Can we do statistical inference in a non-asymptotic way?[1]

Guang Cheng[2]
Statistics@Purdue
www.science.purdue.edu/bigdata/

ONR Review Meeting@Duke
Oct 11, 2017

[1] Acknowledge NSF, ONR and Simons Foundation.
[2] An "exploratory" and ongoing work with Z. Shang and Y. Yang.

Starting from Stat 101

- Given that $X_1, .., X_n \overset{iid}{\sim} N(\theta_0, 1)$, we have

$$P\left(\theta_0 \in \left[\bar{X} \pm \frac{1.96}{\sqrt{n}}\right]\right) = 95\%;$$

Starting from Stat 101

- Given that $X_1, .., X_n \overset{iid}{\sim} N(\theta_0, 1)$, we have

$$P\left(\theta_0 \in \left[\bar{X} \pm \frac{1.96}{\sqrt{n}}\right]\right) = 95\%;$$

- Without knowing the distribution of X_i's, we obtain via CLT

$$P\left(\theta_0 \in \left[\bar{X} \pm \frac{1.96}{\sqrt{n}}\right]\right) \to 95\%.$$

But the price is the "asymptotic" validity;

Starting from Stat 101

- Given that $X_1, .., X_n \overset{iid}{\sim} N(\theta_0, 1)$, we have

$$P\left(\theta_0 \in \left[\bar{X} \pm \frac{1.96}{\sqrt{n}}\right]\right) = 95\%;$$

- Without knowing the distribution of X_i's, we obtain via CLT

$$P\left(\theta_0 \in \left[\bar{X} \pm \frac{1.96}{\sqrt{n}}\right]\right) \to 95\%.$$

  But the price is the "asymptotic" validity;

- What if the distribution of X is unknown and n is finite?

A first thought: concentration inequality

- Concentration inequalities quantify how a random variable X deviates around its mean or median μ. They usually take the form of two-sided bounds for the tails of $X - \mu$, such as

$$P(|X - \mu| > t) \leq \text{something small.}$$

The simplest concentration inequality is Chebyshev inequality;

A first thought: concentration inequality

- Concentration inequalities quantify how a random variable X deviates around its mean or median μ. They usually take the form of two-sided bounds for the tails of $X - \mu$, such as

$$P(|X - \mu| > t) \leq \text{something small.}$$

  The simplest concentration inequality is Chebyshev inequality;
- Concentration inequality holds for any sample size n.

Does concentration inequality idea really work?

- For example, Hoeffding inequality gives

$$P\left(\theta_0 \in \left[\bar{X} \pm \frac{1.27}{\sqrt{n}} c\|X\|_{\psi_2}\right]\right) \geq 95\%$$

for iid X_i's with sub-Gaussian tails[3]. $\|X\|_{\psi_2}$ is the so-called sub-Gaussian norm, e.g., $\|X\|_{\psi_2} = 1/\sqrt{\log 2}$ for Rademacher X;

[3] Relaxable to sub-exponential tails via Bernstein inequality.

Does concentration inequality idea really work?

- For example, Hoeffding inequality gives

$$P\left(\theta_0 \in \left[\bar{X} \pm \frac{1.27}{\sqrt{n}} c\|X\|_{\psi_2}\right]\right) \geq 95\%$$

  for iid X_i's with sub-Gaussian tails[3]. $\|X\|_{\psi_2}$ is the so-called sub-Gaussian norm, e.g., $\|X\|_{\psi_2} = 1/\sqrt{\log 2}$ for Rademacher X;
- For illustration, let us examine one special cases, i.e., Bernoulli, in the class of sub-Gaussian random variables;

[3] Relaxable to sub-exponential tails via Bernstein inequality.

Does concentration inequality idea really work?

