Tutorial

on

Statistical Inference

Larry Wasserman
Carnegie Mellon University

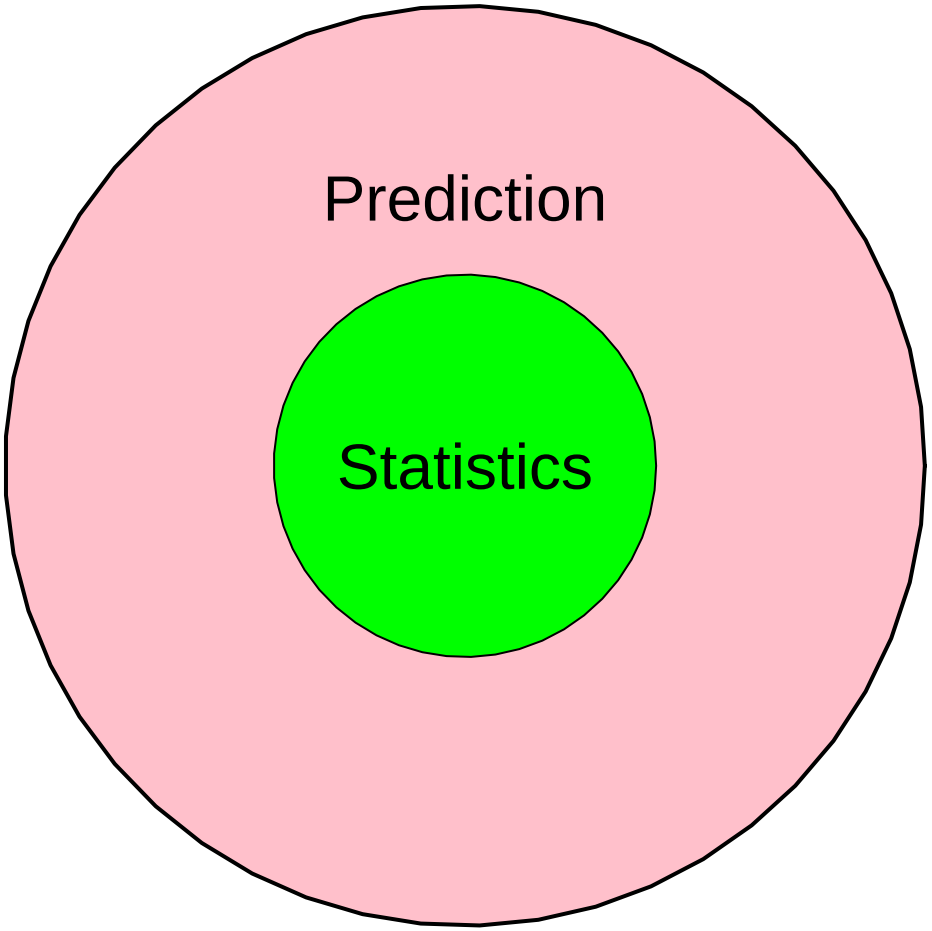
February 2010

Outline

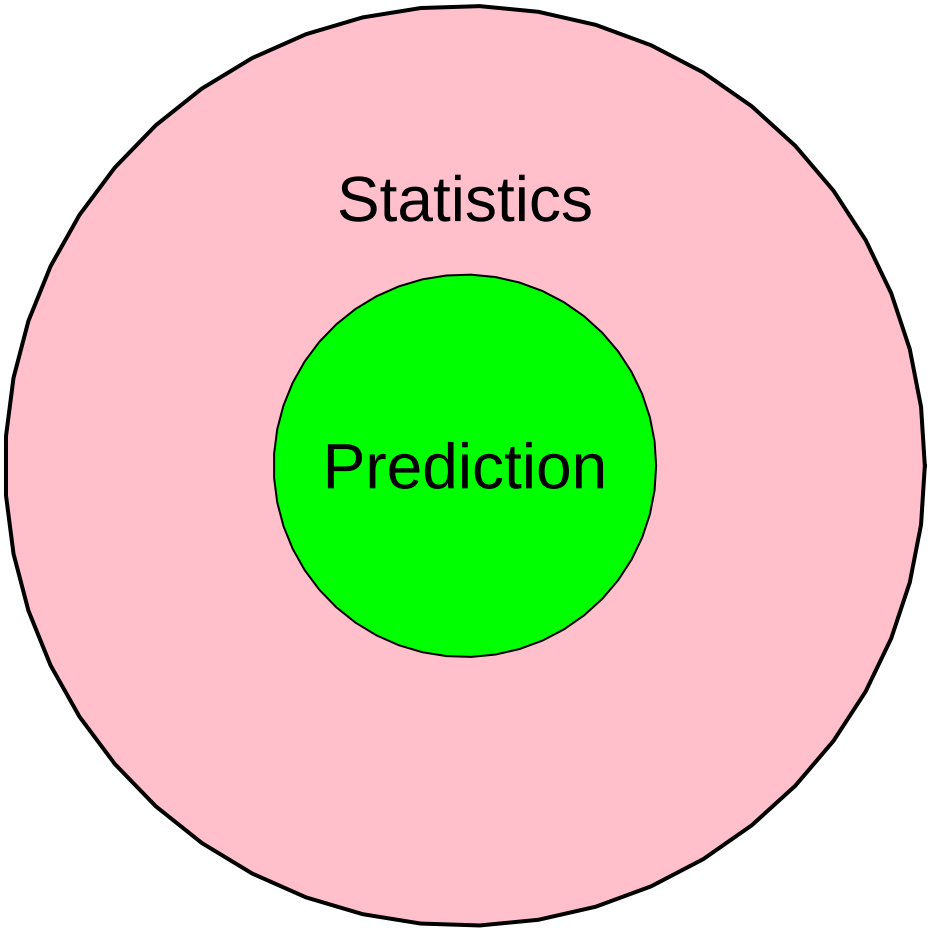
- CS versus Statistics
- Background
- Minimax Theory
- (Confidence Sets)
- (Bayes versus Frequentist)
- (Robustness)
- Statistical View of Differential Privacy

