^*_{n,\lambda}} P_f(\phi_{n,\lambda} = 1) \leq \alpha. \quad \text{(low power)}$$

Minimal projection dimension

Theorem 3

Suppose $s \ll s^*$ and $\lambda \to 0$. Then there exists a positive sequence $\beta_{n,\lambda}$ with $\lim_{n\to\infty} \beta_{n,\lambda} = \infty$, s.t.

$$\limsup_{n\to\infty} \inf_{f\in\mathcal{B}, \|f-f_0\|_{L^2} \geq \beta_{n,\lambda} d^*_{n,\lambda}} P_f(\phi_{n,\lambda} = 1) \leq \alpha. \quad \text{(low power)}$$

- s^* is a *sharp* lower bound for projection dimension

	Polynomial Kernel	Exponential Kernel
s^*	$n^{\frac{2}{4m+1}}$	$(\log n)^{\frac{1}{p}}$
$d^*_{n,\lambda}$	$n^{-\frac{2m}{4m+1}}$	$(\log n)^{\frac{1}{4p}} n^{-1/2}$
λ^*	$n^{-\frac{4m}{4m+1}}$	$(\log n)^{\frac{1}{2p}} n^{-1}$

Can we perform hypothesis testing when kernel smoothness is unknown?

An adaptive test

Consider a polynomial kernel of order m with m unknown

An adaptive test

Consider a polynomial kernel of order m with m unknown

1. For any integer m, find

$$\tau_m = \frac{T_{n,\lambda}(m) - \mu_{n,\lambda}(m)}{\sigma_{n,\lambda}(m)}$$

An adaptive test

Consider a polynomial kernel of order m with m unknown

1. For any integer m, find

$$\tau_m = \frac{T_{n,\lambda}(m) - \mu_{n,\lambda}(m)}{\sigma_{n,\lambda}(m)}$$

2. Find

$$\tau_n^* = \max_{1 \le m \le m_n} \tau_m$$

$\uparrow$ $m_n \asymp (\log n)^{d_0}$, $d_0 \in (0, 1/2)$

An adaptive test

Consider a polynomial kernel of order m with m unknown

1. For any integer m, find

$$\tau_m = \frac{T_{n,\lambda}(m) - \mu_{n,\lambda}(m)}{\sigma_{n,\lambda}(m)}$$

2. Find

$$\tau_n^* = \max_{1 \le m \le m_n} \tau_m$$

$\hookrightarrow$ $m_n \asymp (\log n)^{d_0}$, $d_0 \in (0, 1/2)$

3. Let

$$\tau_{n,m_n} = B_n(\tau_n^* - B_n)$$

$\hookrightarrow$ solution to $2\pi B_n^2 \exp(B_n^2) = m_n^2$

Consistency of adaptive testing

Theorem 4

Suppose $m_n \asymp (\log n)^{d_0}$ for $d_0 \in (0, 1/2)$. Then for any $\alpha \in (0, 1)$, under H_0,

$$P(\tau_{n,m_n} \leq c_\alpha) \to 1 - \alpha, \quad \text{as } n \to \infty,$$

where $c_\alpha = -\log(-\log(1-\alpha))$.

- Proof is based on Stein's leave-one-out method (Stein, 1986)
- Testing rule is

$$\phi^*_{n,\lambda} = I(\tau_{n,m_n} > c_\alpha),$$

$$\phi^*_{n,\lambda} = 1 \Longleftrightarrow \text{reject } H_0$$

Power of adaptive testing

To achieve high power, choose

$$s(m) \gtrsim n^{\frac{2}{4m+1}} (\log\log n)^{-\frac{1}{4m+1}}, \text{ for } m = 1, \ldots, m_n$$

Minimal separation rate:

$$\delta(n, m) \equiv n^{-\frac{2m}{4m+1}} (\log\log n)^{\frac{m}{4m+1}}$$

price for adaptivity

$\delta(n, m)$ is optimal (Spokoiny, 1996)

