

# (In)-consistency of the Bootstrap in non-standard problems<sup>1</sup>

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

31 July, 2009

---

<sup>1</sup>Joint work with Moulinath Banerjee, Michael Woodroffe, 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
- Example: 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

- 1 **Bootstrap in  $n^{1/3}$ -rate problems**
  - Monotone density estimation
  
- 2 **Some other non-standard problems**
  - A change-point model
  - Point impact functional linear model

# Outline

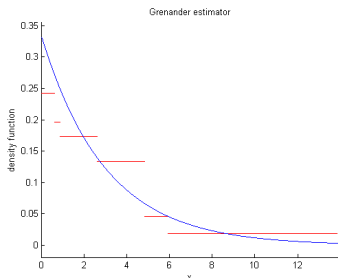
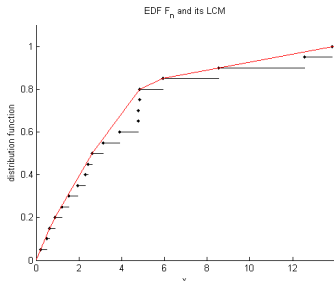
- 1 **Bootstrap in  $n^{1/3}$ -rate problems**
  - Monotone density estimation
- 2 **Some other non-standard problems**
  - A change-point model
  - Point impact functional linear model

# Monotone density estimation

- $X_1, X_2, \dots, X_n$  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, *Sankhāya*)
- Demography, Astronomy, Renewal theory

# The Grenander estimator

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



$F_n$  (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*  $\Delta_n$  consistently?

# Bootstrapping $\Delta_n$

- Want to *approximate* the *sampling distribution*  $H_n$  of  $\Delta_n$
- *Bootstrap sample*:  $X_{n,1}^*, X_{n,2}^*, \dots, X_{n,n}^* \sim \hat{F}_n$ , conditionally independent
- $\tilde{f}_n^*$ : NPMLE of the bootstrap sample
- *Bootstrap statistic*:  $\Delta_n^* = n^{1/3} \left\{ \tilde{f}_n^*(t_0) - \hat{f}_n(t_0) \right\} \sim \hat{H}_n$
- Idea: Approximate the  $H_n$  by  $\hat{H}_n$
- Does  $L(\hat{H}_n, H_n) \xrightarrow{P} 0$  (*weak consistency* of bootstrap)?

## Main results (Sen, Banerjee & Woodroffe, 2009)

- Bootstrapping from  $\mathbb{F}_n$  and  $\tilde{F}_n$  (MLE) are *inconsistent*
- In fact, we argue that  $\Delta_n^*$  does not have *any weak limit*, in probability
- Is there *any consistent* bootstrap 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

- 1 Bootstrap in  $n^{1/3}$ -rate problems
  - Monotone density estimation
- 2 Some other non-standard problems
  - A change-point model
  - Point impact functional linear model

# A change-point model

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

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

where  $\epsilon \perp X$ ,  $E(\epsilon) = 0$

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

## Preliminary findings

- 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

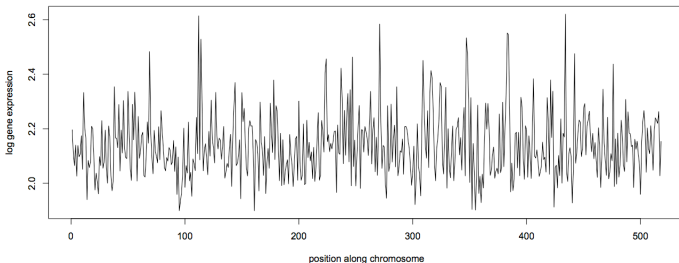
- Bootstrap sample:  $\{(Y_i^*, X_i^*)\}_{i=1}^n$
- Generate  $X_i^* \sim \hat{f}_n$ , where  $\hat{f}_n$  is a *smooth density*
- Residual:  $\hat{\epsilon}_i = Y_i - \hat{\alpha}_n \mathbf{1}\{X_i \leq \hat{d}_n\} - \hat{\beta}_n \mathbf{1}\{X_i > \hat{d}_n\}$
- Generate  $\epsilon_i^*$  from the EDF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

- 1 **Bootstrap in  $n^{1/3}$ -rate problems**
  - Monotone density estimation
  
- 2 **Some other non-standard problems**
  - A change-point model
  - Point impact functional linear model

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

## 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  $\theta_0$  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  $\epsilon_i^*$  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  $\theta_0$ ; data simulated from the linear model with  $\theta_0 = 1/2$ ,  $\alpha_0 = 0$  and  $\beta_0 = 1$ .

$n$	$\sigma$	$H$	Wald- $H$		R Bootstrap		NP Bootstrap	
			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

## 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

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

*Thank You!*  
*Questions?*