Bootstrap Consistency for General Semiparametric M-Estimation

Guang Cheng

Department of Statistics, Purdue University

Statistics Seminar at University of Michigan
March 12th, 2010
Joint Work with Jianhua Huang at Texas A& M University

Outline

- As a general data-resampling based method, the bootstrap has been applied to the semiparametric models in a wide variety of contexts, e.g. Biostatistics, Survival Analysis and Econometrics, for a long time.

- As a general data-resampling based method, the bootstrap has been applied to the semiparametric models in a wide variety of contexts, e.g. Biostatistics, Survival Analysis and Econometrics, for a long time.
- A long existing problem for applying bootstrap to semiparametric models is that there is no theoretical justifications !

- As a general data-resampling based method, the bootstrap has been applied to the semiparametric models in a wide variety of contexts, e.g. Biostatistics, Survival Analysis and Econometrics, for a long time.
- A long existing problem for applying bootstrap to semiparametric models is that there is no theoretical justifications !
- Cited from an Annals paper by Kosorok et al (2004) "*Unfortunately, there is no sufficiently general theory, as far as we are aware, available for the nonparametric bootstrap in the penalized (semiparametric) maximization likelihood setting.*"

- As a general data-resampling based method, the bootstrap has been applied to the semiparametric models in a wide variety of contexts, e.g. Biostatistics, Survival Analysis and Econometrics, for a long time.
- A long existing problem for applying bootstrap to semiparametric models is that there is no theoretical justifications !
- Cited from an Annals paper by Kosorok et al (2004) "*Unfortunately, there is no sufficiently general theory, as far as we are aware, available for the nonparametric bootstrap in the penalized (semiparametric) maximization likelihood setting.*"
- Cited from a JASA paper by Lee et al (2005) "*The bootstrap is a possible solution to these problems, but theoretical justification is not available for semiparametric models.*"

Major Contributions

- We establish the theoretical validity of a broad class of bootstrap methods as an inferential tool for the general semiparametric models.

Major Contributions

- We establish the theoretical validity of a broad class of bootstrap methods as an inferential tool for the general semiparametric models.
- Specifically, we prove that
  - the bootstrap distribution asymptotically imitates the distribution of the M-estimate, i.e., bootstrap distributional consistency;

Major Contributions

- We establish the theoretical validity of a broad class of bootstrap methods as an inferential tool for the general semiparametric models.
- Specifically, we prove that
  - the bootstrap distribution asymptotically imitates the distribution of the M-estimate, i.e., bootstrap distributional consistency;
  - the bootstrap confidence set has the asymptotically correct coverage probability, i.e., bootstrap confidence set consistency.

Major Contributions

- We establish the theoretical validity of a broad class of bootstrap methods as an inferential tool for the general semiparametric models.
- Specifically, we prove that
  - the bootstrap distribution asymptotically imitates the distribution of the M-estimate, i.e., bootstrap distributional consistency;
  - the bootstrap confidence set has the asymptotically correct coverage probability, i.e., bootstrap confidence set consistency.

  for the *nonparametric bootstrap, bayesian bootstrap.....*

Major Contributions

- We establish the theoretical validity of a broad class of bootstrap methods as an inferential tool for the general semiparametric models.
- Specifically, we prove that
  - the bootstrap distribution asymptotically imitates the distribution of the M-estimate, i.e., bootstrap distributional consistency;
  - the bootstrap confidence set has the asymptotically correct coverage probability, i.e., bootstrap confidence set consistency.

  for the *nonparametric bootstrap, bayesian bootstrap.....*
- Most (*but not all*) details of this talk can be found in a forthcoming Annals paper.

Semiparametric Models

- Random Variable $X \sim \{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in \mathcal{H}\}$

Semiparametric Models

- Random Variable $X \sim \{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in \mathcal{H}\}$
  - θ: Euclidean parameter of interest;

Semiparametric Models

- Random Variable $X \sim \{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in \mathcal{H}\}$
  - θ: Euclidean parameter of interest;
  - η: a *possibly* infinite dimensional *nuisance* parameter, e.g. some function.

Semiparametric Models

- Random Variable $X \sim \{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in \mathcal{H}\}$
  - θ: Euclidean parameter of interest;
  - η: a *possibly* infinite dimensional *nuisance* parameter, e.g. some function.
- Example I: The Cox regression model with survival data

Semiparametric Models

- Random Variable $X \sim \{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in \mathcal{H}\}$
  - θ: Euclidean parameter of interest;
  - η: a *possibly* infinite dimensional *nuisance* parameter, e.g. some function.
- Example I: The Cox regression model with survival data
  - θ: regression parameter;
  - η: cumulative hazard function.

Semiparametric Models

- Random Variable $X \sim \{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in \mathcal{H}\}$
  - θ: Euclidean parameter of interest;
  - η: a *possibly* infinite dimensional *nuisance* parameter, e.g. some function.
- Example I: The Cox regression model with survival data
  - θ: regression parameter;
  - η: cumulative hazard function.
- Example II: The partly linear model: $Y = W'\theta + f(T) + \epsilon$

Semiparametric Models

- Random Variable $X \sim \{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in \mathcal{H}\}$
  - θ: Euclidean parameter of interest;
  - η: a *possibly* infinite dimensional *nuisance* parameter, e.g. some function.
- Example I: The Cox regression model with survival data
  - θ: regression parameter;
  - η: cumulative hazard function.
- Example II: The partly linear model: $Y = W'\theta + f(T) + \epsilon$
  - θ: linear regression parameter;
  - f: nonlinear smooth function.