An Oversimplified Description of Statistics versus CS

The View From Computer Science



The View From Statistics



CS:

o what algorithm should I use (or invent)?

o what are the properties of the algorithm? (running time, complexity)

Statistics:

o what assumptions about the data are reasonable?

o what is the best we can do under those assumptions?

o how do I design an estimator (predictor, algorithm) to achieve this performance?

Prediction

- CS view:

o training data $\Longrightarrow$ algorithm

$$X \Longrightarrow \boxed{\text{Algorithm}} \Longrightarrow \text{ prediction}$$

o properties of the algorithm

- Statistics view:

o model: $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P \in \mathcal{P}$

o what $\mathcal{P}$ is reasonable?

o optimal method: find $\widehat{m}$ such that

$$\sup_{P \in \mathcal{P}} \mathbb{E}(Y - \widehat{m}(X))^2 = \inf_{\widehat{m}} \sup_{P \in \mathcal{P}} \mathbb{E}(Y - \widehat{m}(X))^2$$

$$X \Longrightarrow \boxed{\text{Algorithm}} \Longrightarrow \text{prediction } \widehat{m}(X)$$

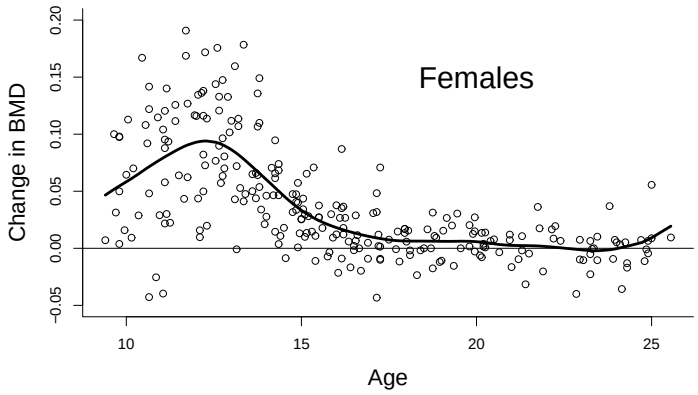
o Also interested in $\widehat{m}(x)$ and confidence interval for $\widehat{m}(x)$

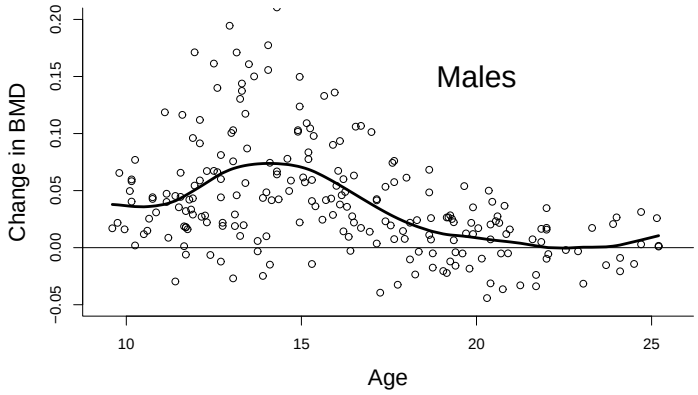
* "optimal" refers to optimality in the minimax sense. There are other notions of optimality.

- Example:

Predict change in bone mineral density Y from age X

o best predictor is the regression function $m(x) = \mathbb{E}(Y|X = x)$





In addition to predicting Y from X we want to know:

o what is the best estimator of m given smoothness assumptions?
o what features in $\widehat{m}$ are real?
o confidence band for $m(x)$
o how different are men and woman?