1. Nonparametric testing based on randomized sketches

2. Nonparametric testing based on parallel computing

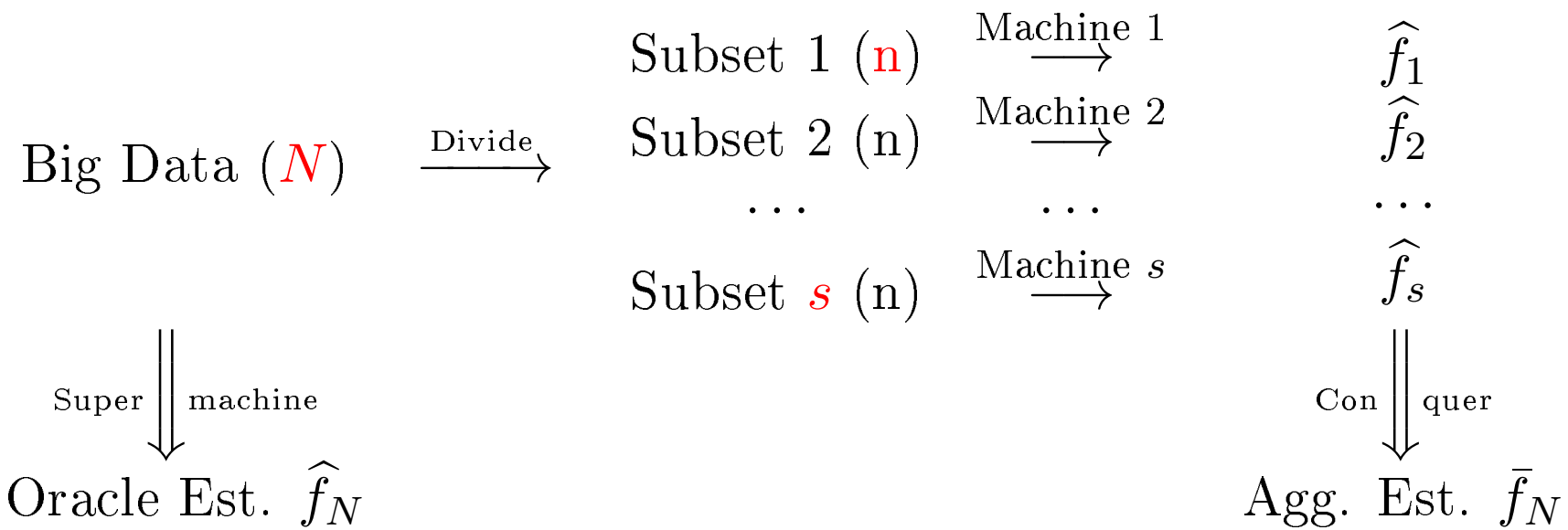
3. Numerical results

What if we can use parallel computing?

Divide-and-Conquer (DC)

Consider the same nonparametric regression model:

