Bootstrap in some non-standard problems[1]

Bodhisattva Sen
Department of Statistics
Columbia University
bodhi@stat.columbia.edu

14 September, 2009

[1] Joint work with Moulinath Banerjee, Michael Woodroofe, Ian McKeague and Emilio Seijo; supported by NSF grant DMS-0906597

Non-standard problems

- Estimators converge at *non-$\sqrt{n}$* rate and/or have *non–normal* limit distributions
- Examples: Estimation of monotone functions, change-point models...

Goal

- Inference in *non-standard* problems
- Want: *Confidence intervals* (CI) for the parameter
- Asymptotic distributions contain *nuisance* parameters (and sometimes unknown) that are difficult to estimate
- Investigate *bootstrap* based methods

Outline

Outline

Current status model

- $\{(X_i, T_i)\}_{i=1}^n$ i.i.d. $X_i \sim F$, $T_i \sim G$; $X_i \perp T_i$
- X_i: *Unobserved* time of *onset* of a disease
- T_i: *Check-up* time
- Data: $\{(\Delta_i, T_i)\}_{i=1}^n$ where $\Delta_i = \mathbf{1}\{X_i \leq T_i\}$
- Goal: Construct a CI for $F(t)$

- $\hat{F}_n = \arg\max_{F\uparrow} \sum_{i=1}^n \{\Delta_i \log F(T_i) + (1-\Delta_i)\log(1-F(T_i))\}$

- $n^{1/3}\{\hat{F}_n(t) - F(t)\} \xrightarrow{d} [4F(t)\{1-F(t)\}f(t)/g(t)]^{1/3}\mathbb{C}$
(Groeneboom & Wellner, 1992)

- Difficulty: Need to estimate the *nuisance parameters* f, g

Maximum score estimator

- Binary response model: $\{(Y_i, X_i)\}_{i=1}^n$ i.i.d.
- $Y_i = \mathbf{1}\{X_i'\beta_0 - \epsilon_i \geq 0\}$ where Median$(\epsilon_i|X_i) = 0$
- If $\epsilon_i \sim f(x) = e^x/(1+e^x) \Rightarrow$ Logistic regression

- $\hat{\beta}_n = \arg\max_{\|\beta\|=1} \sum_{i=1}^n (2Y_i - 1)\mathbf{1}\{X_i'\beta \geq 0\}$
- Goal: Construct a CI for β_0

- $n^{1/3}(\hat{\beta}_n - \beta_0) \xrightarrow{d} \gamma\mathbb{C}$ (Kim and Pollard, 1990)

- Difficulty: Need to estimate the *nuisance parameter* γ – depends on $F(\cdot|x)$ and its density and the distribution of X_i