Background

Terminology

Statistics	CS
Estimation	Learning
Classifier	Hypothesis
Classification	Classification/learning
Regression	Regression
Confidence Interval	????
Kernels	Parzen windows
Mercer Kernels	Kernels
????	semisupervised learning
distribution free	agnostic
sequential design	active learning
$O(\cdot)$	$O(\cdot)$
$a_n \asymp b_n$	$a_n = \Theta(b_n)$
$b_n = O(a_n)$	$a_n = \Omega(b_n)$

Models

$$X_1, \ldots, X_n \sim P$$

- A statistical model is a collection $\mathcal{P}$ of probability distributions.
- Parametric model: $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$
- Example: $\mathcal{P} =$ all Gaussian distributions
- Nonparametric model: $\mathcal{P}$ is infinite dimensional.
- Example: $\mathcal{P}$ is all distributions.
- Example: Sobolev space:

$$\mathcal{P} = \left\{ P : \ p = \frac{dP}{d\mu}, \ \int (p''(x))^2 dx \leq C \right\}$$

Typical Statistical Problems

- Estimate a parameter $\theta = T(P)$ such as the mean:

  minimax risk $$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P ||\widehat{\theta} - \theta(P)||^2$$

- Construct a confidence interval C_n:

  coverage $$\inf_{P \in \mathcal{P}} P(\theta(P) \in C_n) \geq 1 - \alpha.$$

- Nonparametric Density Estimation: $Y_1, \ldots, Y_n \sim p$:

  minimax risk $$\inf_{\widehat{p}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \int (\widehat{p} - p)^2.$$

Typical Statistical Problems

- Nonparametric Regression: $(X_1, Y_1), \ldots, (X_n, Y_n)$: Estimate $m(x) = \mathbb{E}(Y| = x)$. i.e. Predict $Y \in \mathbb{R}$.

  minimax risk $\inf_{\widehat{m}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \int (\widehat{m} - m)^2.$

- Predict $Y \in \{0, 1\}$ i.e. classification.

  minimax risk $\inf_{\widehat{h}} \sup_{P \in \mathcal{P}} P(h(X) \neq Y).$

Empirical Measures

- Empirical measure $\widehat{P}_n$ puts mass $1/n$ at each X_i.

- Glivenko-Cantelli Theorem:

$$\sup_{A \in \mathcal{A}} |P(A) - \widehat{P}_n(A)| \xrightarrow{P} 0$$

of $\mathcal{A}$ is a VC class.

- Exponential inequality:

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon \right) \leq Ce^{-nc\epsilon^2}$$

- Donsker's Theorem:

$$\{\sqrt{n}(\widehat{P}_n(A) - P(A)) : \ A \in \mathcal{A}\} \rightsquigarrow \mathbb{B} = \text{Brownian bridge}$$

This means that $\{\sqrt{n}(\widehat{P}_n(A) - P(A))$ is approximately Normal, uniformly over A.

- The concern with asymptotic Normality seems to be an important difference between CS and Statistics.

Empirical Measures

- Parameter (or functional) $\theta = T(P)$
- Plug-in estimator: $\widehat{\theta} = T(\widehat{P}_n)$
- Example: If θ = mean then $\widehat{\theta} = T(P_n) = \overline{X}$ = sample mean
- Example: If θ = first eigenvector of $\Sigma = \mathbb{E}(X - \mu)(X - \mu)^T$ then $\widehat{\theta} = T(P_n)$ = PCA

Many Normal Means

$$Y_i = \mu_i + \sigma\epsilon_i, \quad i = 1, 2, \ldots, \qquad \epsilon_i \sim N(0, 1)$$

This is a surprisingly rich "laboratory" for doing theory. It can be shown that many problems (density estimation, nonparametric regression, etc) are "statistically isomorphic" to this model.

The Sobolev space of order p in function space corresponds to the Sobolev ellipsoid:

$$\Theta = \left\{ \mu : \sum_{i=1}^{\infty} \mu_i^2 \, i^{2p} \leq C \right\}.$$

Minimax Estimation

Minimax Estimation

- $\theta = \theta(P)$
- θ can be a parameter, a function, a prediction etc.
- estimator $\widehat{\theta} = g(X_1, \ldots, X_n)$
- loss function $L(\theta, \widehat{\theta})$: example: $(\widehat{\theta} - \theta)^2$
- minimax risk:

$$R_n = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \Big(L(\theta(P), \widehat{\theta}) \Big)$$

Minimax Estimation

Goals:

- compute the minimax risk
- find lower and upper bounds on R_n
- compute the minimax rate: $R_n \asymp r_n$
- find an estimator that achieves the minimax risk (or at least the minimax rate)