- For example, Hoeffding inequality gives

$$P\left(\theta_0 \in \left[\bar{X} \pm \frac{1.27}{\sqrt{n}} c\|X\|_{\psi_2}\right]\right) \geq 95\%$$

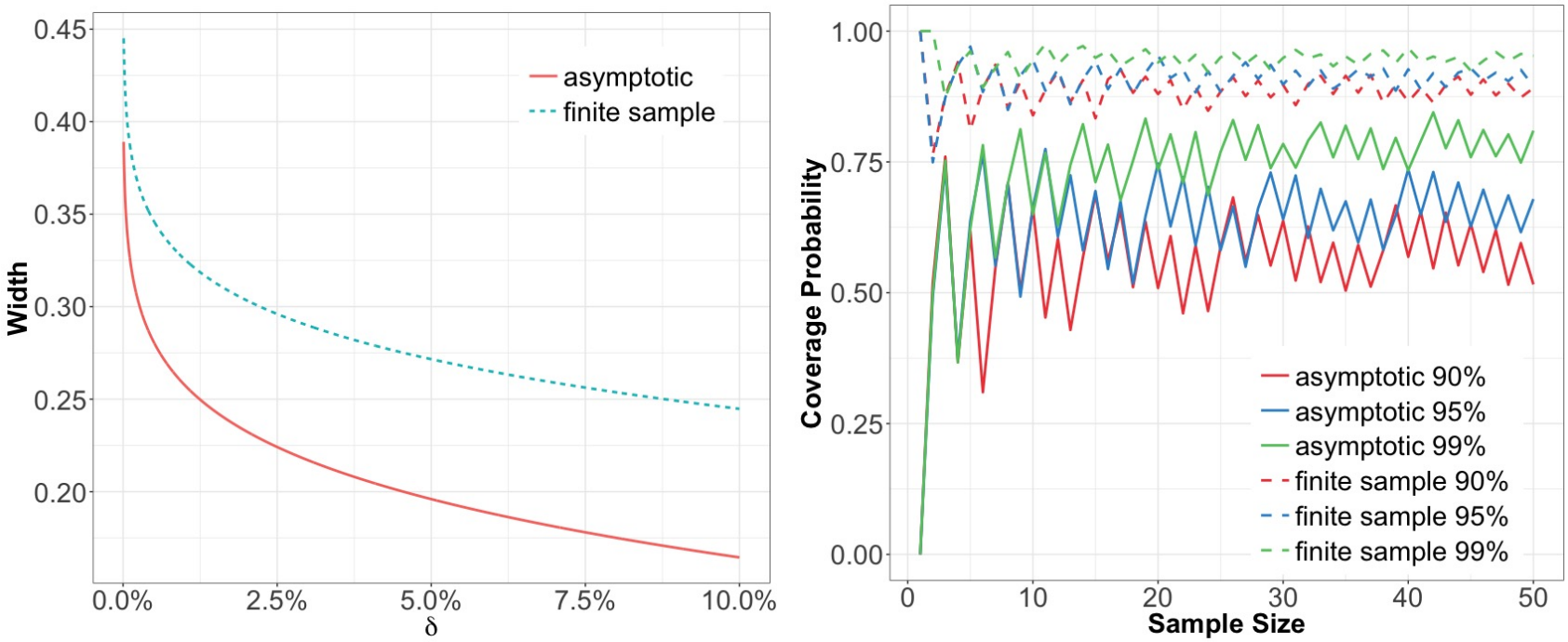
  for iid X_i's with sub-Gaussian tails[3]. $\|X\|_{\psi_2}$ is the so-called sub-Gaussian norm, e.g., $\|X\|_{\psi_2} = 1/\sqrt{\log 2}$ for Rademacher X;
- For illustration, let us examine one special cases, i.e., Bernoulli, in the class of sub-Gaussian random variables;
- When X_i's $\overset{iid}{\sim}$ Ber(θ_0), we have for any sample size n

$$P\left(\theta_0 \in [\bar{X} \pm 1.36/\sqrt{n}]\right) \geq 95\%,$$

  which is sharp in the rate but not the constant in comparison with the asymptotic CI, i.e., $[\bar{X} \pm 0.98/\sqrt{n}]$. Set $\theta_0 = 0.5$.

[3] Relaxable to sub-exponential tails via Bernstein inequality.

Empirical comparison for Bernoulli



Left plot compares width of confidence intervals of two methods; right plot demonstrate their coverage probability versus (small) sample size. Simulation were repeated 1000 times.

Two competing effects: Conservativness vs Asymptotics

How far can we go beyond these simple cases?