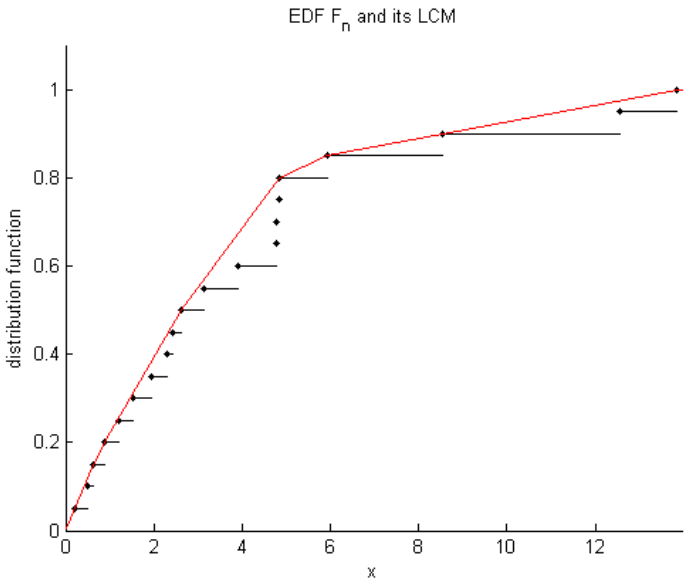
Outline

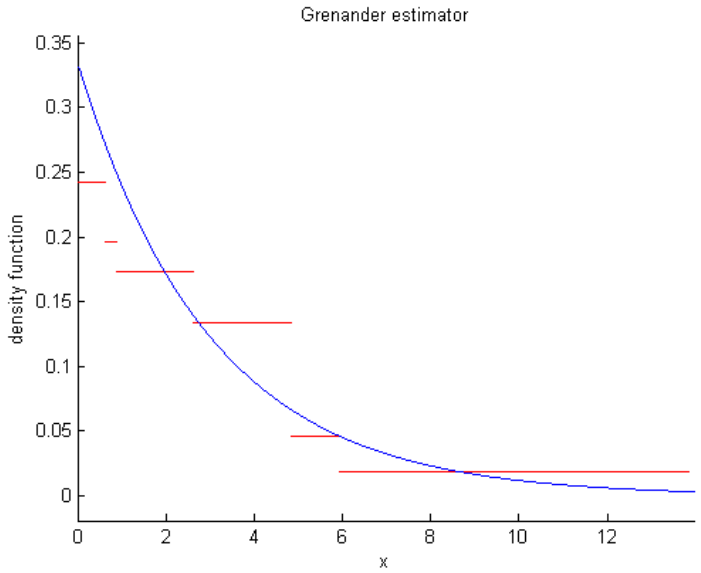
Monotone density estimation

- $X_1, X_2, \ldots, X_n$ i.i.d. $f \downarrow$ (unknown) on $[0, \infty)$
- Want to estimate f
- *Likelihood*: $L(f) = \prod_{i=1}^{n} f(X_i)$
- *Grenander* estimator $\tilde{f}_n$: *NPMLE* of f
- Grenander (1956, *Skand. Akt.*), Prakasa Rao (1969, *Sankhyā*)
- Demography, Astronomy, Renewal theory

The Grenander estimator

- $\tilde{f}_n$, the NPMLE of f, is the *left derivative* of $\tilde{F}_n$





$\mathbb{F}_n$ (e.d.f.) and its *Least Concave Majorant* (LCM), $\tilde{F}_n$ $\tilde{f}_n$ and f (true density)

Theorem (Prakasa Rao, 1969):

Let $t_0 \in (0, \infty)$ and $f'(t_0) \neq 0$ then

$$\Delta_n := n^{1/3} \left\{ \tilde{f}_n(t_0) - f(t_0) \right\} \xrightarrow{d} \kappa \mathbb{C}$$

where $\kappa = 2 \left| \frac{1}{2} f(t_0) f'(t_0) \right|^{1/3}$, $\mathbb{C}$ has *Chernoff's* distribution.

Some features

- $n^{1/3}$-rate of convergence
- *Non-normal* limit distribution
- *Nuisance* parameters
- *Non-standard* asymptotics

- Question: Can we *bootstrap* Δ_n *consistently*?

Bootstrap: An Introduction

- $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ i.i.d. F *unknown*
- Want to estimate H_n: *sampling distribution* of $R_n(\mathbf{X}_n, F)$, e.g., $R_n(\mathbf{X}_n, F) = \Delta_n$

Idea

1. Estimate F by $\hat{F}_n$
2. With $\hat{F}_n$ *fixed*, generate $\mathbf{X}_n^* = (X_1^*, X_2^*, \ldots, X_{m_n}^*)$, *bootstrap sample* of size m_n from $\hat{F}_n$
3. *Approximate* H_n by $\hat{H}_n$, the distribution of $R_{m_n}(\mathbf{X}_n^*, \hat{F}_n)$

Consistency of Bootstrap

- $\mathbf{X} = (X_1, X_2, \ldots)$ defined on $(\Omega, \mathcal{A}, P)$
- $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$
- $H_n(x) = P\{R_n(\mathbf{X}_n, F) \leq x\}$
- $\hat{H}_n(x) = P\{R_n(\mathbf{X}_n^*, \hat{F}_n) \leq x | \mathbf{X}_n\}$
- *Consistency:* $L(\hat{H}_n(\omega), H_n) \xrightarrow{P} 0$ where L is the Levy metric (or any metric metrizing convergence in distribution)
- If $H_n \xrightarrow{d} H$, for the bootstrap procedure to be consistent it is enough to show that $L(\hat{H}_n, H) \xrightarrow{P} 0$

Bootstrapping Δ_n

- Want to *approximate* the *sampling distribution* of $\Delta_n \sim H_n$
- *Bootstrap sample:* $X^*_{n,1}, X^*_{n,2}, \ldots, X^*_{n,n} \sim \hat{F}_n$, conditionally independent
- $\tilde{f}^*_n$: NPMLE of the bootstrap sample
- *Bootstrap statistic:* $\Delta^*_n = n^{1/3} \left\{ \tilde{f}^*_n(t_0) - \tilde{f}_n(t_0) \right\} \sim \hat{H}_n$
- Want to *investigate the limit behavior* of Δ^*_n
- Does $L(\hat{H}_n, \kappa\mathbb{C}) \xrightarrow{P} 0$?

Result (Sen, Banerjee & Woodroofe, 2009)

- Bootstrapping from $\mathbb{F}_n$ and $\tilde{F}_n$ (MLE) are *inconsistent*

Exponential(1) $|Z|, Z \sim \text{Normal}(0,1)$

n	EDF	NPMLE
50	0.747	0.720
100	0.776	0.755
200	0.802	0.780
500	0.832	0.797

n	EDF	NPMLE
50	0.761	0.739
100	0.778	0.757
200	0.780	0.762
500	0.788	0.755

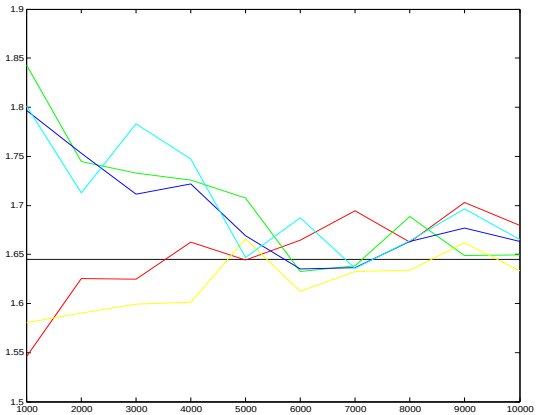
Table: Estimated *coverage probabilities* of nominal *95%* confidence intervals for $f(1)$

