How many iterations are sufficient for semiparametric estimation?

Guang Cheng

Department of Statistics, Purdue University

Statistics Department Colloquium
Northwestern University
March 9th, 2011

Outline

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: an Euclidean parameter of interest;

Semiparametric Models

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

Semiparametric Models

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

Semiparametric Models

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

Semiparametric Models

- Random Variable $X \sim \left\{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in \mathcal{H}\right\}$
  - θ: an Euclidean parameter of interest;
  - η: a *possibly* infinite dimensional nuisance parameter.
- Example I: The Cox regression model with survival data
  - θ: regression covariate;
  - η: cumulative hazard function.
- Example II: The conditional normal model
  - Conditional distribution $Y|(W = w, Z = z) \sim N(\theta' w, \eta(z))$

Semiparametric Models

- Random Variable $X \sim \left\{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in \mathcal{H}\right\}$
  - θ: an Euclidean parameter of interest;
  - η: a *possibly* infinite dimensional nuisance parameter.
- Example I: The Cox regression model with survival data
  - θ: regression covariate;
  - η: cumulative hazard function.
- Example II: The conditional normal model
  - Conditional distribution $Y|(W = w, Z = z) \sim N(\theta' w, \eta(z))$
- Example III: The partly linear model

Semiparametric Models

- Random Variable $X \sim \left\{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in \mathcal{H}\right\}$
  - θ: an Euclidean parameter of interest;
  - η: a *possibly* infinite dimensional nuisance parameter.
- Example I: The Cox regression model with survival data
  - θ: regression covariate;
  - η: cumulative hazard function.
- Example II: The conditional normal model
  - Conditional distribution $Y|(W = w, Z = z) \sim N(\theta' w, \eta(z))$
- Example III: The partly linear model
  - $Y = W'\theta + \eta(Z) + \epsilon$.

- Under regularity conditions, the semiparametric MLE $\widehat{\theta}$ is shown to be efficient in the sense that it achieves the minimal asymptotic variance over all regular estimates.

- Under regularity conditions, the semiparametric MLE $\widehat{\theta}$ is shown to be efficient in the sense that it achieves the minimal asymptotic variance over all regular estimates.
- A common practice to obtain the MLE $\widehat{\theta}$ is through maximizing its log-profile likelihood

$$\log pl_n(\theta) = \sup_{\eta \in \mathcal{H}} \log lik_n(\theta, \eta)$$

via some optimization algorithm.

- Under regularity conditions, the semiparametric MLE $\widehat{\theta}$ is shown to be efficient in the sense that it achieves the minimal asymptotic variance over all regular estimates.
- A common practice to obtain the MLE $\widehat{\theta}$ is through maximizing its log-profile likelihood

$$\log pl_n(\theta) = \sup_{\eta \in \mathcal{H}} \log lik_n(\theta, \eta)$$

via some optimization algorithm.

- For example, the Newton-Raphson (NR) algorithm is applied to the partial likelihood of the Cox model to obtain $\widehat{\theta}$ in **R**.

A general iterative procedure is implemented as follows

A general iterative procedure is implemented as follows

(I) Identify an initial estimate $\widehat{\theta}^{(0)}$;

A general iterative procedure is implemented as follows

(I) Identify an initial estimate $\widehat{\theta}^{(0)}$;

(II) Construct the corresponding nuisance estimate $\widehat{\eta}(\widehat{\theta}^{(0)})$ either by pure nonparametric approach, e.g., isotonic estimation, or under some regularization, e.g., kernel or sieve estimation;

A general iterative procedure is implemented as follows

(I) Identify an initial estimate $\widehat{\theta}^{(0)}$;

(II) Construct the corresponding nuisance estimate $\widehat{\eta}(\widehat{\theta}^{(0)})$ either by pure nonparametric approach, e.g., isotonic estimation, or under some regularization, e.g., kernel or sieve estimation;

(III) Apply NR (or other optimization) algorithm to the generalized profile likelihood (Severini and Wong, 1992)

$$\widehat{S}(\theta) \equiv \log lik_n(\theta, \widehat{\eta}(\theta))$$

at $\theta = \widehat{\theta}^{(0)}$ to obtain $\widehat{\theta}^{(1)}$;

A general iterative procedure is implemented as follows

(I) Identify an initial estimate $\widehat{\theta}^{(0)}$;

(II) Construct the corresponding nuisance estimate $\widehat{\eta}(\widehat{\theta}^{(0)})$ either by pure nonparametric approach, e.g., isotonic estimation, or under some regularization, e.g., kernel or sieve estimation;

(III) Apply NR (or other optimization) algorithm to the generalized profile likelihood (Severini and Wong, 1992)

$$\widehat{S}(\theta) \equiv \log lik_n(\theta, \widehat{\eta}(\theta))$$

at $\theta = \widehat{\theta}^{(0)}$ to obtain $\widehat{\theta}^{(1)}$;

(IV) Repeat k^* iterations until $|\widehat{S}(\widehat{\theta}^{(k^*)}) - \widehat{S}(\widehat{\theta}^{(k^*-1)})| \leq \epsilon$ for some pre-determined sufficiently small ϵ.