There are some recent results on more complicated "parametric models" obtained via various notion of concentration inequality such as Chernozhukov et al (2013) and Strawn et al (2014).

How far can we go beyond these simple cases?

There are some recent results on more complicated "parametric models" obtained via various notion of concentration inequality such as Chernozhukov et al (2013) and Strawn et al (2014).

- Chernozhukov et al (2013) presented non-asymptotic results on bootstrap validity for high dimensional problems such as $\|(1/\sqrt{n})\sum_{i=1}^{n} X_i\|_\infty$ for $X_i \in \mathbb{R}^p$ with $p \gg n$;

How far can we go beyond these simple cases?

There are some recent results on more complicated "parametric models" obtained via various notion of concentration inequality such as Chernozhukov et al (2013) and Strawn et al (2014).

- Chernozhukov et al (2013) presented non-asymptotic results on bootstrap validity for high dimensional problems such as $\|(1/\sqrt{n})\sum_{i=1}^{n} X_i\|_\infty$ for $X_i \in \mathbb{R}^p$ with $p \gg n$;
- Strawn et al (2014) obtained non-asymptotic bounds on the expected concentration of a posterior (around the true parameter) in high dimensional Bayesian linear models;

How far can we go beyond these simple cases?

There are some recent results on more complicated "parametric models" obtained via various notion of concentration inequality such as Chernozhukov et al (2013) and Strawn et al (2014).

- Chernozhukov et al (2013) presented non-asymptotic results on bootstrap validity for high dimensional problems such as $\|(1/\sqrt{n})\sum_{i=1}^{n} X_i\|_\infty$ for $X_i \in \mathbb{R}^p$ with $p \gg n$;
- Strawn et al (2014) obtained non-asymptotic bounds on the expected concentration of a posterior (around the true parameter) in high dimensional Bayesian linear models;
- As far as we are aware, these bounds were mostly derived to justify (asymptotic) statistical estimation/inference in a non-asymptotic manner, rather than used to construct finite-sample inference.

Outline

Smoothing spline models

- Consider a nonparametric regression model

$$y_i = f(x_i) + \epsilon_i.$$

For simplicity, we assume that $x \in \mathbb{I} := [0, 1]$, and that x and ϵ are independent with $E\epsilon = 0$ and $\mathrm{Var}(\epsilon) = 1$;

Smoothing spline models

- Consider a nonparametric regression model

$$y_i = f(x_i) + \epsilon_i.$$

  For simplicity, we assume that $x \in \mathbb{I} := [0,1]$, and that x and ϵ are independent with $E\epsilon = 0$ and $\mathrm{Var}(\epsilon) = 1$;

- The standard smoothing spline estimate (or more generally, kernel ridge estimate) is obtained as

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

reproducing kernel Hilbert space (RKHS)

A non-asymptotic Bahadur representation

Starting from a "naive" concentration inequality

We first investigate the concentration property of $T_n := \|\widehat{f}_n - f_0\|$.

Lemma 1

Under proper non-asymptotic conditions on $\lambda > 0$ and $M > 0$, it holds that

$$\sup_{f_0 \in H^m(1)} P_{f_0}\left(T_n \geq d_n(M,\lambda)\right) \leq 2\exp(-M), \tag{1}$$

where $d_n(M,\lambda) = 2\lambda^{1/m} + c_K(\sqrt{2M}+1)(n\lambda^{\frac{1}{2m}})^{-1/2}$.

Note that Lemma 1 holds uniformly over an unit ball

$$H^m(1) := \{f \in S^m(\mathbb{I}) : \|f\|_{\mathcal{H}} \leq 1\}$$

for any sample size.