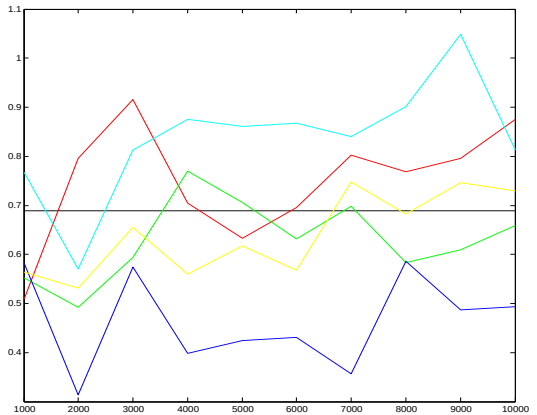
Result (cont...)

- In fact, we argue that Δ_n^* does not have *any weak limit*, in probability

Plot of *95 percentile* of the *bootstrap distribution*

Sample mean	Grenander estimator
$\sqrt{n}(\bar{X}_n - \mu)$	$n^{1/3}\{\tilde{f}_n(1) - f(1)\}$





Results(Cont...)

- Is there *any consistent* BS procedure?
- A version of *smoothed Bootstrap*, m out of n bootstrap from $\mathbb{F}_n$ and $\tilde{F}_n$ are *consistent*
- Derived *sufficient* conditions for *consistency*

Why is bootstrap inconsistent?

- One obvious problem with drawing the bootstrap samples from the e.d.f. $\mathbb{F}_n$ is that $\mathbb{F}_n$ *does not have a density*
- $\mathbb{F}_n$ or $\tilde{F}_n$ are not *smooth* enough

Outline

Wicksell's second problem

Problem

- $\mathbf{X} = (X_1, X_2, X_3) \sim \rho(z)$, $Z = X_1^2 + X_2^2 + X_3^2$
- Goal: Estimate F, the distribution function of Z
- Only observe their *projected positions*, i.e., (X_1, X_2)

- Wicksell's (1925, *Biometrika*) Corpuscle Problem
- *Inverse* problem (missing data)
- Groeneboom and Jongbloed (1995, AOS)

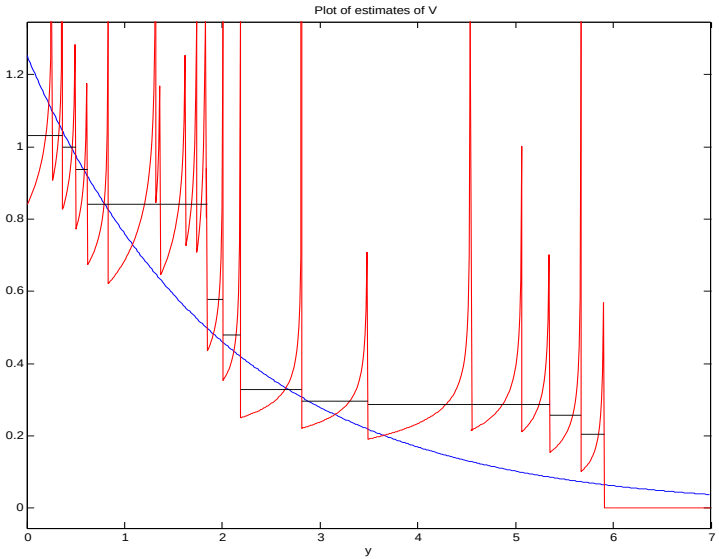
Description

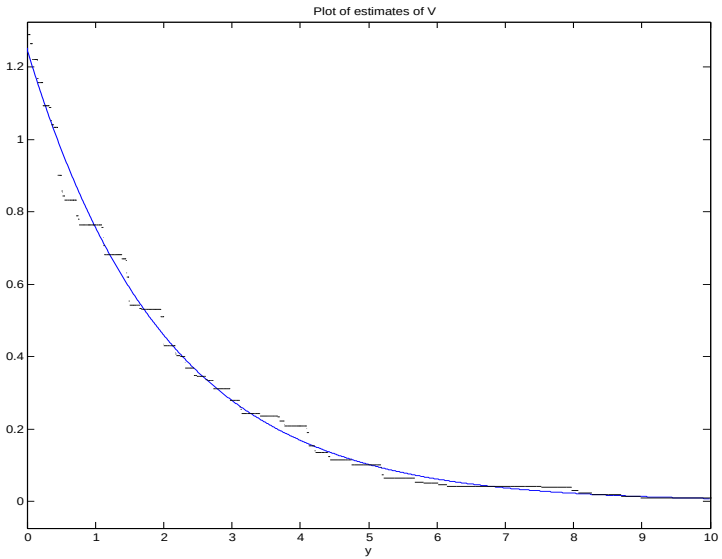
- If $Y := X_1^2 + X_2^2 \sim g$, $g(y) = \pi \int_y^\infty \frac{\rho(z)}{\sqrt{z-y}} dz$
- Define $V(y) := \int_y^\infty \frac{g(u)}{\sqrt{u-y}} du = \pi^2 \int_y^\infty \rho(z) dz$; V is $\downarrow$
- $F(z) = 1 + \frac{2}{\pi} \int_z^\infty \sqrt{y}\, dV(y)$