Semiparametric Models

- Random Variable $X \sim \{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in \mathcal{H}\}$
  - θ: Euclidean parameter of interest;
  - η: a *possibly* infinite dimensional *nuisance* parameter, e.g. some function.
- Example I: The Cox regression model with survival data
  - θ: regression parameter;
  - η: cumulative hazard function.
- Example II: The partly linear model: $Y = W'\theta + f(T) + \epsilon$
  - θ: linear regression parameter;
  - f: nonlinear smooth function.
- Example III: Generalized Estimating Equation (GEE)

Semiparametric Models

- Random Variable $X \sim \{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in \mathcal{H}\}$
  - θ: Euclidean parameter of interest;
  - η: a *possibly* infinite dimensional *nuisance* parameter, e.g. some function.
- Example I: The Cox regression model with survival data
  - θ: regression parameter;
  - η: cumulative hazard function.
- Example II: The partly linear model: $Y = W'\theta + f(T) + \epsilon$
  - θ: linear regression parameter;
  - f: nonlinear smooth function.
- Example III: Generalized Estimating Equation (GEE)
  - β: mean regression parameter;
  - α: finite dimensional nuisance correlation parameter.

- The semiparametric MLE $(\widehat{\theta}, \widehat{\eta})$ is defined as

$$(\widehat{\theta}, \widehat{\eta}) = \arg \sup_{\theta \in \Theta, \eta \in \mathcal{H}} \sum_{i=1}^{n} \log lik(\theta, \eta)(X_i).$$

- The semiparametric MLE $(\widehat{\theta}, \widehat{\eta})$ is defined as

$$(\widehat{\theta}, \widehat{\eta}) = \arg \sup_{\theta \in \Theta, \eta \in \mathcal{H}} \sum_{i=1}^{n} \log lik(\theta, \eta)(X_i).$$

- Given regularity conditions, it is proven that MLE $\widehat{\theta}$ is semiparametric efficient (minimal asymptotic variance) in the sense that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \tilde{I}_0^{-1}), \tag{1}$$

where $\tilde{I}_0$ is the *efficient information matrix*.

The asymptotic inferences for θ based on the above asymptotic distribution theory face the following practical challenges:

The asymptotic inferences for θ based on the above asymptotic distribution theory face the following practical challenges:

1. The confidence set construction and asymptotic covariance estimation for θ both involve estimating and inverting a hard-to-estimate infinite-dimensional operator.

The asymptotic inferences for θ based on the above asymptotic distribution theory face the following practical challenges:

1. The confidence set construction and asymptotic covariance estimation for θ both involve estimating and inverting a hard-to-estimate infinite-dimensional operator.
2. Convergence rates for θ and η may be different, i.e., θ has a parametric rate and η has a nonparametric rate. Thus, we can not treat (θ, η) as a whole component.

The asymptotic inferences for θ based on the above asymptotic distribution theory face the following practical challenges:

1. The confidence set construction and asymptotic covariance estimation for θ both involve estimating and inverting a hard-to-estimate infinite-dimensional operator.

2. Convergence rates for θ and η may be different, i.e., θ has a parametric rate and η has a nonparametric rate. Thus, we can not treat (θ, η) as a whole component.

3. See Bickel et al (1998) and Kosorok (2008) for more details.

Three semiparametric inferential tools are thus motivated:

Three semiparametric inferential tools are thus motivated:

1. The Profile Likelihood Approach

Three semiparametric inferential tools are thus motivated:

1. The Profile Likelihood Approach

   - Murphy and van der Vaart (2000) proved that the log-profile likelihood defined as

$$\log pl_n(\theta) = \sup_{\eta \in \mathcal{H}} \sum_{i=1}^{n} \log lik(\theta, \eta)(X_i)$$

     asymptotically possesses the parabolic form, like a parametric likelihood, when making inferences for θ.

Three semiparametric inferential tools are thus motivated:

1. The Profile Likelihood Approach

- Murphy and van der Vaart (2000) proved that the log-profile likelihood defined as

$$\log pl_n(\theta) = \sup_{\eta \in \mathcal{H}} \sum_{i=1}^{n} \log lik(\theta, \eta)(X_i)$$

asymptotically possesses the parabolic form, like a parametric likelihood, when making inferences for θ.

- For example, we can use the second order numerical derivative (curvature) of the $\log pl_n(\theta)$ at $\widehat{\theta}$, called *observed information*, to estimate $-\tilde{I}_0$.

2. The Profile Sampler

2. The Profile Sampler

- The inferences of θ are based on the sampling from the posterior of the profile likelihood.

2. The Profile Sampler

- The inferences of θ are based on the sampling from the posterior of the profile likelihood.
- MCMC is used for sampling from the above posterior distribution. For example, we can use the chain mean (the inverse of chain variance) to approximate MLE $\widehat{\theta}$ ($\tilde{I}_0$).

2. The Profile Sampler

- The inferences of θ are based on the sampling from the posterior of the profile likelihood.
- MCMC is used for sampling from the above posterior distribution. For example, we can use the chain mean (the inverse of chain variance) to approximate MLE $\widehat{\theta}$ ($\tilde{I}_0$).
- The theoretical validity of the profile sampler depends on the parabolic form of the profile likelihood proven in Murphy and van der Vaart (2000).

However,

- The profile likelihood may not approximate well the desired parabolic form when the sample size is small;

However,

- The profile likelihood may not approximate well the desired parabolic form when the sample size is small;
- The profile likelihood based procedures are not automatic:

However,

- The profile likelihood may not approximate well the desired parabolic form when the sample size is small;
- The profile likelihood based procedures are not automatic:
  - we need to specify a prior in the profile sampler;

However,

- The profile likelihood may not approximate well the desired parabolic form when the sample size is small;
- The profile likelihood based procedures are not automatic:
  - we need to specify a prior in the profile sampler;
  - we need to choose the step size in computing the observed information.