Tools for lower bounds:

- Fano's inequality
- Assouad's lemma
- LeCam's lemma
- Bayes estimators with constant risk

Fano's Inequality

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} E_P(d(\widehat{\theta}, \theta(P))) \geq \inf_{\widehat{\theta}} \sup_{P \in F} E_P(d(\widehat{\theta}, \theta(P)))$$

where $F = \{P_0, P_1, \ldots, P_M\}$

Suppose that

$$d(\theta_j, \theta_k) \geq \psi_n$$

and

$$\frac{1}{M} \sum_{j=1}^{M} K(P_j, P_0) \leq \frac{\log M}{16}$$

where $K(P, Q) = \int p \log(p/q)$, then

$$\inf_{\widehat{\theta}} \sup_{\theta \in F} E_\theta(d(\widehat{\theta}, \theta)) \geq C\psi_n$$

(This version is due to Tsybakov (2003).)

Examples

- the mean of a Normal
- maximum likelihood
- functionals
- densities
- Sobolev spaces
- high dimensional classification
- semi-supervised learning
- manifold learning

Example: Normal

$X_1, \ldots, X_n \sim N(\theta, 1)$

$L(\theta, \widehat{\theta}) = \ell(||\theta - \widehat{\theta}||)$ where ℓ is bowl-shaped (convex, symmetric level sets)

The unique minimax estimator (over all bowl-shaped loss functions) is

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}.$$

Example: Maximum Likelihood

$X_1, \ldots, X_n \sim P$

$P \in \mathcal{P} = \{P_\theta : \ \theta \in \Theta\}$

$$\widehat{\theta}_n = \arg\max_{\theta \in \Theta} L(\theta)$$

where

$$L(\theta) = \prod_{i=1}^{n} p_\theta(X_i)$$

where p_θ is the density of P_θ.

When is Maximum Likelihood Minimax?

Short answer: For typical parametric models, fixed dimension, large sample sizes (and some regularity conditions) the mle is (aproximately) minimax.

Long answer: (Le Cam and Hajek): Under certions conditions:

$$\sup_F \liminf_{n\to\infty} \sup_{h\in F} \mathbb{E}_{\theta_n} \ell(\sqrt{n}(\widehat{\theta}_n - \theta_n)) \geq \text{Risk(mle)}$$

where $\theta_n = \theta + h/\sqrt{n}$ and F varies over all finite sets.

- This fails apart for nonparametric problems
- It also fails apart for high dimensional parametric problems (see Martin's tutorial)

Example: Functionals

(Donoho and Liu, 1991).

Let $\theta = T(P)$. Define the modulus of continuity

$$\omega(\epsilon) = \sup\{|T(P) - T(Q)| : \ H(P,Q) \leq \epsilon, \quad P \in \mathcal{P}\}$$

where $H^2(P,Q) = \int(\sqrt{p} - \sqrt{q})^2$. If T is a linear functional and $\mathcal{P}$ is convex then

$$\inf_{T_n} \sup_{P \in \mathcal{P}} \mathbb{E}_P \ell(T_n - T(F)) = \Theta(\omega(n^{-1/2})).$$

The lower bound is valid over all functionals.

Examples:
o $T(P) = p(x)$
o $T(P) = \int p^2(x)dx$
o $T(P) =$ mode of p

Example: Estimating a Density Function

Let $X_1, \ldots, X_n \sim P$ where $X_i \in \mathbb{R}^d$ and

$$P \in \mathcal{P} = \left\{ P : \ p = \frac{dP}{d\mu}, \ \int (p''(x))^2 dx \leq C \right\}.$$

Let $L(p, \widehat{p}) = \int (p(x) - \widehat{p}(x))^2 dx$. Then, there exists a universal constant $C > 0$ such that

$$\inf_{\widehat{p}} \sup_{P \in \mathcal{P}} \mathbb{E}_P L(p, \widehat{p}) \geq \frac{C}{n^{4/(4+d)}}.$$

Furthermore, if we use the kernel estimator

$$\widehat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} \ K\left(\frac{||x - X_i||}{h} \right)$$

with kernel K and bandwidth $h \asymp n^{-1/(4+d)}$ then

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P L(p, \widehat{p}) \leq \frac{c}{n^{4/(4+d)}}.$$

Example: Sobolev Spaces

$Y_i = N(\mu_i, \sigma^2),\ i = 1, 2, \ldots,$

$$\Theta = \left\{ \mu : \ \sum_{i=1}^{\infty} \mu_i^2 i^{2p} \leq C \right\}.$$

Pinsker's theorem:

$$\inf_{\widehat{\mu}} \sup_{\mu \in \Theta} \mathbb{E}_\mu ||\widehat{\mu} - \mu||^2 \sim \frac{A_p}{n^{2p/(2p+1)}}$$

where

$$A_p = \left(\frac{\sigma}{\pi}\right)^{2p/(2p+1)} e^{2/(2p+1)} \left(\frac{p}{p+1}\right)^{2p/(2p+1)} (2p+1)^{1/(2p+1)}$$

There is a known estimator that achieves the minimax risk.