$$y = f(x) + \epsilon$$



$$\bar{f}_N = \frac{1}{s}\sum_{j=1}^{s} \widehat{f}_j$$

Wald type test

- Consider the same hypothesis testing problem

$$H_0 : f = f_0 \text{ vs. } H_1 : f \neq f_0$$

- Consider a Wald type test statistic

$$T_{N,\lambda} = \|\bar{f}_N - f_0\|_{L^2}^2$$

- Prefer a large number of divisions to reduce computational cost

Computing time vs. number of divisions

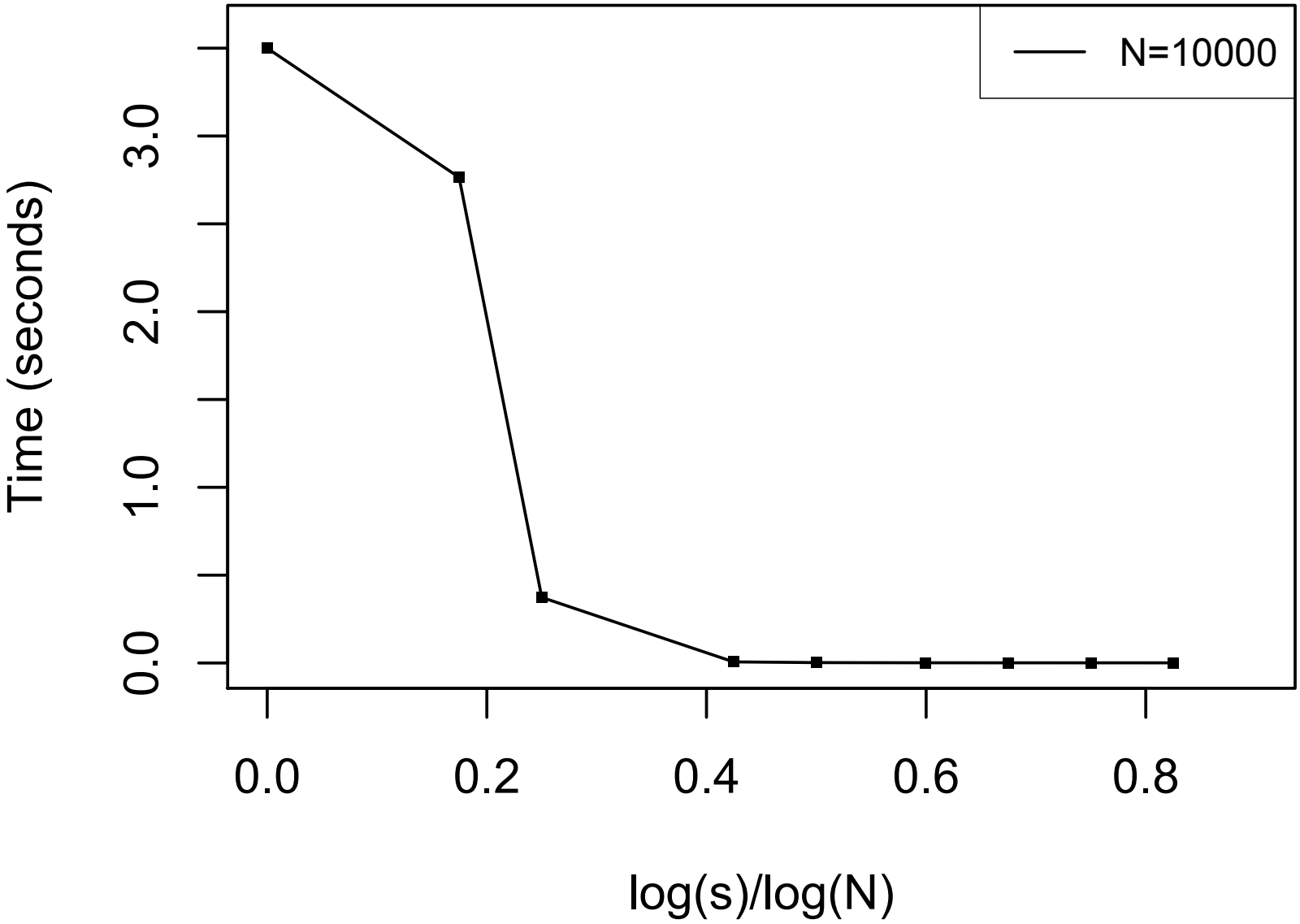


Figure 3: Computing time is decreasing with s

Existence of maximal number of divisions

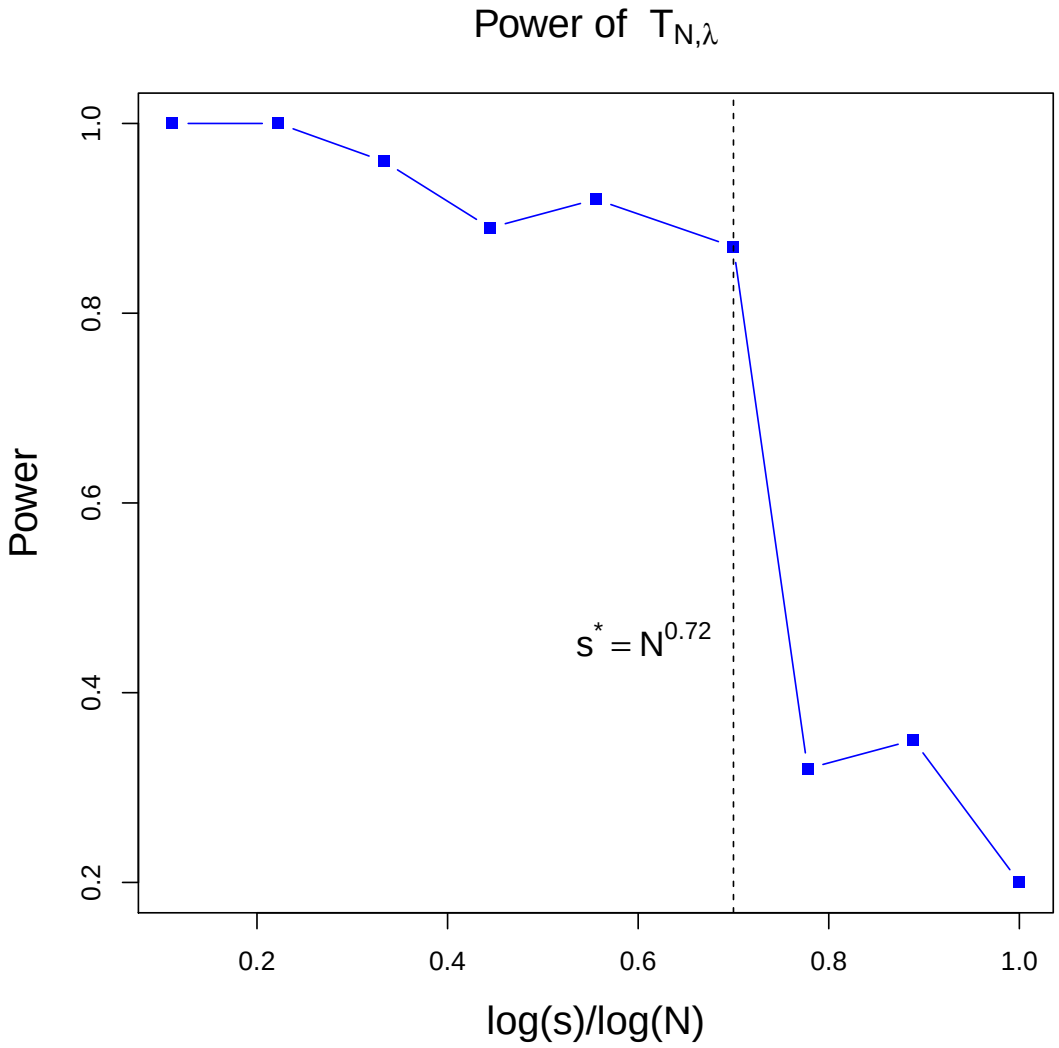


Figure 4: N=10,000. Significance level 0.05. Run 1,000 replicates

Phase transition diagram

Consider smoothing spline regression with a smoothing parameter λ and smoothness $m \geq 1$.