However,

- The profile likelihood may not approximate well the desired parabolic form when the sample size is small;
- The profile likelihood based procedures are not automatic:
  - we need to specify a prior in the profile sampler;
  - we need to choose the step size in computing the observed information.
- MCMC convergence in the profile sampler is not well studied.

However,

- The profile likelihood may not approximate well the desired parabolic form when the sample size is small;
- The profile likelihood based procedures are not automatic:
  - we need to specify a prior in the profile sampler;
  - we need to choose the step size in computing the observed information.
- MCMC convergence in the profile sampler is not well studied.
- It takes long time to run Markov chain to get accurate inferences for θ when η has slow convergence rate (Cheng and Kosorok, 2008a, b);

3. The last approach is the Bootstrap Sampling
   The bootstrap method has already been used before the above approaches were invented.

3. The last approach is the Bootstrap Sampling
   The bootstrap method has already been used before the above approaches were invented.

   Methodological Features:

3. The last approach is the Bootstrap Sampling
   The bootstrap method has already been used before the above approaches were invented.

   Methodological Features:

   A. A lot of well studied resampling techniques are available;

3. The last approach is the Bootstrap Sampling
   The bootstrap method has already been used before the above approaches were invented.

   Methodological Features:

   A. A lot of well studied resampling techniques are available;
   B. It is conceptually simple to implement, just resampling;

3. The last approach is the Bootstrap Sampling
   The bootstrap method has already been used before the above approaches were invented.

   Methodological Features:

   A. A lot of well studied resampling techniques are available;
   B. It is conceptually simple to implement, just resampling;
   C. we can make bootstrap inferences for both θ and η based on the bootstrap sample;

3. The last approach is the Bootstrap Sampling
   The bootstrap method has already been used before the above approaches were invented.

   Methodological Features:

   A. A lot of well studied resampling techniques are available;
   B. It is conceptually simple to implement, just resampling;
   C. we can make bootstrap inferences for both θ and η based on the bootstrap sample;
   D. It is an automatic procedure, e.g., needless to specify prior;

3. The last approach is the Bootstrap Sampling
   The bootstrap method has already been used before the above approaches were invented.

   Methodological Features:

   A. A lot of well studied resampling techniques are available;
   B. It is conceptually simple to implement, just resampling;
   C. we can make bootstrap inferences for both θ and η based on the bootstrap sample;
   D. It is an automatic procedure, e.g., needless to specify prior;
   E. It has small sample advantages.

- However, there is no theoretical justifications available now.

In this talk, we consider the general class of *exchangeably weighted bootstrap*. This general resampling scheme was first proposed by Rubin (1981), i.e., Bayesian Bootstrap, and then extensively studied by Newton and Mason (1992), Praestgaard and Wellner (1993) and Bertail and Barbe (1995) in the parametric and nonparametric context.

The class of *exchangeably weighted bootstrap* includes

- *Standard Nonparametric Bootstrap:*
  Sample with replacement (sampling from the empirical distribution)

The class of *exchangeably weighted bootstrap* includes

- *Standard Nonparametric Bootstrap:*
  Sample with replacement (sampling from the empirical distribution)
- *Multiplier Bootstrap: (Generalization of Bayesian Bootstrap)*
  A smooth alternative to the nonparametric bootstrap (useful for bootstrapping censored data).

The class of *exchangeably weighted bootstrap* includes

- *Standard Nonparametric Bootstrap:*
  Sample with replacement (sampling from the empirical distribution)
- *Multiplier Bootstrap: (Generalization of Bayesian Bootstrap)*
  A smooth alternative to the nonparametric bootstrap (useful for bootstrapping censored data).
- *Double Bootstrap:*
  Sampling from the nonparametric bootstrap sample.

The class of *exchangeably weighted bootstrap* includes

- *Standard Nonparametric Bootstrap:*
  Sample with replacement (sampling from the empirical distribution)
- *Multiplier Bootstrap: (Generalization of Bayesian Bootstrap)*
  A smooth alternative to the nonparametric bootstrap (useful for bootstrapping censored data).
- *Double Bootstrap:*
  Sampling from the nonparametric bootstrap sample.
- Others: Multivariate Hypergeometric Bootstrap, Polya-Eggenberger Bootstrap, Bootstrap generated from deterministic weights......

- The bootstrap estimator is defined as

$$(\widehat{\theta}^*, \hat{\eta}^*) = \arg \sup_{\theta\in\Theta, \eta\in\mathcal{H}} \sum_{i=1}^{n} \log lik(\theta, \eta)(X_i^*), \qquad (2)$$

where $(X_1^*, \ldots, X_n^*)$ is the bootstrap sample.

- The bootstrap estimator is defined as

$$(\widehat{\theta}^*, \hat{\eta}^*) = \arg \sup_{\theta \in \Theta, \eta \in \mathcal{H}} \sum_{i=1}^{n} \log lik(\theta, \eta)(X_i^*), \tag{2}$$

where $(X_1^*, \ldots, X_n^*)$ is the bootstrap sample.

- In the exchangeably weighted bootstrap sampling, $(\widehat{\theta}^*, \widehat{\eta}^*)$ can be reformulated as

$$(\widehat{\theta}^*, \hat{\eta}^*) = \arg \sup_{\theta \in \Theta, \eta \in \mathcal{H}} \sum_{i=1}^{n} W_{ni} \log lik(\theta, \eta)(X_i), \tag{3}$$

where $W_n \equiv (W_{n1}, \ldots, W_{nn})$ is the bootstrap weight.