There is also a known estimator that achieves the minimax risk without knowledge of p. This is called an adaptive minimax estimator.

Example: High-Dimensonal Classification

Why can we classify well in high dimensions? Here is a minimax explanation. (Audibert and Tsybakov 2007, Kohler and Krzyzak 2006).

Risk:

$$R(h) = \mathbb{P}(Y \neq h(X)) - \mathbb{P}(Y \neq h_{\text{Bayes}}(X)).$$

One expects a minimax rate of

$$O\left(\frac{1}{n^{\beta/(2\beta+d)}}\right)$$

where β is the smoothness of

$$m(x) = \mathbb{E}(Y|X = x)$$

and $d =$ dimension(X). This is $O(1)$ as $d \to \infty$.

But ...

High-Dimensional Classification

Recall that the Bayes classifier is

$$h(x) = \begin{cases} 1 & \text{if } m(x) \geq 1/2 \\ 0 & \text{if } m(x) < 1/2. \end{cases}$$

Low noise condition (large margin):

$$\mathbb{P}\left(\left|m(X) - \frac{1}{2}\right| \leq t\right) \leq Ct^{\alpha}.$$

If α is large, the classes are well-separated.

Then:

$$\inf_h \sup_{P \in \mathcal{P}} R(h) \geq Cn^{-\beta(1+\alpha)/(2\beta+d)}$$

High-Dimensional Classification

If we use the plug-in classifier:

$$h(x) = I(\widehat{m}(x) > 1/2)$$

where $\widehat{m}$ is the kernel regression estimator with bandwidth $h = n^{-1/(2\beta+d)}$ then

$$\sup_{P \in \mathcal{P}} R(h) \leq C' n^{-\beta(1+\alpha)/(2\beta+d)}.$$

This rate behaves like $O(1/n)$ when α is large. In fact, when $\alpha = \infty$, the rate is e^{-cn}.

Moral: Fast rates come from the assumption not the classifier.

Example: Semi-supervised Inference

Two minimax analyses: Lafferty and Wasserman (2007) and Singh, Nowak and Zhu (2008).

Labeled data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and unlabeled data $(X_1, \ldots, X_N)$. Want to classify or to estimate

$$m(x) = \mathbb{E}(Y|X = x).$$

Cluster assumption: m is smooth over clusters of the marginal $p(x)$.

Lafferty and Wasserman (2007) showed that may/may not improve the minimax rate of convergence depending on the assumptions.

Example: Semi-supervised Inference

Singh, Nowak and Zhu (2008) obtained the following upper and lower bounds, based on distance γ between clusters:

γ	semi	non-semi	SSL helps?
I	$n^{-2\alpha/(2\alpha+d)}$	$n^{-2\alpha/(2\alpha+d)}$	NO
II	$n^{-2\alpha/(2\alpha+d)}$	$n^{-2\alpha/(2\alpha+d)}$	NO
III	$n^{-2\alpha/(2\alpha+d)}$	$n^{1/d}$	YES
IV	$n^{-1/d}$	$n^{1/d}$	NO
V	$n^{-2\alpha/(2\alpha+d)}$	$n^{1/d}$	YES
VI	$n^{-2\alpha/(2\alpha+d)}$	$n^{1/d}$	YES

Benefit of stating models precisely: can say when an algorithm does/does not work.

Example: Estimating a Manifold

(From Genovese, Perone-Pacifico, Verdinelli, Wasserman 2010).

$$Y_i = f(U_i) + \epsilon, \qquad i = 1, \ldots, n$$

where $Y_i \in \mathbb{R}^D$. Here, $U_1, \ldots, U_n$ are unobserved and $U_i \in [0,1]^d$ with $d < D$.

$$f : [0,1]^d \to \mathbb{R}^D$$

and the image of f is a smooth manifold M. Suppose that ϵ_i has support on a ball of radius σ.

How well can we estimate M?

Example: Estimating a Manifold

Specific case: $D = 2$ and $d = 1$ so $\mathcal{M}$ is a curve.

If a ball of radius Δ can roll freely on $\mathcal{M}$ and if $\sigma < \Delta$. Then

$$\inf_{\widehat{M}} \sup_{P \in \mathcal{P}} \mathbb{E}_p d_H(M, \widehat{M}) \geq \frac{C}{n^{2/3}}$$

where

$$d_H(A, B) = \inf \{\epsilon : \ A \subset B^\epsilon, \ B \subset A^\epsilon\}$$

is the Hausdorff distance and

$$A^\epsilon = \bigcup_{x \in A} B(x, \epsilon).$$

Example: Estimating a Manifold

Originally, we estimated the support S of Y_i by

$$\widehat{S} = \bigcup_{i=1}^{n} B(Y_i, \epsilon_n)$$

and then took $\widehat{M}$ to be the medial axis (middle) of $\widehat{S}$.