Estimates of V

- *Unbiased* estimate of V: $V_n(y) := \frac{1}{n} \sum_{i=1}^n \frac{\mathbf{1}\{Y_i > y\}}{\sqrt{Y_i - y}}$
- *Isotonic* estimate of V: $\tilde{V}_n$, closest non-increasing function to V_n (in L_2-sense)
- $\tilde{V}_n$: *slope of the LCM of* $U_n(x) = \int_0^x V_n(z)\, dz$

Plot of V, V_n and $\tilde{V}_n$ from synthetic data



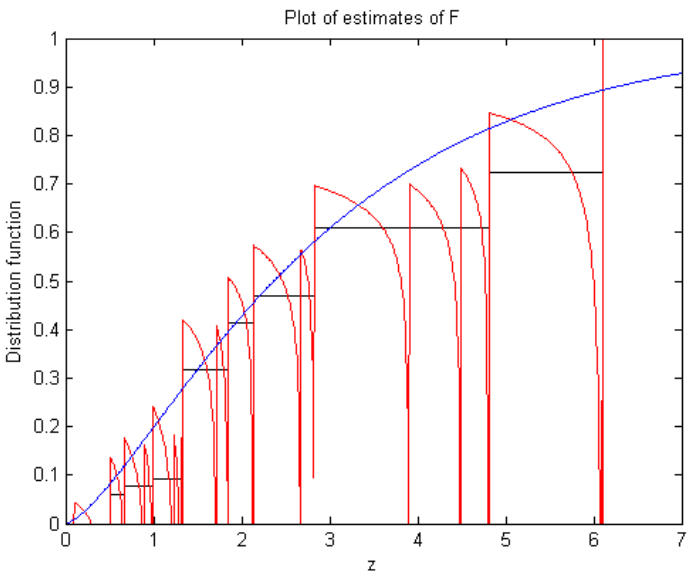


$n = 20$

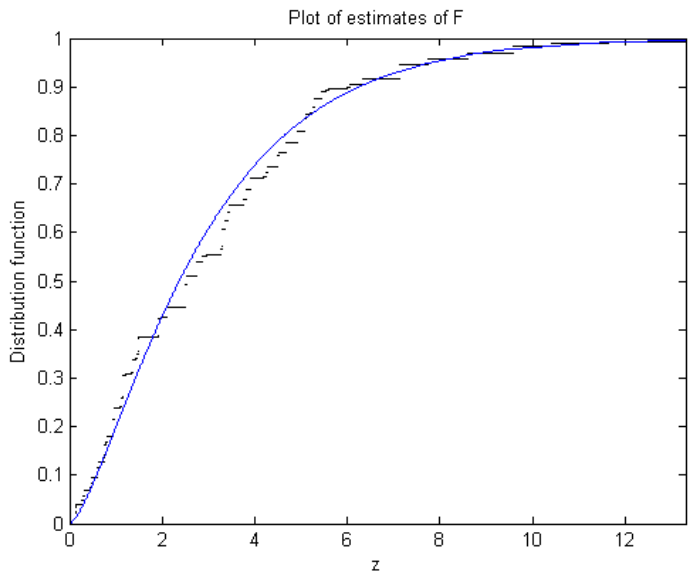
$n = 500$

Estimates of F

- $F_n(z) = 1 + \frac{2}{\pi} \int_z^\infty \sqrt{y}\, dV_n(y)$
- $\tilde{F}_n(z) = 1 + \frac{2}{\pi} \int_z^\infty \sqrt{y}\, d\tilde{V}_n(y)$



F, F_n and $\tilde{F}_n$ $(n = 20)$



F and $\tilde{F}_n$ $(n = 500)$

Theorem (Sen & Woodroofe, 2009)

Under certain assumptions, for $x_0 \in (0, \infty)$ and $\epsilon_n = \sqrt{\frac{\log n}{n}}$,

$$\epsilon_n^{-1} \{F_n(x_0) - F(x_0)\} \xrightarrow{d} N(0, \tfrac{4}{\pi^2} x_0 g(x_0))$$
$$\epsilon_n^{-1} \left\{\tilde{F}_n(x_0) - F(x_0)\right\} \xrightarrow{d} N(0, \tfrac{2}{\pi^2} x_0 g(x_0))$$

- ϵ_n^{-1}-rate of convergence to a *normal* limit
- Limiting *variance*

Bootstrap

- Bootstrapping (from the e.d.f.) $\tilde{V}_n$ and $\tilde{F}_n$ are *consistent*
- We derived sufficient conditions for any bootstrap method to be *consistent*