Different bootstrap sampling schemes correspond to different bootstrap weights W_n, e.g.,

Different bootstrap sampling schemes correspond to different bootstrap weights W_n, e.g.,

- Standard Nonparametric Bootstrap: $W_n \sim Multinomial\,(n, (n^{-1}, \ldots, n^{-1}))$.

Different bootstrap sampling schemes correspond to different bootstrap weights W_n, e.g.,

- Standard Nonparametric Bootstrap:
  $W_n \sim \mathit{Multinomial}\,(n, (n^{-1}, \ldots, n^{-1}))$.
- Multiplier Bootstrap:
  $W_{nj} = \alpha_j/\bar{\alpha}$, *where* $(\alpha_1, \ldots, \alpha_n) \overset{iid}{\sim}$ *nonnegative continuous r.v.s.* For example, $\alpha_j \sim \mathrm{Exp}(1)$ in Bayesian Bootstrap.

Different bootstrap sampling schemes correspond to different bootstrap weights W_n, e.g.,

- Standard Nonparametric Bootstrap:
  $W_n \sim Multinomial\,(n, (n^{-1}, \ldots, n^{-1}))$.

- Multiplier Bootstrap:
  $W_{nj} = \alpha_j/\bar{\alpha}$, *where* $(\alpha_1, \ldots, \alpha_n) \overset{iid}{\sim}$ *nonnegative continuous r.v.s.* For example, $\alpha_j \sim \mathrm{Exp}(1)$ in Bayesian Bootstrap.

- Double Bootstrap:
  $W_n \sim Multinomial(n, (M_{n1}/n, \ldots, M_{nn}/n))$
  $(M_{n1}, \ldots, M_{nn}) \sim Multinomial(n, (n^{-1}, \ldots, n^{-1}))$.

Bootstrap Consistency Assumptions

- Regularity conditions (those in guaranteeing the asymptotic normality of MLE $\widehat{\theta}$).

Bootstrap Consistency Assumptions

- Regularity conditions (those in guaranteeing the asymptotic normality of MLE $\widehat{\theta}$).
- Bootstrap weights satisfy:
  1. *The vector* $W_n = (W_{n1}, \ldots, W_{nn})'$ *is exchangeable*;
  2. $W_{ni} \geq 0$ *and* $\sum_{i=1}^{n} W_{ni} = n$;
  3. $(1/n)\sum_{i=1}^{n}(W_{ni} - 1)^2 \xrightarrow{P_W} c^2 > 0$;
  4. *Some lower moment conditions for* W_{ni}.

Bootstrap Consistency Assumptions

- Regularity conditions (those in guaranteeing the asymptotic normality of MLE $\widehat{\theta}$).
- Bootstrap weights satisfy:
  1. *The vector* $W_n = (W_{n1}, \ldots, W_{nn})'$ *is exchangeable*;
  2. $W_{ni} \geq 0$ *and* $\sum_{i=1}^{n} W_{ni} = n$;
  3. $(1/n)\sum_{i=1}^{n}(W_{ni} - 1)^2 \xrightarrow{P_W} c^2 > 0$;
  4. *Some lower moment conditions for* W_{ni}.
- Some measurability conditions (so that the Fubini's Theorem can be used freely), envelop conditions and empirical processes conditions (the entropy number of some function class is manageable).

Bootstrap Consistency Theorem (Theorem I):

Assuming the above conditions, we have

1. *$\sqrt{n}$ consistency of $\widehat{\theta}^*$:*

$$\|\widehat{\theta}^* - \theta_0\| = O_{P_W}(n^{-1/2}) \quad \text{in } P_X - \text{probability}.$$

Bootstrap Consistency Theorem (Theorem I):

Assuming the above conditions, we have

1. *$\sqrt{n}$ consistency of $\widehat{\theta}^*$:*

$$\|\widehat{\theta}^* - \theta_0\| = O_{P_W}(n^{-1/2}) \quad \text{in } P_X - \text{probability.}$$

2. *Asymptotic Normality for $\widehat{\theta}^*$:*

$$(\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta}) \Longrightarrow N(0, \tilde{I}_0^{-1}) \quad \text{in } P_X - \text{probability,}$$

where $\Longrightarrow$ denotes *conditional weak convergence*.

Bootstrap Consistency Theorem (Theorem I):

Assuming the above conditions, we have

1. *$\sqrt{n}$ consistency of $\widehat{\theta}^*$:*

$$\|\widehat{\theta}^* - \theta_0\| = O_{P_W}(n^{-1/2}) \quad \text{in } P_X - \text{probability.}$$

2. *Asymptotic Normality for $\widehat{\theta}^*$:*

$$(\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta}) \Longrightarrow N(0, \tilde{I}_0^{-1}) \quad \text{in } P_X - \text{probability,}$$

where $\Longrightarrow$ denotes *conditional weak convergence*.

3. *Bootstrap distribution consistency for θ*

$$\sup_{x \in R^d} \left| P_{W|\mathcal{X}_n}((\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta}) \leq x) - P_X(\sqrt{n}(\widehat{\theta} - \theta_0) \leq x) \right| \xrightarrow{P_X} 0.$$

Remark 1:

- $c = 1$ for the *nonparametric bootstrap* in which $W_n \sim \text{Multinomial}\,(n, (n^{-1}, \ldots, n^{-1}))$.