But... it turns out that this does not achieve the minimax rate.

The minimax bound is achieved by:
o using a smooth estimate of S
o finding the center (medial axis) of the support.

Adaptivity

Let $\{\Theta_\alpha\}$ be a collection of spaces. Suppose that the minimax risk is $r_n(\alpha)$ for Θ_α:

$$r_n(\alpha) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_\alpha} \mathbb{E}_\theta L(\widehat{\theta}, \theta).$$

Can we find an estimator $\widehat{\theta}$ such that, for each α,

$$\sup_{\theta \in \Theta_\alpha} \mathbb{E}_\theta L(\widehat{\theta}, \theta) \asymp r_n(\alpha)$$

without knowledge of α?

Example: Wavelet Regression

$$Y_i = f(X_i) + \epsilon_i$$

Assume that $f \in B^{\sigma}_{p,q}$ (Besov space).

The minimax rate depends on (p, q, σ).

Donoho et al:

- expand f in a wavelet basis: $f = \sum_{j,k} \beta_{j,k} \psi_{j,k}(x)$.
- Set $Z_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \psi_j(X_i)$.
- Set $\widehat{\beta}_j = \text{soft}(Z_j)$ where

$$\text{soft}(x) = \text{sign}(x)(|x| - \lambda)_+$$

then $\widehat{f}(x) = \sum_j \widehat{\beta}_j \psi_j(x)$ is adaptive minimax, that is, it achieves the minimax risk over $B^{\sigma}_{p,q}$ without knowledge of (p, q, σ).

Confidence Sets

Find $C_n = C_n(X_1, \ldots, X_n)$ such that

$$\inf_{\theta \in \Theta} P_\theta(\theta \in C_n) \geq 1 - \alpha$$

Very different from prediction. There are NO adaptive confidence sets (Low 1997, Genovese and Wasserman 2007): If

$$Y = m(X) + \sigma\epsilon$$

with $m \in$ Lipschitz(s):

$$|m(x) - m(y)| \leq s|x - y|$$

and $0 \leq s \leq L$ then, if

$$\mathbb{P}(\ell \leq m \leq u) \geq 1 - \alpha$$

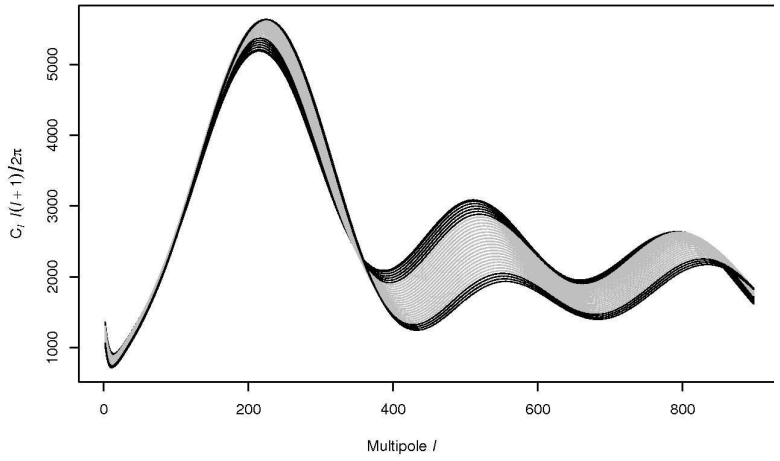
then

$$||u - l||_\infty = \left(\frac{\log n}{n}\right)^{1/3} \times \left(\frac{L\sigma^2}{2}\right)^{1/3} \times (1 + o(1)).$$

Thus, cannot adapt to $s < L$.

Example: CMB

Example: Estimating the power spectrum of the cosmic microwave backgound radiation (CMB) from WMAP. Peaks give vital information about dark matter, comsological parameters etc. (Genovese, Miller, Nichol, Arjunwadkar, Wasserman 2004).