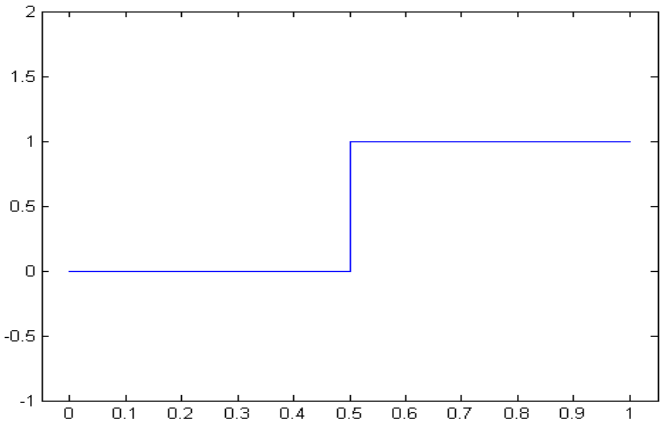
Outline

A change-point model

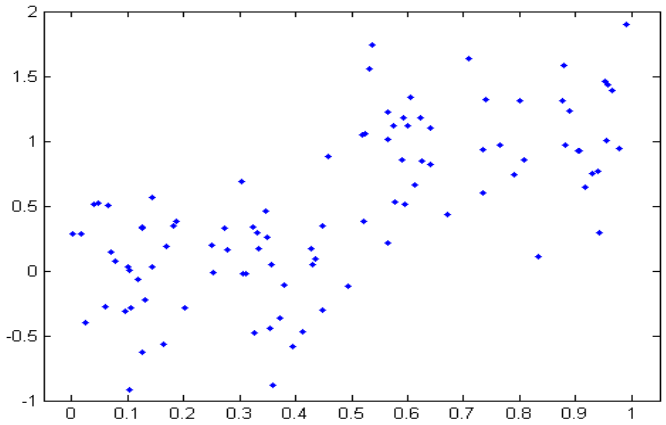
- $\{(Y_i, X_i)\}_{i=1}^n \in \mathbb{R}^2$ i.i.d. from the *regression* model

$$Y = \alpha_0 \mathbf{1}\{X \leq d_0\} + \beta_0 \mathbf{1}\{X > d_0\} + \epsilon$$

where $\epsilon \perp X$, $E(\epsilon) = 0$; d_0 is the *change point*



Regression function



Scatter Plot

A change-point model

- $(\hat{\alpha}_n, \hat{\beta}_n, \hat{d}_n) = \arg\min \sum_{i=1}^{n} [Y_i - \alpha \mathbf{1}\{X_i \leq d\} - \beta \mathbf{1}\{X_i > d\}]^2$
- $n(\hat{d}_n - d_0) \xrightarrow{d} \arg\min$ *Two sided Compound poisson process* that depends on
  - the *distribution* of ϵ!
  - the *density* of the covariate X
- How do we construct a CI for d_0?
- Does *bootstrap* work?

Results (Seijo and Sen, 2009)

- Usual *Nonparametric* bootstrap ("bootstrapping pairs") is *not consistent*
- Bootstrapping *residuals* keeping X_i's fixed *does not work*
- Need to *smooth* the distribution of X! Intuition?