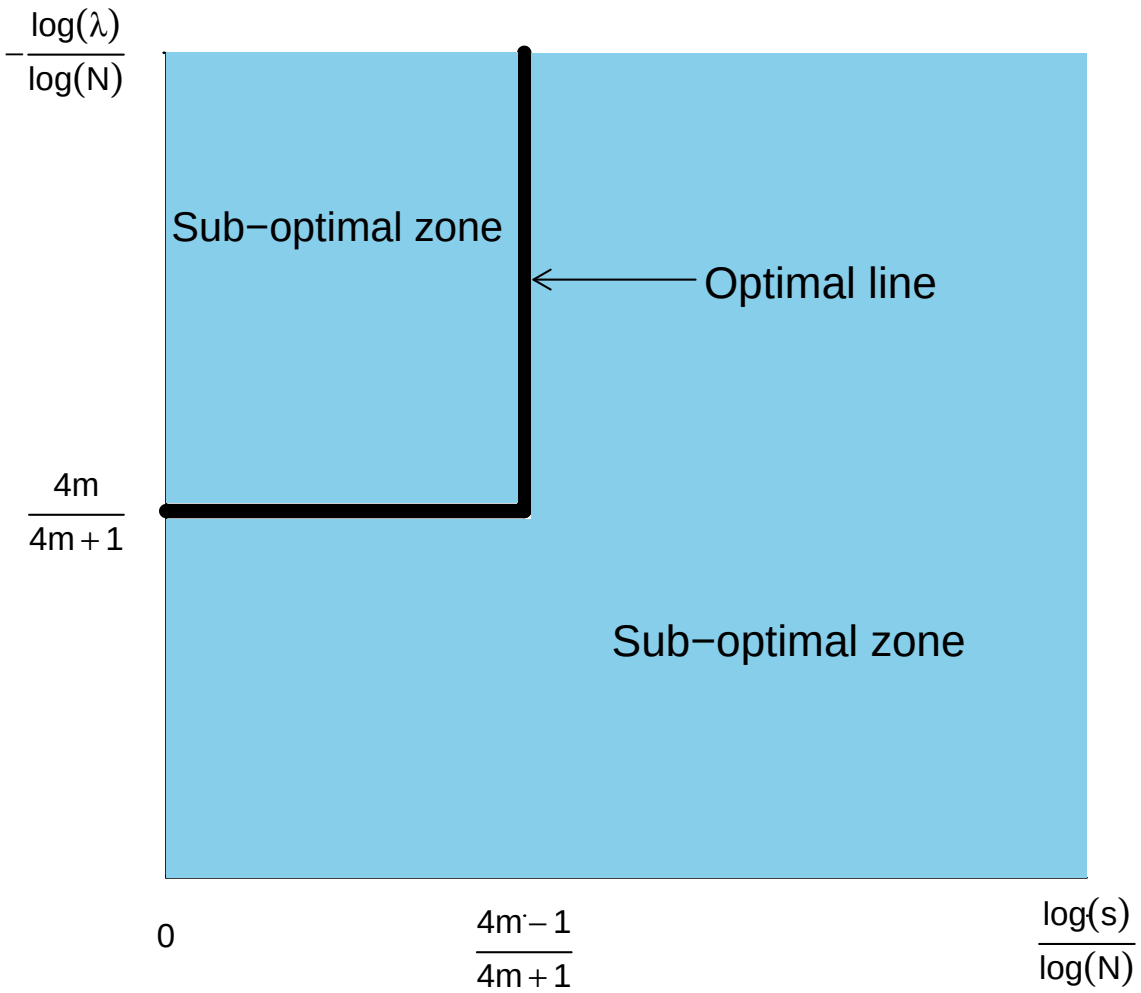


Figure 5: *Locations of* (λ, s) *achieving testing optimality*

Main theorem: testing consistency

Theorem 6

Suppose $\lambda \to 0$, $n \to \infty$ when $N \to \infty$, and $\lim_{N\to\infty} n\lambda^{1/2m}$ exists (which could be infinity). Then, we have under H_0,

$$\frac{T_{N,\lambda} - \mu_{N,\lambda}}{\sigma_{N,\lambda}} \xrightarrow{d} N(0,1), \quad \text{as } N \to \infty,$$

where $\mu_{N,\lambda} = \mathrm{E}_{H_0}\{T_{N,\lambda}\}$ and $\sigma^2_{N,\lambda} = \mathrm{Var}_{H_0}\{T_{N,\lambda}\}$.