A closer examination of the naive concentration

- We notice that

$$\begin{aligned} E\{\widehat{f}_n\} &\approx f_0 - \mathcal{P}_\lambda f_0 \\ &\neq f_0! \end{aligned}$$

The term $\mathcal{P}_\lambda f_0$ is the estimation bias caused by penalization;

A closer examination of the naive concentration

- We notice that

$$\begin{aligned} E\{\widehat{f}_n\} &\approx f_0 - \mathcal{P}_\lambda f_0 \\ &\neq f_0! \end{aligned}$$

  The term $\mathcal{P}_\lambda f_0$ is the estimation bias caused by penalization;
- If using $T_n = \|\widehat{f}_n - f_0\|$ to test $H_0 : f = f_0$, we prove that its minimal separation rate is $n^{-\frac{m}{2m+1}}$, which is sub-optimal in comparison with the minimax rate of testing $n^{-\frac{2m}{4m+1}}$;

A closer examination of the naive concentration

- We notice that

$$\begin{aligned} E\{\widehat{f}_n\} &\approx f_0 - \mathcal{P}_\lambda f_0 \\ &\neq f_0! \end{aligned}$$

  The term $\mathcal{P}_\lambda f_0$ is the estimation bias caused by penalization;
- If using $T_n = \|\widehat{f}_n - f_0\|$ to test $H_0 : f = f_0$, we prove that its minimal separation rate is $n^{-\frac{m}{2m+1}}$, which is sub-optimal in comparison with the minimax rate of testing $n^{-\frac{2m}{4m+1}}$;
- To make inference, we need to develop a second order expansion, which we call as "non-asymptotic" Bahadur representation.

Review on (asymptotic) Bahadur representation

- Bahadur representations (Bahadur, 1966) are often useful to study the "asymptotic" properties of statistical estimators;

Review on (asymptotic) Bahadur representation

- Bahadur representations (Bahadur, 1966) are often useful to study the "asymptotic" properties of statistical estimators;
- For example, He and Shao (1996) considered a general M-estimation framework:

$$(1/n)\sum_{i=1}^{n}\psi(x_i,\widehat{\theta}_n) = o(\delta_n)$$

for some $\delta_n \to 0$, and obtained that as $n \to \infty$

$$\widehat{\theta}_n - \theta_0 - \sum_{i=1}^{n} D_n^{-1}\psi(x_i,\theta_0) = O(R_n) \text{ almost surely.}$$

for some $R_n \ll o(n^{-1/2})$.

Non-asymptotic Bahadur representation

Inspired by Shang and Cheng (2013), we define

$$\widehat{T}_n := \left\| \widehat{f}_n - (I - \mathcal{P}_\lambda) f_0 - \frac{1}{n} \sum_{i=1}^{n} \epsilon_i K_{X_i} \right\|$$

and study its concentration property.

Theorem 2

Under similar conditions as Lemma 1, it holds that

$$\sup_{f_0 \in H^m(1)} P_{f_0} \left(\widehat{T}_n \geq \widehat{d}_n(M, \lambda) \right) \leq 2 \exp(-M),$$

where $\widehat{d}_n(M, \lambda) = c_K^2 \sqrt{M} n^{-1/2} \lambda^{-1/(2m)} A(\lambda) d_n(M, \lambda)$.

Statistical application: hypothesis testing

Nonparametric hypothesis of interest

Consider the following hypothesis testing problem:

$$H_0 : f = f_0, \ vs \ H_1 : f \neq f_0,$$

e.g., $f_0 = 0$.

Test statistic

- Clearly, $\widehat{T}_n = \|\widehat{f}_n - (I - \mathcal{P}_\lambda) f_0 - n^{-1} \sum_{i=1}^n \epsilon_i K_{X_i}\|$ cannot be used as a test statistic since ϵ_i's are not observable;

Test statistic

- Clearly, $\widehat{T}_n = \|\widehat{f}_n - (I - \mathcal{P}_\lambda) f_0 - n^{-1} \sum_{i=1}^n \epsilon_i K_{X_i}\|$ cannot be used as a test statistic since ϵ_i's are not observable;
- Rather, we propose to use $\widetilde{T}_n$ as test statistic:

$$\begin{aligned} \widetilde{T}_n &= \|\widehat{f}_n - (I - \mathcal{P}_\lambda) f_0\|^2 - E\|n^{-1} \sum_{i=1}^n \epsilon_i K_{X_i}\|^2, \\ &= \|\widehat{f}_n - (I - \mathcal{P}_\lambda) f_0\|^2 - \frac{1}{n} \sum_{\nu \geq 1} \frac{1}{1 + \lambda/\rho_\nu}; \end{aligned}$$