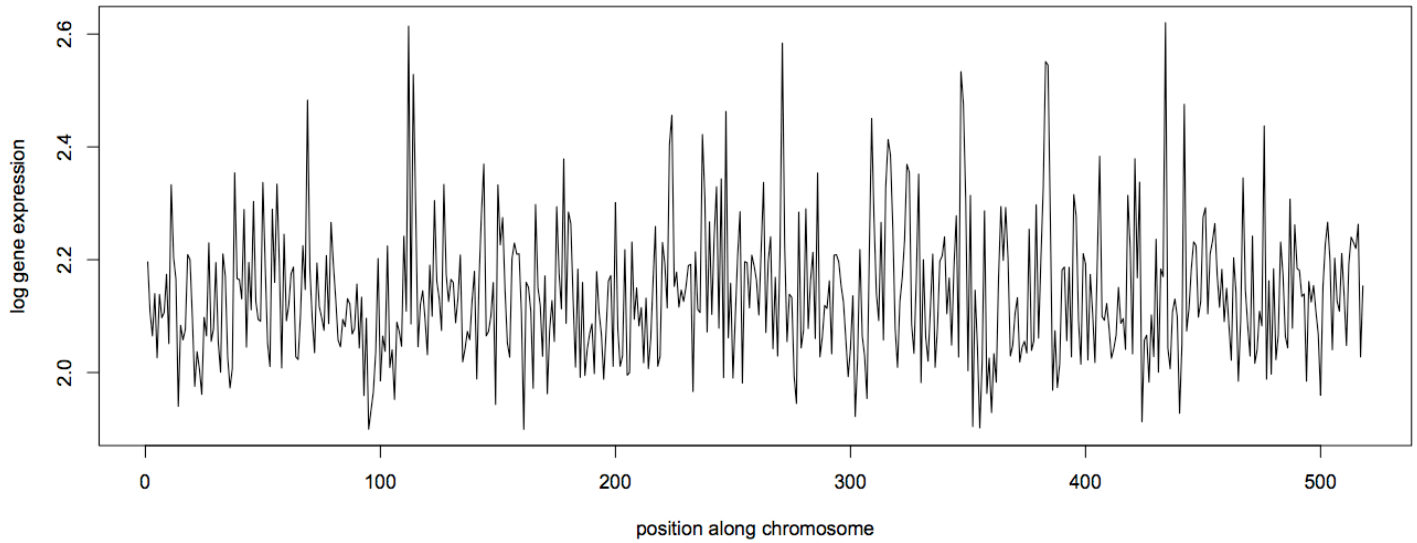
Consistent Bootstrap

- Bootstrap sample: $\{(Y_i^*, X_i^*)\}_{i=1}^n$
- Generate $X_i^* \sim \hat{f}_n$, where $\hat{f}_n$ is a *smooth density*
- Residual: $\hat{\epsilon}_i = Y_i - \hat{\alpha}_n \mathbf{1}\{X_i \leq \hat{d}_n\} - \hat{\beta}_n \mathbf{1}\{X_i > \hat{d}_n\}$
- Generate ϵ_i^* from the EDF of $\{\hat{\epsilon}_i - \bar{\epsilon}\}_{i=1}^n$
- $Y_i^* = \hat{\alpha}_n \mathbf{1}\{X_i^* \leq \hat{d}_n\} + \hat{\beta}_n \mathbf{1}\{X_i^* > \hat{d}_n\} + \epsilon_i^*$

Outline

Point impact functional linear model

- Genome-wide *expression studies*
- Goal: Locate *genes* associated with *clinical outcomes*, e.g., BMI, bio-markers, etc.
- Gene expression profile across a chromosome can be regarded as a *functional predictor*



- X can be modeled as a *fractional Brownian* motion (fBm) on $[0, 1]$ with *Hurst* index $0 < H < 1$

- *Model*: $Y = \alpha_0 + \beta_0 X(\theta_0) + \epsilon$, where $\epsilon \perp X$, $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$
- θ_0 is the *sensitive* time-point
- Data: $\{(Y_i, X_i)\}_{i=1}^n$ i.i.d. with X_i being a fBm
- $(\hat{\alpha}_n, \hat{\beta}_n, \hat{\theta}_n) = \arg\min_{(\alpha,\beta,\theta)} \sum_{i=1}^n [Y_i - \alpha - \beta X_i(\theta)]^2$
- Can we construct a CI for θ_0?

Result (McKeague & Sen, 2009)

If B_H is a fBm with Hurst index H, then
$n^{1/(2H)}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \arg\max_t\{2\sigma B_H(t) + |t|^{2H}\}$

- Can we construct a CI for θ_0 that *avoids* estimation of H?

- The *Non-parametric* bootstrap method is *inconsistent*
- Bootstrapping *residuals* is *consistent*

Procedure

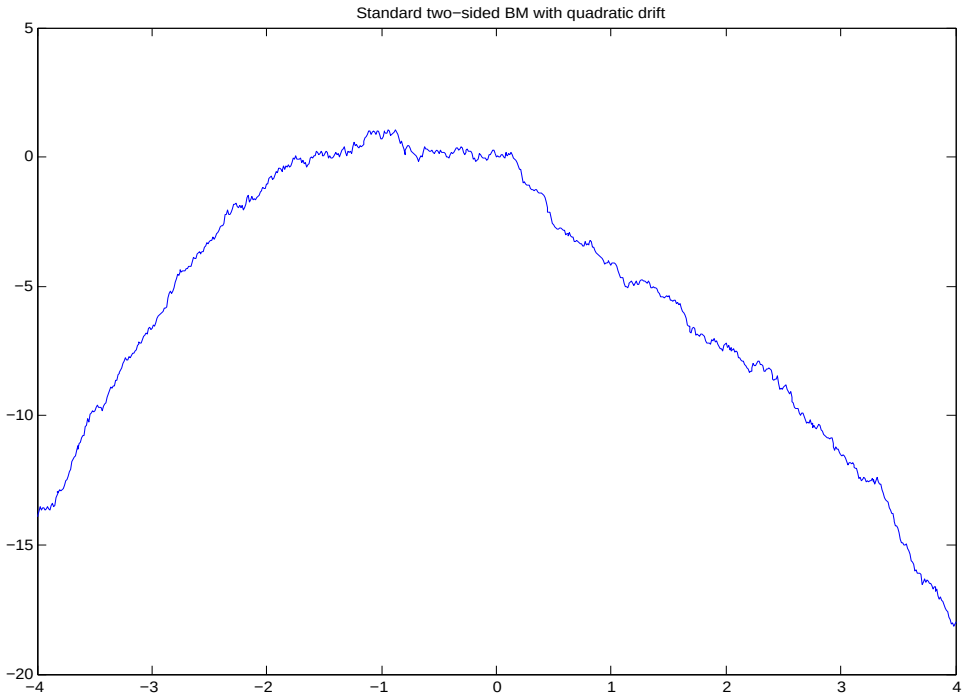
- Bootstrap sample $\{(Y_i^*, X_i)\}_{i=1}^n$
- Residual: $\hat{\epsilon}_i = Y_i - \hat{\alpha}_n - \hat{\beta}_n X_i(\hat{\theta}_n)$
- *Fix* X_i, generate ϵ_i^* from the EDF of $\{\hat{\epsilon}_i - \bar{\epsilon}\}_{i=1}^n$
- $Y_i^* = \hat{\alpha}_n + \hat{\beta}_n X_i(\hat{\theta}_n) + \epsilon_i^*$

Table: Monte Carlo results for coverage probabilities and average widths of nominal 95% confidence intervals for θ_0; data simulated from the linear model with $\theta_0 = 1/2$, $\alpha_0 = 0$ and $\beta_0 = 1$.