Remark 1:

- $c = 1$ for the *nonparametric bootstrap* in which $W_n \sim \text{Multinomial}\,(n, (n^{-1}, \ldots, n^{-1}))$.
- $c = \sigma(\alpha_1)/E(\alpha_1)$ for the *multiplier bootstrap* in which $W_{nj} = \alpha_j/\bar{\alpha}$. [$c = 1$ for the bayesian bootstrap]

Remark 1:

- $c = 1$ for the *nonparametric bootstrap* in which $W_n \sim \text{Multinomial}\,(n, (n^{-1}, \ldots, n^{-1}))$.
- $c = \sigma(\alpha_1)/E(\alpha_1)$ for the *multiplier bootstrap* in which $W_{nj} = \alpha_j/\bar{\alpha}$. [$c = 1$ for the bayesian bootstrap]
- $c = \sqrt{2}$ for the *double bootstrap* in which $W_n \sim \text{Multinomial}(n, (M_{n1}/n, \ldots, M_{nn}/n))$ $(M_{n1}, \ldots, M_{nn}) \sim \text{Multinomial}(n, (n^{-1}, \ldots, n^{-1}))$

Remark 2:

The bootstrap consistency assumptions are only slightly stronger than those (regularity conditions) needed in showing the asymptotic normality of $\widehat{\theta}$. Therefore, we can conclude that the bootstrap consistency for $\widehat{\theta}^*$ is almost automatically guaranteed.

Definition for *conditional weak convergence* $\Longrightarrow$ is as follows

$$\widehat{X}_n^* \Longrightarrow X \text{ conditional on data } \mathcal{X}_n$$

$$\text{if} \sup_{h \in BL_1(B)} |E_{\cdot|\mathcal{X}_n} h(\widehat{X}_n^*) - Eh(X)| = o_{P_X}(1),$$

where $BL_1(B)$ is a collection of all Lipschitz continuous functions $h : B \mapsto R$ bounded by 1 and having Lipschitz constant 1 (Hoffmann-Jorgensen, 1984).

Technical tools used in proving Bootstrap Consistency Theorem include

Definition for *conditional weak convergence* $\Longrightarrow$ is as follows

$$\widehat{X}_n^* \Longrightarrow X \text{ conditional on data } \mathcal{X}_n$$

$$\text{if} \sup_{h \in BL_1(B)} |E_{\cdot|\mathcal{X}_n} h(\widehat{X}_n^*) - Eh(X)| = o_{P_X}(1),$$

where $BL_1(B)$ is a collection of all Lipschitz continuous functions $h : B \mapsto R$ bounded by 1 and having Lipschitz constant 1 (Hoffmann-Jorgensen, 1984).

Technical tools used in proving Bootstrap Consistency Theorem include

- Bootstrapped Empirical Processes (Gine and Zinn, 1990);

Definition for *conditional weak convergence* $\Longrightarrow$ is as follows

$$\widehat{X}_n^* \Longrightarrow X \text{ conditional on data } \mathcal{X}_n$$

$$\text{if } \sup_{h\in BL_1(B)} |E_{\cdot|\mathcal{X}_n} h(\widehat{X}_n^*) - Eh(X)| = o_{P_X}(1),$$

where $BL_1(B)$ is a collection of all Lipschitz continuous functions $h : B \mapsto R$ bounded by 1 and having Lipschitz constant 1 (Hoffmann-Jorgensen, 1984).

Technical tools used in proving Bootstrap Consistency Theorem include

- Bootstrapped Empirical Processes (Gine and Zinn, 1990);
- Multiplier Inequality (Wellner and Zhan, 1996);

Definition for *conditional weak convergence* $\Longrightarrow$ is as follows

$$\widehat{X}_n^* \Longrightarrow X \text{ conditional on data } \mathcal{X}_n$$

$$\text{if} \sup_{h \in BL_1(B)} |E_{\cdot|\mathcal{X}_n} h(\widehat{X}_n^*) - Eh(X)| = o_{P_X}(1),$$

where $BL_1(B)$ is a collection of all Lipschitz continuous functions $h : B \mapsto R$ bounded by 1 and having Lipschitz constant 1 (Hoffmann-Jorgensen, 1984).

Technical tools used in proving Bootstrap Consistency Theorem include

- Bootstrapped Empirical Processes (Gine and Zinn, 1990);
- Multiplier Inequality (Wellner and Zhan, 1996);
- Hoffmann-Jorgensen Inequality (van der Vaart and Wellner, 1996).

Bootstrap Confidence Set

- Advantages: Replace the tedious theoretical derivations in semiparametric inferences with routine simulations of bootstrap samples.

Bootstrap Confidence Set

- Advantages: Replace the tedious theoretical derivations in semiparametric inferences with routine simulations of bootstrap samples.
- The bootstrap distributional consistency Theorem implies the consistency of bootstrap confidence sets of the following types:
  - percentile type
  - hybrid type
  - t type

- Let $P_{W|\mathcal{X}_n}$ denote the conditional distribution given observations $\mathcal{X}_n$. We define $\tau^*_{n\alpha}$ satisfying

$$P_{W|\mathcal{X}_n}\left(\widehat{\theta}^* \leq \tau^*_{n\alpha}\right) = \alpha.$$

- Let $P_{W|\mathcal{X}_n}$ denote the conditional distribution given observations $\mathcal{X}_n$. We define $\tau^*_{n\alpha}$ satisfying

$$P_{W|\mathcal{X}_n}\left(\widehat{\theta}^* \leq \tau^*_{n\alpha}\right) = \alpha.$$