Test statistic

- Clearly, $\widehat{T}_n = \|\widehat{f}_n - (I - \mathcal{P}_\lambda) f_0 - n^{-1} \sum_{i=1}^n \epsilon_i K_{X_i}\|$ cannot be used as a test statistic since ϵ_i's are not observable;
- Rather, we propose to use $\widetilde{T}_n$ as test statistic:

$$\begin{aligned} \widetilde{T}_n &= \|\widehat{f}_n - (I - \mathcal{P}_\lambda) f_0\|^2 - E\|n^{-1} \sum_{i=1}^n \epsilon_i K_{X_i}\|^2, \\ &= \|\widehat{f}_n - (I - \mathcal{P}_\lambda) f_0\|^2 - \frac{1}{n} \sum_{\nu \geq 1} \frac{1}{1 + \lambda/\rho_\nu}; \end{aligned}$$

- In practice, replace $(I - \mathcal{P}_\lambda) f_0$ by a "noiseless" version of $\widehat{f}_n$:

$$\widehat{f}_n^{NL} = \arg \min_{f \in S^m(\mathbb{I})} \frac{1}{n} \sum_{i=1}^n (f_0(X_i) - f(X_i))^2 + \lambda \|f\|_{S^m(\mathbb{I})}.$$

Exact Type I and II errors control

Based on the non-asymptotic Bahadur representation, we can control on Type I and II errors for any finite sample size as follows.

Theorem 3

Under proper non-asymptotic conditions on λ, α and β, we have

$$\textit{Type I error}: \quad P_{f_0}(\widetilde{T}_n \geq \widetilde{d}_n(\log(15/\alpha), \lambda)) \leq \alpha,$$

$$\textit{Type II error}: \quad \sup_{f - f_0 \in H^m_{\alpha,\beta}(1)} P_f(\widetilde{T}_n \leq \widetilde{d}_n(\log(15/\alpha), \lambda)) \leq \beta,$$

for any $\alpha, \beta \in (0,1)$. The cut-off value

$$\widetilde{d}_n(M, \lambda) \approx \frac{4\sqrt{M}\rho_K}{n\lambda^{1/(4m)}}.$$

The separation set
$H^m_{\alpha,\beta}(1) := \{g \in H^m(1) : \|g\| \geq \rho_n(\log(15/\alpha), \log(30/\beta), \lambda)\}$, *where*
$\rho_n(M, L, \lambda) \approx \sqrt{\zeta_K \lambda} + \sqrt{4\rho_K\sqrt{M}/(nh^{1/(4m)})}$.

Asymptotic minimax optimality

- An important implication of Theorem 3 is that $\widetilde{T}_{n,\lambda}$ is asymptotically minimax optimal;

Asymptotic minimax optimality

- An important implication of Theorem 3 is that $\widetilde{T}_{n,\lambda}$ is asymptotically minimax optimal;
- To see that, we minimize the separation function $\rho_n(\log(15/\alpha), \log(30/\beta), \lambda)$ over λ, and obtain the minimal separation rate $n^{-2m/(4m+1)}$ when λ is chosen as

$$\begin{aligned} \lambda^* &:= \left(\left(\frac{4\rho_K}{\zeta_K} \right)^2 \log(15/\alpha) \right)^{2m/(4m+1)} n^{-4m/(4m+1)} \\ &\asymp n^{-4m/(4m+1)}. \end{aligned}$$

Asymptotic minimax optimality

- An important implication of Theorem 3 is that $\widetilde{T}_{n,\lambda}$ is asymptotically minimax optimal;
- To see that, we minimize the separation function $\rho_n(\log(15/\alpha), \log(30/\beta), \lambda)$ over λ, and obtain the minimal separation rate $n^{-2m/(4m+1)}$ when λ is chosen as

$$\begin{aligned} \lambda^* &:= \left(\left(\frac{4\rho_K}{\zeta_K} \right)^2 \log(15/\alpha) \right)^{2m/(4m+1)} n^{-4m/(4m+1)} \\ &\asymp n^{-4m/(4m+1)}. \end{aligned}$$

- Following similar derivations, we prove that the test statistic based on the naive concentration, i.e., $T_n = \|\widehat{f}_n - f_0\|$, is actually sub-optimal. So, our intuition of not using it is correct.