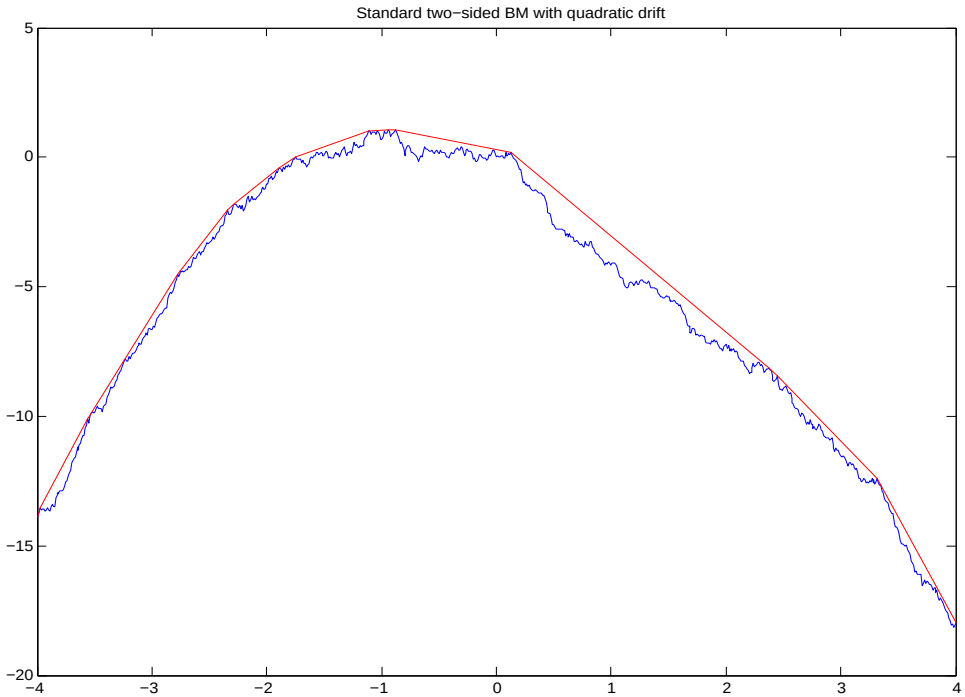
			Wald-H		R Bootstrap		NP Bootstrap	
n	σ	H	cover	width	cover	width	cover	width
20	0.3	0.3	0.874	0.023	0.924	0.044	1.000	0.174
		0.5	0.880	0.088	0.946	0.119	0.992	0.220
		0.7	0.822	0.170	0.912	0.249	0.970	0.360
	0.5	0.3	0.806	0.129	0.912	0.211	0.998	0.410
		0.5	0.852	0.256	0.924	0.333	0.988	0.487
		0.7	0.834	0.352	0.938	0.510	0.962	0.591
40	0.3	0.3	0.984	0.007	0.986	0.002	1.000	0.022
		0.5	0.892	0.048	0.942	0.053	0.992	0.087
		0.7	0.898	0.108	0.930	0.138	0.976	0.182
	0.5	0.3	0.900	0.039	0.928	0.054	0.998	0.149
		0.5	0.908	0.134	0.950	0.165	0.990	0.251
		0.7	0.856	0.229	0.946	0.332	0.962	0.386