Bayesian Inference

Frequentist View:

o probability means long run frequency

o θ is fixed, X is random

o procedures have frequency guarantees

Frequentist confidence interval: C_n:

$$\inf_{\theta \in \Theta} P_\theta(\theta \in C_n) = 1 - \alpha$$

Bayesian View:

o probability is subjective degree-of-belief

o θ is a random variable

o procedures do not have frequency guarantees

Bayesian interval:

$$P(\theta \in C_n | D) = 1 - \alpha$$

Bayesian Inference

The two approches are not always compatible. That is, we can have $P(\theta \in C_n|D) = 1 - \alpha$ and yet

$$\inf_{\theta \in \Theta} P_\theta(\theta \in C_n) \approx 0.$$

Bayes versus Frequentist

Example: (Robins and Ritov)

$$(X_1, R_1, Y_1), \ldots, (X_n, R_n, Y_n)$$

where

$$X_i \in \mathbb{R}^{20000}, \;\; R_i \in \{0, 1\}, \quad Y_i \in \{0, 1\}.$$

- We observe X_i.
- We generate $R_i \sim$ Bernoulli$(\pi(X_i))$ where $\pi : \mathbb{R}^{20000} \to [0, 1]$ is a known function.
- If $R_i = 1$ we observe Y_i. If $R_i = 0$ we do not observe Y_i.

Goal: estimate $\theta = \mathbb{P}(Y_i = 1)$.

When $R_i = 0$ then we have missing data since the Y_i's are not observed. This problem is a simplification of a real situation that occurs in some randomized clinical trials.

Bayes versus Frequentist

Frequentist Analysis. Let

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i R_i}{\pi(X_i)}.$$

Then

$$\mathbb{E}(\widehat{\theta}) = \theta$$

and

$$\sqrt{n}(\widehat{\theta} - \theta) \rightsquigarrow N(0, \tau^2).$$

Use Hoeffding's inequality to get a finite sample 95 percent confidence interval. The length of the interval is $O(n^{-1/2})$. No assumptions at all about the 20,000 dimensional regression function $m(x)$:

$$m(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x).$$

Bayes versus Frequentist

Bayesian Analysis. To make the problem simpler, let us assume that $f(x)$ is known. Note that

$$\theta = \mathbb{P}(Y = 1) = \int m(x) f(x) dx.$$

The likelihood is

$$\begin{aligned} L(m) &= \prod_{i=1}^{n} f(X_i, R_i, Y_i) \\ &= \prod_{i=1}^{n} f(X_i) f(R_i|X_i) f(Y_i|X_i)^{R_i} \\ &\propto \prod_{i} f(Y_i|X_i)^{R_i} \\ &\propto \prod_{i=1}^{n} \Big[m(X_i)^{Y_i} (1 - m(X_i))^{1-Y_i} \Big]^{R_i} \end{aligned}$$

where $m \in \mathcal{M}$ all 20,000 dimensional functions. The likelihood has no information.

Bayes versus Frequentist

The Bayes (or likelihood) estimator is not consistent.

The likelihood is useless in some high-dimensional problems. The Bayesian analysis ignores the randomization probabilities $\pi(X_i)$ since they drop out of the likelihood. But the frequentist estimator is explicitly a function of the $\pi(X_i)$'s.

Also,

$$\inf_P \mathbb{P}_P(\theta \in C_n) \geq 1 - \alpha$$

but if $\mathbb{P}(\theta \in B|\text{Data}) = 1 - \alpha$ then

$$\inf_P \mathbb{P}_P(\theta \in B) \approx 0.$$

Robustness and Influence Functions

Modern robust statistical theory was developed by Huber, Hampel, Tukey and others in the 1960's and 1970's. It seems relevant for privacy theory and was used explicity in Dwork and Lei (2009).

Intuitively, an estimator is robust if making a small change in the data does not affect the estimator too much. (Similar to differential privacy.)

Let $\theta = T(P)$. The influence function:

$$\psi_P(x) = \lim_{\epsilon \longrightarrow 0} \frac{T((1-\epsilon)P + \epsilon\delta_x) - T(P)}{\epsilon}$$

where δ_x is a point mass at x. Want a bounded (or even redescending) influence function.

Robustness and Influence Functions

If $\theta = T(P) =$ the mean then $\psi_P(x) = x - \theta$ and

$$\sup_x |\psi_P(x)| = \infty.$$

If $\theta = T(P) =$ the median then

$$\psi_P(x) = \frac{\text{sign}(x - \theta)}{2f(\theta)}$$

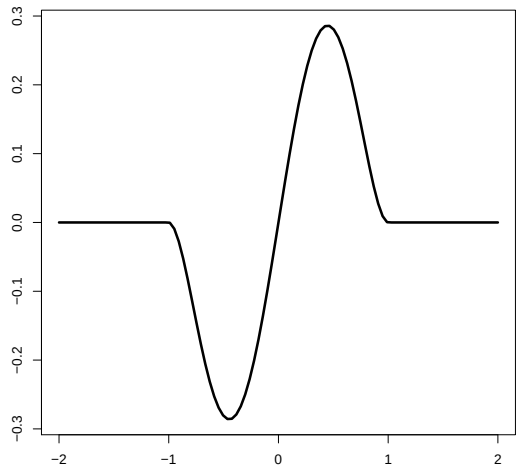
and

$$\sup_x |\psi_P(x)| = \frac{1}{2f(\theta)} < \infty$$

assuming $f > 0$.