Extensions

- Composite Hypothesis, e.g.,

$$H_0 : f \in \mathcal{F}_0 \text{ vs } H_1 : f \notin \mathcal{F}_0,$$

where $\mathcal{F}_0 = \{f : f \text{ is linear on } \mathbb{I}\}$;

Extensions

- Composite Hypothesis, e.g.,

$$H_0 : f \in \mathcal{F}_0 \text{ vs } H_1 : f \notin \mathcal{F}_0,$$

  where $\mathcal{F}_0 = \{f : f \text{ is linear on } \mathbb{I}\}$;
- Kernel ridge regression with different decaying rates of eigen-values;

Extensions

- Composite Hypothesis, e.g.,

$$H_0 : f \in \mathcal{F}_0 \text{ vs } H_1 : f \notin \mathcal{F}_0,$$

  where $\mathcal{F}_0 = \{f : f \text{ is linear on } \mathbb{I}\}$;
- Kernel ridge regression with different decaying rates of eigen-values;
- Non-Gaussian regression with a smooth log-concave density such as logistic regression.

Simulations

Simulations

Consider the following hypothesis:

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

where $f_0 = 5(x^2 - x + \frac{1}{6})$.

Simulations

Consider the following hypothesis:

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

where $f_0 = 5(x^2 - x + \frac{1}{6})$.

- Data were generated as follows

$$Y_i = f_c(X_i) + \epsilon_i, \qquad \text{and} \quad f_c(x) = \frac{1}{2} c\, x^2 + f_0(x);$$

Simulations

Consider the following hypothesis:

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

where $f_0 = 5(x^2 - x + \frac{1}{6})$.

- Data were generated as follows

$$Y_i = f_c(X_i) + \epsilon_i, \qquad \text{and} \quad f_c(x) = \frac{1}{2} c\, x^2 + f_0(x);$$

- Set $\epsilon_i \overset{iid}{\sim} N(0,1)$ and $X_i \overset{iid}{\sim} Unif(0,1)$;

Simulations

Consider the following hypothesis:

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

where $f_0 = 5(x^2 - x + \frac{1}{6})$.

- Data were generated as follows

$$Y_i = f_c(X_i) + \epsilon_i, \qquad \text{and} \quad f_c(x) = \frac{1}{2} c\, x^2 + f_0(x);$$

- Set $\epsilon_i \overset{iid}{\sim} N(0,1)$ and $X_i \overset{iid}{\sim} Unif(0,1)$;
- Set Type I and II errors as $\alpha = \beta = 0.05$.

Remark: cutoff value

- Recall that the cutoff value $\tilde{d}_n(\log(15/\alpha), \lambda)$ is provided in Corollary 3. However, we found that it is quite conservative due to the use of some loose concentration inequalities;

Remark: cutoff value

- Recall that the cutoff value $\tilde{d}_n(\log(15/\alpha), \lambda)$ is provided in Corollary 3. However, we found that it is quite conservative due to the use of some loose concentration inequalities;
- Fortunately, we can simulate $\widetilde{T}_n$ as follows. By conditioning on X, we simulate a number of synthetic datasets $\{\mathbf{Y}^{(k)}\}_{k=1}^N$ from the null model

$$\mathbf{Y}_i^{(k)} = f_0(X_i) + N(0,1),$$

each of which yields a new test statistic $\widetilde{T}_n^{(k)}$. The cutoff value is set as the $(1-\alpha)$-th sample quantile of $\{\widetilde{T}_n^{(k)}\}_{k=1}^N$.