Testing rule is

$$\phi_{N,\lambda} = I(|T_{N,\lambda} - \mu_{N,\lambda}| \geq z_{1-\alpha/2}\sigma_{N,\lambda}),$$

$$\phi_{N,\lambda} = 1 \Longleftrightarrow \text{reject } H_0$$

Main theorem: power

Consider m-order smoothing splines

Let

$$d_{N,\lambda} = \sqrt{\lambda + n^{-2m} + \sigma_{N,\lambda}}$$

Theorem 7

Suppose $s \lesssim N^{\frac{4m-1}{4m+1}}$, $\lambda \to 0$, $n \to \infty$ when $N \to \infty$, and $\lim_{N\to\infty} n\lambda^{1/2m}$ exists (which could be infinity). Then for any $\varepsilon > 0$, there exist $C_\varepsilon, N_\varepsilon > 0$ s.t. for any $N \geq N_\varepsilon$,

$$\inf_{\substack{f\in\mathcal{B} \\ \|f-f_0\|_2 \geq C_\varepsilon d_{N,\lambda}}} P_f\left(\phi_{N,\lambda} = 1\right) \geq 1 - \varepsilon, \quad \text{(high power)}$$

where $\mathcal{B} = \{f \in S^m(\mathbb{I}) : \|f\|_{\mathcal{H}} \leq C\}$ for a positive constant C.

- $\lambda^* = N^{-\frac{4m}{4m+1}}$

Main theorem: maximal number of divisions

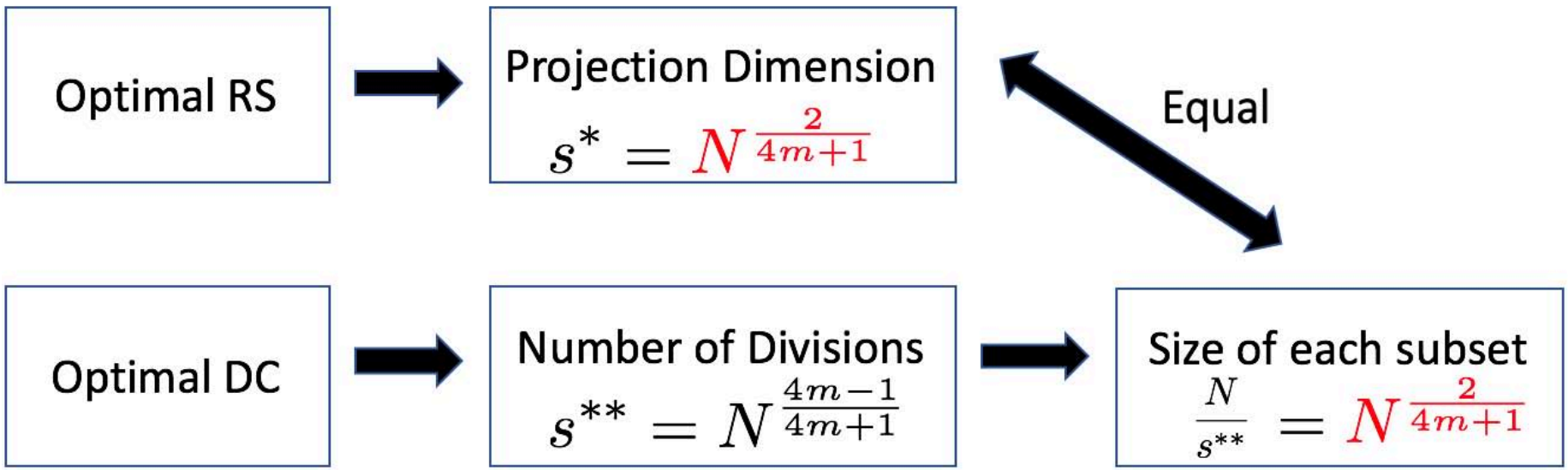
Theorem 8

Suppose $s \gg N^{\frac{4m-1}{4m+1}}$, $\lambda \to 0$, $n \to \infty$ when $N \to \infty$, and $\lim_{N\to\infty} n\lambda^{1/2m}$ exists (possibly infinity). Then there exists a positive sequence $\beta_{N,\lambda}$ with $\lim_{N\to\infty} \beta_{N,\lambda} = \infty$ s.t.

$$\limsup_{N\to\infty} \inf_{\substack{f\in\mathcal{B} \\ \|f-f_0\|_2 \geq \beta_{N,\lambda} d^*_{N,\lambda}}} P_f\left(\phi_{N,\lambda} = 1\right) \leq \alpha. \quad \text{(low power)}$$

- $s^{**} = N^{\frac{4m-1}{4m+1}}$ is a *sharp* upper bound for number of divisions to maintain testing optimality

A "theoretical" suggestion



1. Nonparametric testing based on randomized sketches

2. Nonparametric testing based on parallel computing

3. Numerical results

Simulation study

Consider the hypothesis testing problem

$$H_0 : f = 0 \text{ vs. } H_1 : f \neq 0$$

- $y_i = f(x_i) + \epsilon_i,\ i = 1, \ldots, n,\ n = 2^9, 2^{10}, 2^{11}, 2^{12}$
- $\epsilon_i \overset{iid}{\sim} N(0,1)$ and $x_i \overset{iid}{\sim} Unif[0,1]$
- $f(x) = c(3\beta_{30,17}(x) + 2\beta_{3,11}(x))$ with $c = 0, 0.01, 0.02, 0.03$