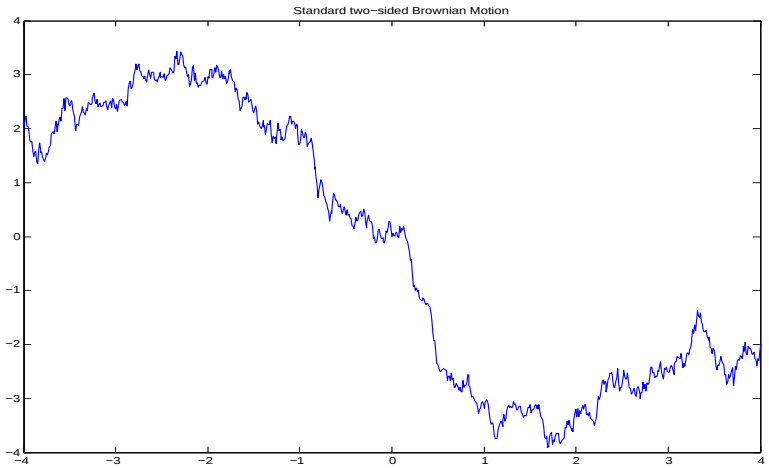
Outline

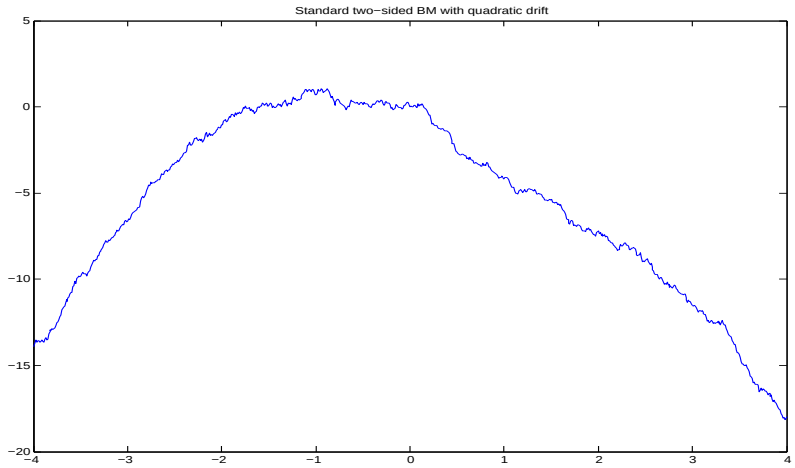
- *LCM* is the operator that maps a function $g : \mathbb{R} \to \mathbb{R}$ into the LCM of g



- *LCM* is the operator that maps a function $g : \mathbb{R} \to \mathbb{R}$ into the LCM of g







$\mathbb{W}$ is a *std. two-sided BM* $\qquad \mathbb{Z}(h) = \mathbb{W}(f(t_0)h) + \frac{1}{2}h^2 f'(t_0)$

- Recall: $\Delta_n = n^{1/3}\{\tilde{f}_n(t_0) - f(t_0)\} \xrightarrow{d} \kappa\mathbb{C}$
- $\kappa\mathbb{C} \stackrel{d}{=} LCM[\mathbb{Z}]'(0)$

Outline

- BS sample: $X^*_{n,1}, X^*_{n,2}, \ldots, X^*_{n,n} \sim \mathbb{F}_n$, conditionally independent; $\mathbb{F}^*_n$, $\tilde{f}^*_n$ defined previously
- The quantity of interest $\Delta^*_n := n^{1/3}\{\tilde{f}^*_n(t_0) - \tilde{f}_n(t_0)\}$
- We study the *limiting* behavior of the process
$$\mathbb{Z}^*_n(h) := n^{2/3}\left\{\mathbb{F}^*_n(t_0 + hn^{-1/3}) - \mathbb{F}^*_n(t_0) - \tilde{f}_n(t_0)hn^{-1/3}\right\}$$
- $\Delta^*_n = LCM[\mathbb{Z}^*_n]'(0)$; recall: $\Delta_n \xrightarrow{d} LCM[\mathbb{Z}]'(0)$
- Use *continuous mapping* arguments to deduce the limiting behavior of Δ^*_n

Details

- $\mathbb{Z}_n^* = \mathbb{Z}_{n,1}^* + \mathbb{Z}_{n,2}^*$ where
  $\mathbb{Z}_{n,1}^*(h) = n^{2/3}\left\{(\mathbb{F}_n^* - \mathbb{F}_n)(t_0 + hn^{-1/3}) - (\mathbb{F}_n^* - \mathbb{F}_n)(t_0)\right\}$
  $\mathbb{Z}_{n,2}^*(h) = n^{2/3}\left\{\mathbb{F}_n(t_0 + hn^{-1/3}) - \mathbb{F}_n(t_0) - \tilde{f}_n(t_0)hn^{-1/3}\right\}$
- $\mathbb{Z}_{n,1}^*(h) \xrightarrow{d} \mathbb{W}_1^*(f(t_0)h) =: \mathbb{Z}_1^*(h)$ *conditional* on data *a.s.*
- Note that conditionally $\mathbb{Z}_{n,2}^*$ is a *fixed function*

Result

- *Conditional* on **X**, the distribution of $\mathbb{Z}_n^*$ *does not* have any weak limit in probability
- Then, $\Delta_n^* = LCM[\mathbb{Z}_n^*]'(0)$ should *not* converge weakly to $LCM[\mathbb{Z}]'(0)$ (not so immediate!)

Take home points

- Usual *with replacement* bootstrap does not work in some non-standard problems
- Sometimes more explicit use of the underlying *model* can be make bootstrap inference *valid*
- In $n^{1/3}$-convergence problems *smoothing* is required; such examples are plenty, e.g., Current status model, monotone regression, Manski's maximum score estimator, etc.

References

- Sen, B., Banerjee, M. & Woodroofe, M. (2009). Inconsistency of Bootstrap: the Grenander estimator. (in revision with the *Ann. Statist.*)
- Sen, B. & Woodroofe, M. (2009). A Stereological problem (to be submitted)
- McKeague, I. & Sen, B. (2009). Trajectories with point impact in functional linear regression. (submitted)
- Seijo, E. & Sen, B. (2009). Bootstrap in change-point model (to be submitted)

Thank You!

Questions?