Remark: the choice of λ

- Rather, our theoretical analysis is useful in choosing λ;

Remark: the choice of λ

- Rather, our theoretical analysis is useful in choosing λ;
- In simulations, we choose λ by numerically minimizing the separation function

$$\rho_n(\log(15/\alpha), \log(30/\beta), \lambda)$$

over λ. The resulting λ is denoted as λ_{FS};

Remark: the choice of λ

- Rather, our theoretical analysis is useful in choosing λ;
- In simulations, we choose λ by numerically minimizing the separation function

$$\rho_n(\log(15/\alpha), \log(30/\beta), \lambda)$$

  over λ. The resulting λ is denoted as λ_{FS};
- This non-asymptotic approach is much computationally cheaper than the conventional generalized cross validation;

Remark: the choice of λ

- Rather, our theoretical analysis is useful in choosing λ;
- In simulations, we choose λ by numerically minimizing the separation function

$$\rho_n(\log(15/\alpha), \log(30/\beta), \lambda)$$

  over λ. The resulting λ is denoted as λ_{FS};
- This non-asymptotic approach is much computationally cheaper than the conventional generalized cross validation;
- Note that all constants in $\rho_n(M, L, \lambda)$ and $d_n(M, \lambda)$ such as ρ_K, ζ_K and c_K only depend on λ and the eigenvalues ρ_ν's, which can be well approximately by the empirical eigenvalues of the reproducing kernel matrix.

Empirical comparison with "asymptotically" valid test

For simple hypothesis, we compare four testing procedures:

(S1) The proposed $\widetilde{T}_n$ with the smoothing parameter λ_{FS};

(S2) Asymptotically valid penalized likelihood ratio test (Shang and Cheng, 2013), denoted as $PLRT(f_0)$, with the same λ_{FS};

(S3) The proposed $\widetilde{T}_n$ with λ selected by GCV, denoted as λ_{GCV};

(S4) PLRT statistic $PLRT(f_0)$ with the same λ_{GCV}.

Note that the cut-off value for PLRT in S2 and S4 is obtained from Monte Carlo simulation and the null limit distribution given in Shang and Cheng (2013), respectively.

Simulation results

n	c	$\lambda_{FS}^{1/4}$	RP_{FS}	RP_{PLRT}	$\lambda_{GCV}^{1/4}$	RP_{GCV}	RP'_{GCV}
50	0	0.126	0.046	0.052	0.142(0.008)	0.052	0.054
	1		0.100	0.092	0.138(0.009)	0.088	0.084
	2		0.396	0.340	0.134(0.009)	0.323	0.310
	3		0.822	0.764	0.128(0.009)	0.752	0.748
100	0	0.108	0.048	0.048	0.122(0.007)	0.053	0.052
	1		0.264	0.170	0.117(0.008)	0.167	0.144
	2		0.650	0.558	0.112(0.007)	0.493	0.474
	3		0.976	0.934	0.110(0.003)	0.924	0.918
200	0	0.092	0.050	0.048	0.104(0.006)	0.051	0.050
	1		0.368	0.334	0.102(0.007)	0.325	0.290
	2		0.896	0.862	0.098(0.006)	0.832	0.816
	3		1.00	1.00	0.094(0.003)	1.00	1.00
300	0	0.084	0.046	0.048	0.096(0.006)	0.048	0.048
	1		0.426	0.404	0.094(0.006)	0.397	0.394
	2		0.968	0.946	0.092(0.005)	0.930	0.914
	3		1.00	1.00	0.090(0.005)	1.00	1.00
400	0	0.079	0.052	0.050	0.091(0.004)	0.049	0.054
	1		0.668	0.640	0.087(0.005)	0.631	0.618
	2		1.00	1.00	0.085(0.004)	1.00	1.00
	3		1.00	1.00	0.084(0.003)	1.00	1.00

TABLE 1

Simulation results for simple hypothesis testing. λ_{GCV} *is an average value over* 500 *replicates (that varies as* c*).* RP_{FS}, RP_{PLRT}, RP_{GCV} *and* RP'_{GCV} *are average rejection proportions by* $\widetilde{T}_n$ *with* λ_{FS}, *PLRT with* λ_{FS}, $\widetilde{T}_n$ *with* λ_{GCV} *and PLRT with* λ_{GCV} *respectively, over* 500 *replicates.*

Empirical observations

- Overall, all four procedures have comparable type I errors, i.e., $c = 0$, for any sample size;

Empirical observations

- Overall, all four procedures have comparable type I errors, i.e., $c = 0$, for any sample size;
- As for power performances, we note that