Robustness and Influence Functions

Tukey's biweight estimator has a redescending influence function:



As far as I know, Dwork and Lei (2009) is the only paper linking robustness and privacy.

Statistical View of Differential Privacy

- Zhou and Wasserman, (JASA, 2010).

- Rinaldo, Wasserman and Zhou (in progress)

Differential Privacy

Database $x = (x_1, \ldots, x_n)$. Empirical distribution P^x.

Release $z = (z_1, \ldots, z_k)$. Empirical distribution P^z.

Mechanism: $M = \{Q_x : \ x \in \mathcal{X}\}$

$x \longrightarrow Q_x \longrightarrow Z$

Require that

$$Q_x(Z \in A) \leq e^{\alpha} Q_y(Z \in A) \quad \text{for all } A$$

whenever $x \sim y$ (x and y differ in one coordinate).

Differential Privacy

Conditional minimax risk:

$$\inf_{Q_x} \sup_{x \in D} \mathbb{E}_{Q_x} d(P^x, P^Z)$$

Marginal minimax risk when $X = (X_1, \ldots, X_n) \sim P$:

$$\inf_{\{Q_x\}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \mathbb{E}_{Q_x} d(P, P^Z)$$

Differential Privacy

Some distances:

Kolmogorov-Smirnov (KS) distance:

$$d(P,Q) = \sup_{t_1,\ldots,t_d} |P(X_1 \leq t_1, \ldots, X_d \leq t_d) - Q(X_1 \leq t_1, \ldots, X_d \leq t_d)|.$$

L_2 distance:

$$d(P,Q) = \int (\widetilde{p} - \widetilde{q})^2$$

where $\widetilde{p}$ is the density of a smoothed verion of P.

Wasserstein (Mallow, earth-mover) distance:

$$d(P,Q) = \inf_R \mathbb{E}_R ||X - Y||^2$$

where $X \sim P$, $Y \sim Q$ and the infimum is over all joint distributions R with marginals P and Q.

Differential Privacy

Exponential mechanism

Draw:

$$Z = (Z_1, \ldots, Z_k) \sim q(z|x) \propto e^{-c\alpha d(P^x, P^z)}$$

where c depends on d, n and k.

Differential Privacy

(Zhou and Wasserman, JASA, 2010).

$X = (X_1, \ldots, X_n) \sim P$

$X = x \longrightarrow Q_x \longrightarrow Z = (Z_1, \ldots, Z_k)$

Compare:

$d(P, P^x)$ to $d(P, P^z)$.

Distance	Data Release mechanism			minimax rate
	smoothed histogram	perturbed histogram	exp. mech.	
L_2	$n^{-2/(2r+3)}$	$n^{-2/(2+r)}$	NA	$n^{-2/(2+r)}$
KS	$n^{-2/(6+r)}$	$n^{-2/(2+r)}$	$n^{-1/3}$	$n^{-1/2}$

Are these rates optimal? Let's take a minimax view ...

Differential Privacy

Let $d(P^x, P^z) = ||P^x - P^z||_\infty$ be KS distance. For simplicity, assume that $X_i \in [0,1]$. Recall that the Conditional risk is

$$\sup_{x \in D} \mathbb{E}_{Q_x} d(P^x, F^Z)$$

Let $D = \{x : \ ||P^x - U||_\infty \leq \delta\}$ where U is the uniform. Then

$$\inf_{\text{mechanisms}} \sup_{x \in D} \mathbb{E}_{Q_x} d(P^x, F^Z) \geq \frac{\delta}{2}$$

and if $q(z|x) \propto \exp\{-n\alpha||P^x - P^z||_\infty\}$ then the bound is achieved.

Differential Privacy

Marginal risk: $X = (X_1, \ldots, X_n) \sim P$:

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \mathbb{E}_{Q_x} d(P, P^z)$$

where $\mathcal{P} = \{P : \ p \text{ is bounded}\}$ Then

$$\inf_{\text{mechanisms}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \mathbb{E}_{Q_x} d(P, F^Z) \geq \frac{C}{\sqrt{n}}$$

and again is achieved by the exponential mechanism.

This as the same rate as the non-privatized data. We are currently extending the results to other distances.

THE END

References

- *All of Statistics* (Wasserman)
- *All of Nonparametric Statistics* (Wasserman)
- *Statistical Inference* (Casella and Berger)
- *Asymptotic Statistics* (van der Vaart)
- *Introduction to Nonparametric Estimation* (Tsybakov)
- *Statistical Machine Learning* (Lafferty and Wasserman) (coming soon)
- *A Distribution-Free Theory of Nonparametric Regression* (Gyorfi, Kohler, Krzyzak, Walk)