- Based on the bootstrap distributional consistency Theorem, we can approximate the α-th quantile of the distribution of $(\widehat{\theta} - \theta_0)$ by $(\tau^*_{n\alpha} - \widehat{\theta})/c$. Thus we construct the $(1-\alpha)$ *percentile*-type bootstrap confidence set as

$$BC_p(\alpha) = \left[\widehat{\theta} + \frac{\tau^*_{n(\alpha/2)} - \widehat{\theta}}{c}, \widehat{\theta} + \frac{\tau^*_{n(1-\alpha/2)} - \widehat{\theta}}{c}\right].$$

- We define $\kappa^*_{n\alpha}$ satisfying

$$P_{W|\mathcal{X}_n}\left((\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta}) \leq \kappa^*_{n\alpha}\right) = \alpha.$$

- We define $\kappa^*_{n\alpha}$ satisfying

$$P_{W|\mathcal{X}_n}\left((\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta}) \leq \kappa^*_{n\alpha}\right) = \alpha.$$

- Similarly, we can approximate the α-th quantile of $\sqrt{n}(\widehat{\theta} - \theta_0)$ by $\kappa^*_{n\alpha}$. Thus we construct the $(1-\alpha)$ *hybrid*-type bootstrap confidence set as

$$BC_h(\alpha) = \left[\widehat{\theta} - \frac{\kappa^*_{n(1-\alpha/2)}}{\sqrt{n}}, \widehat{\theta} - \frac{\kappa^*_{n(\alpha/2)}}{\sqrt{n}}\right].$$

Remark 3:

- Both the *percentile* bootstrap confidence set $BC_p(\alpha)$ and *hybrid* bootstrap confidence set $BC_h(\alpha)$ can be computed easily through routine bootstrap sampling.

Remark 3:

- Both the *percentile* bootstrap confidence set $BC_p(\alpha)$ and *hybrid* bootstrap confidence set $BC_h(\alpha)$ can be computed easily through routine bootstrap sampling.
- We can avoid estimating the asymptotic variance of $\widehat{\theta}$, i.e., $\widetilde{I}_0^{-1}$, in both $BC_p(\alpha)$ and $BC_h(\alpha)$.

The distributional consistency Theorem together with the quantile convergence Theorem implies the consistency of percentile-type and hybrid-type bootstrap confidence sets.

Bootstrap Confidence Set Corollary (Corollary I): Under the conditions in Theorem 1, we have

$$\begin{aligned} P_{XW}(\theta_0 \in BC_p(\alpha)) &\longrightarrow 1-\alpha, \\ P_{XW}(\theta_0 \in BC_h(\alpha)) &\longrightarrow 1-\alpha, \end{aligned}$$

as $n \to \infty$.

Remark 4:
Provided the consistent estimator for the asymptotic covariance is available, we can show that the t-type bootstrap confidence set is also consistent by considering the Slutsky's Theorem.

Before presenting more general results, let us summarize our contributions in the likelihood framework so far:

Before presenting more general results, let us summarize our contributions in the likelihood framework so far:

- We first show that the bootstrap distribution of $(\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta})$, conditional on the observed data, asymptotically imitates the unconditional distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ in Theorem 1.

Before presenting more general results, let us summarize our contributions in the likelihood framework so far:

- We first show that the bootstrap distribution of $(\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta})$, conditional on the observed data, asymptotically imitates the unconditional distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ in Theorem 1.
- We next establish in Corollary 1 that the coverage probabilities of the percentile and hybrid bootstrap confidence sets for θ converge to the nominal level as a consequence of Theorem 1.

Before presenting more general results, let us summarize our contributions in the likelihood framework so far:

- We first show that the bootstrap distribution of $(\sqrt{n}/c)(\widehat{\theta}^* - \widehat{\theta})$, conditional on the observed data, asymptotically imitates the unconditional distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ in Theorem 1.
- We next establish in Corollary 1 that the coverage probabilities of the percentile and hybrid bootstrap confidence sets for θ converge to the nominal level as a consequence of Theorem 1.
- *Note that the above conclusions hold no matter η has $\sqrt{n}$ convergence rate or slower than $\sqrt{n}$ convergence rate.*

Our bootstrap consistency conclusions in the likelihood setup can be extended to the more general scenario as follows.

Our bootstrap consistency conclusions in the likelihood setup can be extended to the more general scenario as follows.

$G1.$ (Penalized) M-estimation:

$$(\widehat{\theta}_{\lambda_n}, \widehat{\eta}_{\lambda_n}) = \arg \max_{\theta\in\Theta,\eta\in\mathcal{H}} \sum_{i=1}^{n} m_{\lambda_n}(\theta,\eta)(X_i).$$

Our bootstrap consistency conclusions in the likelihood setup can be extended to the more general scenario as follows.

*G*1. (Penalized) M-estimation:

$$(\widehat{\theta}_{\lambda_n}, \widehat{\eta}_{\lambda_n}) = \arg \max_{\theta \in \Theta, \eta \in \mathcal{H}} \sum_{i=1}^{n} m_{\lambda_n}(\theta, \eta)(X_i).$$

Example II. $Y = W'\theta + f(T) + \epsilon$.

If f is smooth, we can estimate (θ, f) by penalizing its roughness $J(f)$ [partial smoothing spline]:

$$(\widehat{\theta}_{\lambda_n}, \widehat{\eta}_{\lambda_n}) = \arg \min_{\theta \in \Theta, \eta \in \mathcal{H}} \left\{ \sum_{i=1}^{n} (Y_i - W_i'\theta - f(T_i))^2 + \lambda_n J(f) \right\}.$$

*G*2. Estimation Equation:

$$\widehat{\theta} \text{ solves } \sum_{i=1}^{n} z(\theta, \widehat{\eta}(\theta))(X_i) = 0,$$

where $\widehat{\eta}(\theta)$ is an estimate for η given any fixed θ (satisfying some convergence rate condition).