  $\beta_{30,17}$: Beta density

Four testing methods

- Standard KRR

- Randomized sketches (RS):
  - projection dimension $2n^{2/9}$

- Adaptive randomized sketches (Adaptive RS):
  - projection dimension $2n^{2/9}(\log\log n)^{-1/9}$
  - number of multiple tests $\sqrt{\log n}$

- Divide-and-Conquer (DC):
  - number of divisions $0.5n^{7/9}$

Computing time

Method	$n = 2^9$	$n = 2^{10}$	$n = 2^{11}$	$n = 2^{12}$
Standard KRR	1.44	6.67	42.61	332.08
RS	0.31	0.40	0.79	2.98
Adaptive RS	0.53	1.55	4.33	15.93
DC	0.28	0.35	0.54	2.13

Computing time (in seconds) of one trial on a Windows computer with 2GB of memory and a single-threaded 2.70Ghz CPU

Size and power

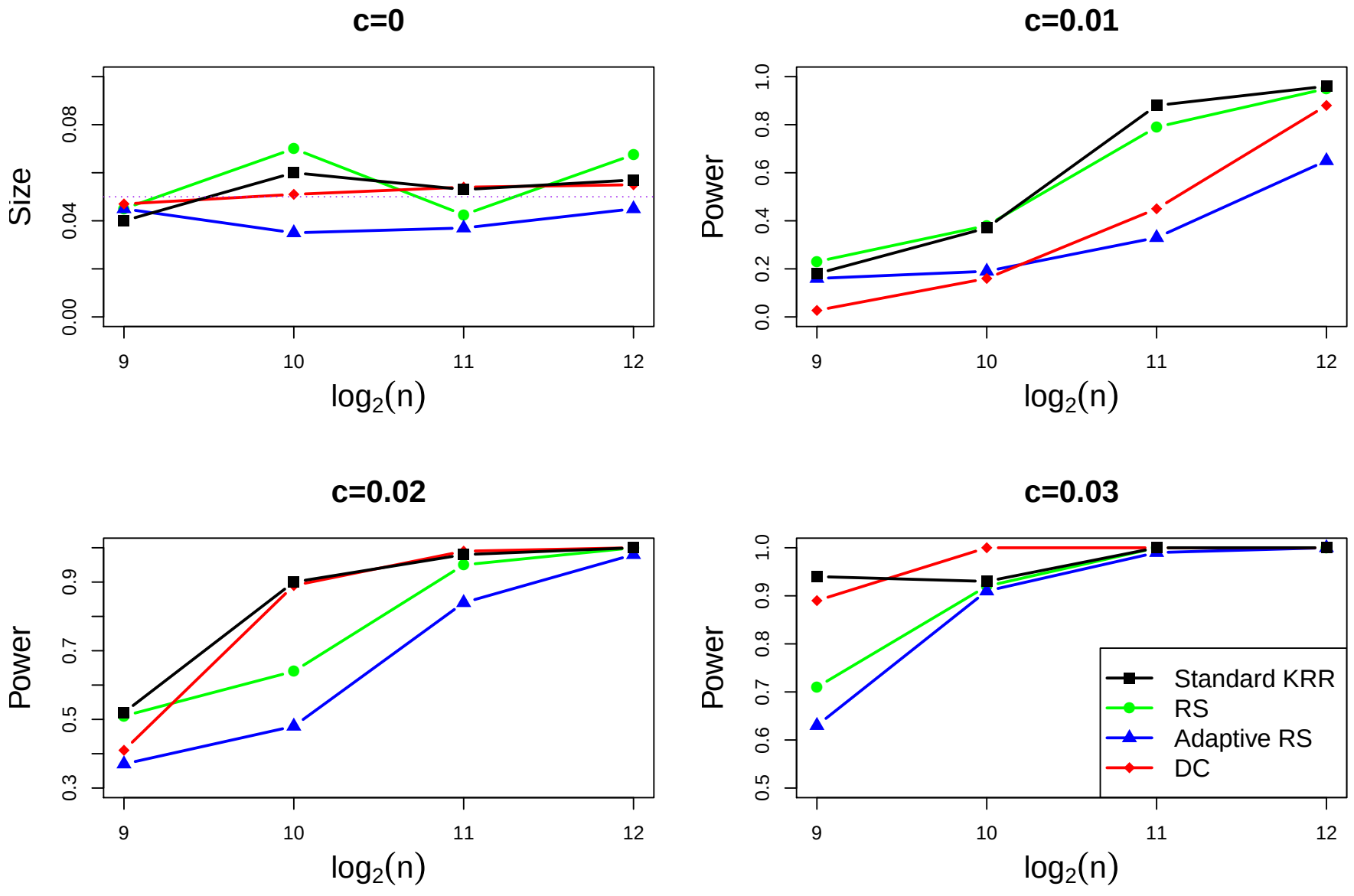


Figure 6: Significance level 0.05. Averages over 1,000 replications

Real data analysis

- Million Song Dataset:
  - 1,000,000 observations (music tracks)
  - year of release y
  - 12 timbre averages $x_1, \ldots, x_{12}$
- Goal: identify significant timbre averages
- Fit $y = f(x_j)$+error for $1 \leq j \leq 12$. Test $H_0 : f$ is constant
- Gaussian kernel $K(x, y) = \exp\left(-\frac{(x-y)^2}{\phi}\right)$; unknown $\phi > 0$
- Estimate $\widehat{\phi} \approx 3$ based on distributed GCV (Xu, Shang, C., 2016) and 463,715 training samples
- Wald test on 536,285 testing samples:
  - projection dimension $536,285^{\frac{2}{9}} \approx 19$
  - parallel processors $536,285^{\frac{7}{9}}/100 \approx 286$

Selection of significant timbre variables

	x_1	x_2	x_3	x_4	x_5	x_6
RS	0.0318	0.0943	0.7740	0.4069	0.7586	0.3699
DC	0.0205	0.0319	0.8212	0.3211	0.4729	0.4103
	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
RS	0.0433	0.2305	0.1642	0.5591	0.8979	0.0201
DC	0.0319	0.1005	0.0449	0.7138	0.7753	0.0037

Table 1: P-values for testing marginal association by two methods

- Matrix multiplication for RS: about 380 seconds
- Wald test: about 10–20 seconds
- One node in a cluster with 20 cores