Empirical observations

- Overall, all four procedures have comparable type I errors, i.e., $c = 0$, for any sample size;
- As for power performances, we note that
  - (i) the test using λ_{FS} is always more powerful than those using λ_{GCV}. This justifies the finite sample advantage of the non-asymptotic formula in selecting λ;

Empirical observations

- Overall, all four procedures have comparable type I errors, i.e., $c = 0$, for any sample size;
- As for power performances, we note that
  - (i) the test using λ_{FS} is always more powerful than those using λ_{GCV}. This justifies the finite sample advantage of the non-asymptotic formula in selecting λ;
  - (ii) $\widetilde{T}_n$ is always more powerful than PLRT given the same choice of λ. In other words, S1 is always the most powerful one. This supports the need of removing estimation bias in nonparametric testing;

Empirical observations

- Overall, all four procedures have comparable type I errors, i.e., $c = 0$, for any sample size;
- As for power performances, we note that
  - (i) the test using λ_{FS} is always more powerful than those using λ_{GCV}. This justifies the finite sample advantage of the non-asymptotic formula in selecting λ;
  - (ii) $\widetilde{T}_n$ is always more powerful than PLRT given the same choice of λ. In other words, S1 is always the most powerful one. This supports the need of removing estimation bias in nonparametric testing;
- The third observation is that as c increases, λ_{GCV} continues decreasing and becomes closer to λ_{FS}, but never reaches λ_{FS}. This is consistent with their different asymptotic orders, i.e., $\lambda_{FS} \asymp n^{-2m/(2m+1/2)}$ and $\lambda_{GCV} \asymp n^{-2m/(2m+1)}$.

Discussions

- The "non-asymptotic Bahadur representation" seems a possible direction to pursue in finite-sample inference;

Discussions

- The "non-asymptotic Bahadur representation" seems a possible direction to pursue in finite-sample inference;
- The practical usefulness of finite-sample inference procedures rely on "computably sharp" concentration inequality. Can this be done through data re-sampling or MCMC?

Discussions

- The "non-asymptotic Bahadur representation" seems a possible direction to pursue in finite-sample inference;
- The practical usefulness of finite-sample inference procedures rely on "computably sharp" concentration inequality. Can this be done through data re-sampling or MCMC?
- Does it exist something like "non-asymptotic" Fisher information?

Discussions

- The "non-asymptotic Bahadur representation" seems a possible direction to pursue in finite-sample inference;
- The practical usefulness of finite-sample inference procedures rely on "computably sharp" concentration inequality. Can this be done through data re-sampling or MCMC?
- Does it exist something like "non-asymptotic" Fisher information?
- Can the non-asymptotic approach be extended to Bayesian domain, e.g., choose hyper-parameter in hierarchical priors?

Questions?

Thanks to ONR for support

Some necessary notation

- In smoothing spline models, we set $\mathcal{H}$ as an m-th order Sobolev space, denoted as $S^m(\mathbb{I})$, and $\|f\|^2_{\mathcal{H}}$ as $\int_{\mathbb{I}}\{f^{(m)}(x)\}^2dx$;
- Define $\langle f, g\rangle = E\{f(X)g(X)\} + \lambda\int_{\mathbb{I}} f^{(m)}(x)g^{(m)}(x)dx$. Endowed with $\langle\cdot,\cdot\rangle$, $S^m(\mathbb{I})$ is an RKHS;
- Let $K(x_1, x_2)$ be the reproducing kernel function satisfying $\langle K_x, f\rangle = f(x)$ with $K_x(\cdot) = K(x,\cdot)$. The underlying eigen-system is denoted as $\{\rho_\nu, \phi_\nu(\cdot)\}_{\nu=1}^{\infty}$, where $\rho_\nu \asymp \nu^{-2m}$;
- For later use, define

$$\begin{aligned} c_K &= \lambda^{\frac{1}{4m}} \sup_{x\in\mathbb{I}} K^{1/2}(x,x), \\ \rho_K &= \lambda^{\frac{1}{2m}} E[K^2(X_1, X_2)] \text{ with } X_1, X_2 \overset{iid}{\sim} X. \end{aligned}$$