$G2.$ Estimation Equation:

$$\widehat{\theta} \text{ solves } \sum_{i=1}^{n} z(\theta, \widehat{\eta}(\theta))(X_i) = 0,$$

where $\widehat{\eta}(\theta)$ is an estimate for η given any fixed θ (satisfying some convergence rate condition).

$G3.$ The nuisance parameter η can be finite dimensional (conditions can be relaxed in this special case).

$G2$. Estimation Equation:

$$\widehat{\theta} \text{ solves } \sum_{i=1}^{n} z(\theta, \widehat{\eta}(\theta))(X_i) = 0,$$

where $\widehat{\eta}(\theta)$ is an estimate for η given any fixed θ (satisfying some convergence rate condition).

$G3$. The nuisance parameter η can be finite dimensional (conditions can be relaxed in this special case).

Example III (GEE). This is a natural example of G2 & G3 in which θ is the mean regression parameter of interest and η is the finite dimensional *nuisance* correlation parameter.

- In the first Cox regression model, $(\widehat{\theta}, \widehat{\eta}) = \arg\max_{\theta\in\Theta,\eta\in\mathcal{H}} \sum_{i=1}^{n} \log lik(\theta, \eta)(X_i)$.

- In the first Cox regression model, $(\widehat{\theta}, \widehat{\eta}) = \arg\max_{\theta\in\Theta,\eta\in\mathcal{H}} \sum_{i=1}^{n} \log lik(\theta,\eta)(X_i)$.
- In the second partial smoothing spline model, $(\widehat{\theta}_{\lambda_n}, \widehat{\eta}_{\lambda_n}) = \arg\max_{\theta\in\Theta,\eta\in\mathcal{H}} \sum_{i=1}^{n} m_{\lambda_n}(\theta,\eta)(X_i)$.

- In the first Cox regression model, $(\widehat{\theta}, \widehat{\eta}) = \arg\max_{\theta \in \Theta, \eta \in \mathcal{H}} \sum_{i=1}^{n} \log lik(\theta, \eta)(X_i)$.
- In the second partial smoothing spline model, $(\widehat{\theta}_{\lambda_n}, \widehat{\eta}_{\lambda_n}) = \arg\max_{\theta \in \Theta, \eta \in \mathcal{H}} \sum_{i=1}^{n} m_{\lambda_n}(\theta, \eta)(X_i)$.
- In the last GEE model, $\widehat{\theta}$ solves $\sum_{i=1}^{n} z(\theta, \widehat{\eta}(\theta))(X_i) = 0$.

Model I. Cox Regression Model with Current Status Data

The hazard function of the survival time T of a subject with covariate Z is modelled as:

$$\lambda(t|z) \equiv \lim_{\Delta\to 0} \frac{1}{\Delta} Pr(t \leq T < t + \Delta | T \geq t, Z = z) = \lambda(t) \exp(\theta' z),$$

where λ is an unspecified baseline hazard function.

Model I. Cox Regression Model with Current Status Data

The hazard function of the survival time T of a subject with covariate Z is modelled as:

$$\lambda(t|z) \equiv \lim_{\Delta \to 0} \frac{1}{\Delta} Pr(t \leq T < t + \Delta | T \geq t, Z = z) = \lambda(t) \exp(\theta' z),$$

where λ is an unspecified baseline hazard function.

- In this model, we are usually interested in the regression parameter θ while treating the cumulative hazard function $\eta(y) = \int_0^y \lambda(t) dt$ as the nuisance parameter.

Model I. Cox Regression Model with Current Status Data

The hazard function of the survival time T of a subject with covariate Z is modelled as:

$$\lambda(t|z) \equiv \lim_{\Delta \to 0} \frac{1}{\Delta} Pr(t \leq T < t + \Delta | T \geq t, Z = z) = \lambda(t) \exp(\theta' z),$$

where λ is an unspecified baseline hazard function.

- In this model, we are usually interested in the regression parameter θ while treating the cumulative hazard function $\eta(y) = \int_0^y \lambda(t)dt$ as the nuisance parameter.
- The nonparametric bootstrap is applied to this model, e.g., Efron and Tibshirani (1986), but without any theoretical justifications.

- Consider the current status data in which the event time T is unobservable but we know whether the event has occurred at the examination time C or not. Thus, we observe $X = (C, \delta, Z)$, where $\delta = I\{T \leq C\}$.

- Consider the current status data in which the event time T is unobservable but we know whether the event has occurred at the examination time C or not. Thus, we observe $X = (C, \delta, Z)$, where $\delta = I\{T \leq C\}$.
- Note that, in the case of current status data, the nonparametric MLE $\widehat{\eta}$ is $n^{1/3}$ convergent.

- Consider the current status data in which the event time T is unobservable but we know whether the event has occurred at the examination time C or not. Thus, we observe $X = (C, \delta, Z)$, where $\delta = I\{T \leq C\}$.
- Note that, in the case of current status data, the nonparametric MLE $\widehat{\eta}$ is $n^{1/3}$ convergent.
- Our theory implies that
  Bootstrapping the regression parameter θ in the *Cox regression model with current status data* based on the exchangeable weights, e.g., *nonparametric bootstrap & bayesian bootstrap*, is consistent.