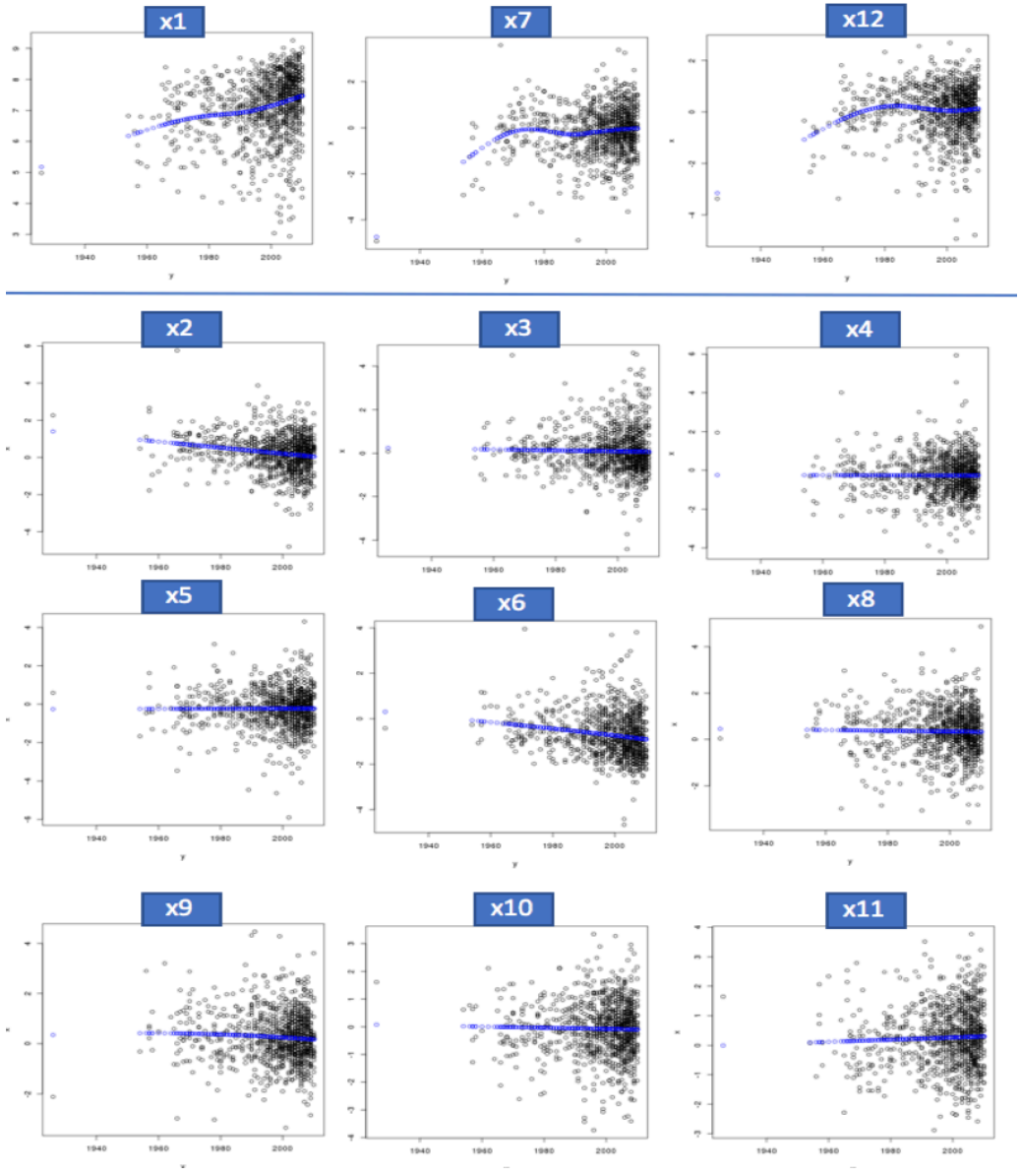


Figure 7: x_1, x_7, x_{12} have significant patterns, with p-values < 0.05 for both RS and DC methods.

Summary

- Nonparametric testing based on RS and DC
- Equally powerful as standard method with less computing cost
- Computational limits are characterized

Thanks! and Questions?

Backup slides

A By-product: minimax optimal estimation

Recall that $T_{n,\lambda} = \|\widehat{f}_R - f_0\|_{L^2}^2$ is the squared estimation error

Corollary (Optimal Estimation Rate)

Suppose that $\lambda \to 0$, $n\lambda \to \infty$. With probability greater than $1 - 2\exp\{-cn\lambda\}$, we have

$$\|\widehat{f}_R - f_0\|_{L^2}^2 \leq c'(\lambda + \frac{s_\lambda}{n})$$

with $s_\lambda = \min\{j : \widehat{\mu}_j \leq \lambda\}$.

- Polynomial kernel:
  $\lambda = n^{-\frac{2m}{2m+1}}$ and $s_\lambda = n^{\frac{1}{2m+1}}$
- Exponential kernel:
  $\lambda = (\log n)^{1/p} n^{-1}$ and $s_\lambda = (\log n)^{1/p}$
- Recover the minimax optimal estimation results in Yang, Pilanci and Wainwright (2016, AoS) under the same set of conditions

Calculation of $\mu_{n,\lambda}$ and $\sigma_{n,\lambda}$

Define $\Delta = \mathbf{K}\mathbf{S}^T(\mathbf{S}\mathbf{K}^2\mathbf{S}^{\mathrm{T}} + \lambda\,\mathbf{S}\mathbf{K}\mathbf{S}^{\mathrm{T}})^{-1}\mathbf{S}\mathbf{K}$. Under $H_0 : f = 0$, we have $y_i = \epsilon_i$ for $i = 1, \ldots, n$, and hence

$$T_{n,\lambda} = \|\widehat{f}_R\|_{L^2}^2 \approx \|\widehat{f}_R\|_n^2 = \frac{1}{n}\epsilon^T\Delta^2\epsilon \quad \text{(for large } n\text{)}$$

So we have the following approximations for practical use:

$$\mu_{n,\lambda} \approx \frac{1}{n}\mathrm{Tr}(\Delta^2) = \frac{1}{n}\mathrm{Tr}(\Gamma^2)$$

and

$$\sigma_{n,\lambda}^2 \approx \frac{2}{n^2}\mathrm{Tr}(\Delta^4) = \frac{2}{n^2}\mathrm{Tr}(\Gamma^4)$$

where

$$\Gamma = \mathbf{S}\mathbf{K}^2\mathbf{S}^T(\mathbf{S}\mathbf{K}^2\mathbf{S}^{\mathrm{T}} + \lambda\,\mathbf{S}\mathbf{K}\mathbf{S}^{\mathrm{T}})^{-1} \quad (s \times s)$$

Projection dimension for adaptive testing

Theorem 5

Suppose

$$s(m) \gtrsim n^{\frac{2}{4m+1}} (\log\log n)^{-\frac{1}{4m+1}}$$

for each $m \in \{1, \ldots, m_n \asymp (\log n)^{d_0}\}$. Then, for any $\varepsilon > 0$, there exist positive constants $\tilde{c}_\varepsilon, \widetilde{n}_\varepsilon$ such that for any $n \geq \widetilde{n}_\varepsilon$

$$\inf_{\substack{f \in \mathcal{B}_{n,m} \\ \|f - f_0\|_{L^2} \geq \tilde{c}_\varepsilon \delta(n,m)}} P_f(\phi^*_{n,\lambda} = 1) \geq 1 - \varepsilon, \text{(high power)}$$

where $\mathcal{B}_{n,m} = \{f \in \mathcal{H}(m) : (\mathbf{f})^T \mathbf{K}^{-1} \mathbf{f} \leq 1\}$ with $\mathbf{f} = (f(x_1), \cdots, f(x_n))$, and $\mathbf{K}$ is the kernel matrix, and

$$\delta(n, m) \equiv n^{-2m/(4m+1)} (\log\log n)^{m/(4m+1)}.$$

price for adaptivity

- $\delta(n, m)$ is optimal (Spokoiny, 1996)