Model II. Partly Linear Model

We assume that

$$Y = \theta' W + f(Z) + \epsilon,$$

where ϵ is independent of (W, Z) and f is an unknown smooth function belonging to second order Sobolev space. The distribution of ϵ is assumed to satisfy some orthogonality condition.

Model II. Partly Linear Model

We assume that

$$Y = \theta' W + f(Z) + \epsilon,$$

where ϵ is independent of (W, Z) and f is an unknown smooth function belonging to second order Sobolev space. The distribution of ϵ is assumed to satisfy some orthogonality condition.

- Estimate (θ, f) using the penalized least square criterion:

$$\sum_{i=1}^{n}(y_i - \theta' w_i - f(z_i))^2 + \lambda_n \int_0^1 [f^{(2)}(s)]^2 ds. \tag{4}$$

Model II. Partly Linear Model

We assume that

$$Y = \theta' W + f(Z) + \epsilon,$$

where ϵ is independent of (W, Z) and f is an unknown smooth function belonging to second order Sobolev space. The distribution of ϵ is assumed to satisfy some orthogonality condition.

- Estimate (θ, f) using the penalized least square criterion:

$$\sum_{i=1}^{n}(y_i - \theta' w_i - f(z_i))^2 + \lambda_n \int_0^1 [f^{(2)}(s)]^2 ds. \qquad (4)$$

- Our theory implies that
  Bootstrapping the *partial smoothing spline estimator* $\widehat{\theta}_{\lambda_n}$ is consistent.

Model III. Generalized Estimating Equation for Longitudinal Data

In longitudinal data sets, we observe m_i repeated measurements for the i-th subject, i.e.

$$\text{Outcome } \; Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{im_i} \end{pmatrix} \; \text{ and } m_i \times p \; \text{ Covariate } \; X_i = \begin{pmatrix} X'_{i1} \\ \vdots \\ X'_{im_i}, \end{pmatrix}$$

for $i = 1, \ldots, n$. Each subject/cluster is assumed to be independent. However, the repeated measurements within each subject/cluster are assumed to be correlated.

- Model assumptions for the longitudinal data:

- Model assumptions for the longitudinal data:

  M1. $E(Y_i|X_i) = \mu_i(\beta_0)$;

- Model assumptions for the longitudinal data:

  *M*1. $E(Y_i|X_i) = \mu_i(\beta_0)$;
  *M*2. $Var(Y_i|X_i) = \Sigma_i(\beta_0)$.

- Model assumptions for the longitudinal data:
  M1. $E(Y_i|X_i) = \mu_i(\beta_0)$;
  M2. $Var(Y_i|X_i) = \Sigma_i(\beta_0)$.

- Due to the difficulty in specifying $\Sigma_i(\beta_0)$, Liang and Zeger (1986) introduced an additional correlation parameter α to form the so called "working covariance matrix" $V_i(\alpha, \beta)$ in order to approximate $\Sigma_i(\beta)$, and thus estimate β by solving

$$\sum_{i=1}^{n} D_i'(\beta) V_i^{-1}(\widehat{\alpha}(\beta), \beta) S_i(\beta) = 0, \tag{5}$$

where $D_i(\beta) = \partial\mu_i(\beta)/\partial\beta$, $S_i(\beta) = Y_i - \mu_i(\beta)$.

- Model assumptions for the longitudinal data:
  M1. $E(Y_i|X_i) = \mu_i(\beta_0)$;
  M2. $Var(Y_i|X_i) = \Sigma_i(\beta_0)$.

- Due to the difficulty in specifying $\Sigma_i(\beta_0)$, Liang and Zeger (1986) introduced an additional correlation parameter α to form the so called "working covariance matrix" $V_i(\alpha, \beta)$ in order to approximate $\Sigma_i(\beta)$, and thus estimate β by solving

$$\sum_{i=1}^{n} D_i'(\beta) V_i^{-1}(\widehat{\alpha}(\beta), \beta) S_i(\beta) = 0, \tag{5}$$

  where $D_i(\beta) = \partial\mu_i(\beta)/\partial\beta$, $S_i(\beta) = Y_i - \mu_i(\beta)$.

- $\widehat{\alpha}(\beta)$ may be obtained by GEE2 (Zhao and Prentice, 1990).

- In Liang and Zeger (1986), the solution of GEE $\widehat{\beta}$ is shown to be asymptotically normal even when $V_i(\alpha, \beta) \neq \Sigma_i(\beta)$ for all α (allowing covariance matrix misspecifications).

- In Liang and Zeger (1986), the solution of GEE $\widehat{\beta}$ is shown to be asymptotically normal even when $V_i(\alpha, \beta) \neq \Sigma_i(\beta)$ for all α (allowing covariance matrix misspecifications).
- In this scenario, it is natural to view the correlation parameter α as the nuisance parameter. Note that, in this particular example, α can be estimated at the parametric rate.

- In Liang and Zeger (1986), the solution of GEE $\widehat{\beta}$ is shown to be asymptotically normal even when $V_i(\alpha, \beta) \neq \Sigma_i(\beta)$ for all α (allowing covariance matrix misspecifications).
- In this scenario, it is natural to view the correlation parameter α as the nuisance parameter. Note that, in this particular example, α can be estimated at the parametric rate.
- Our general theory implies that
  In the framework of GEE, *cluster* bootstrap estimate of the mean regression parameter β [Sherman and le Cessie (1997)] based on the exchangeable weights is consistent.

Thank you for your attention....

Assistant Professor Guang Cheng
Department of Statistics, Purdue University
chengg@purdue.edu