- Extensively applied in the semiparametric literature including

- Extensively applied in the semiparametric literature including
  - Quasi Likelihood Estimation (Severini and Staniwalis 1994);

- Extensively applied in the semiparametric literature including
  - Quasi Likelihood Estimation (Severini and Staniwalis 1994);
  - Semiparametric Mixture Models (Roeder et al 1996);

- Extensively applied in the semiparametric literature including
  - Quasi Likelihood Estimation (Severini and Staniwalis 1994);
  - Semiparametric Mixture Models (Roeder et al 1996);
  - Semiparametric Transformation Model (Linton et al 2008);

- Extensively applied in the semiparametric literature including
  - Quasi Likelihood Estimation (Severini and Staniwalis 1994);
  - Semiparametric Mixture Models (Roeder et al 1996);
  - Semiparametric Transformation Model (Linton et al 2008);
  - Generalized Partly Linear (Single Index) Model (Fan et al 1995);

- Extensively applied in the semiparametric literature including
  - Quasi Likelihood Estimation (Severini and Staniwalis 1994);
  - Semiparametric Mixture Models (Roeder et al 1996);
  - Semiparametric Transformation Model (Linton et al 2008);
  - Generalized Partly Linear (Single Index) Model (Fan et al 1995);
  - Survival Models (Huang 1996; Murphy et al., 1997);

- Extensively applied in the semiparametric literature including
  - Quasi Likelihood Estimation (Severini and Staniwalis 1994);
  - Semiparametric Mixture Models (Roeder et al 1996);
  - Semiparametric Transformation Model (Linton et al 2008);
  - Generalized Partly Linear (Single Index) Model (Fan et al 1995);
  - Survival Models (Huang 1996; Murphy et al., 1997);
- The above iterative procedure can also be adapted to the penalized estimation and selection of semiparametric models, e.g., Cheng and Zhang (2010).

- Extensively applied in the semiparametric literature including
  - Quasi Likelihood Estimation (Severini and Staniwalis 1994);
  - Semiparametric Mixture Models (Roeder et al 1996);
  - Semiparametric Transformation Model (Linton et al 2008);
  - Generalized Partly Linear (Single Index) Model (Fan et al 1995);
  - Survival Models (Huang 1996; Murphy et al., 1997);
- The above iterative procedure can also be adapted to the penalized estimation and selection of semiparametric models, e.g., Cheng and Zhang (2010).
- The choice of ϵ or k^* is quite arbitrary in the above papers.

In this talk, we will answer

"How many iterations do we really need?"

equivalently,

"$k^* =$? for obtaining a semiparametric efficient $\widehat{\theta}^{(k^*)}$?"

from a theoretical point of view.

In this talk, we will answer

"How many iterations do we really need?"

equivalently,

"$k^* =$? for obtaining a semiparametric efficient $\widehat{\theta}^{(k^*)}$?"

from a theoretical point of view.

- k^* depends on the convergence rates of $\widehat{\theta}^{(0)}$ and $\widehat{\eta}(\theta)$.

In this talk, we will answer

"How many iterations do we really need?"

equivalently,

"$k^* =$? for obtaining a semiparametric efficient $\widehat{\theta}^{(k^*)}$?"

from a theoretical point of view.

- k^* depends on the convergence rates of $\widehat{\theta}^{(0)}$ and $\widehat{\eta}(\theta)$.
  - k^* depends on the bandwidth order if kernel approach is used;

In this talk, we will answer

"How many iterations do we really need?"

equivalently,

"$k^* =$? for obtaining a semiparametric efficient $\widehat{\theta}^{(k^*)}$?"

from a theoretical point of view.

- k^* depends on the convergence rates of $\widehat{\theta}^{(0)}$ and $\widehat{\eta}(\theta)$.
  - k^* depends on the bandwidth order if kernel approach is used;
  - k^* depends on the order of the smoothing parameter if the penalization approach is used;

Why are our results useful?

Why are our results useful?

- Knowing the minimal k^* for each bootstrap sample will significantly reduce the bootstrap computational cost for making semiparametric inferences.

Grid Search of the Initial Estimate

- It is critical to initiate the iterations in a suitable neighborhood of the true value θ_0, i.e., $\widehat{\theta}^{(0)}$ is assumed to be n^{ψ}-consistent.

Grid Search of the Initial Estimate

- It is critical to initiate the iterations in a suitable neighborhood of the true value θ_0, i.e., $\widehat{\theta}^{(0)}$ is assumed to be n^{ψ}-consistent.
- Without any prior model information, a natural way is to do grid search of $\widehat{S}(\theta)$. Recall that $\widehat{S}(\theta) = \log lik_n(\theta, \widehat{\eta}(\theta))$.

Grid Search of the Initial Estimate

- It is critical to initiate the iterations in a suitable neighborhood of the true value θ_0, i.e., $\widehat{\theta}^{(0)}$ is assumed to be n^{ψ}-consistent.
- Without any prior model information, a natural way is to do grid search of $\widehat{S}(\theta)$. Recall that $\widehat{S}(\theta) = \log lik_n(\theta, \widehat{\eta}(\theta))$.
- We next present two types of grid search algorithm:

Grid Search of the Initial Estimate

- It is critical to initiate the iterations in a suitable neighborhood of the true value θ_0, i.e., $\widehat{\theta}^{(0)}$ is assumed to be n^{ψ}-consistent.
- Without any prior model information, a natural way is to do grid search of $\widehat{S}(\theta)$. Recall that $\widehat{S}(\theta) = \log lik_n(\theta, \widehat{\eta}(\theta))$.
- We next present two types of grid search algorithm:
  - Deterministic grid search;

Grid Search of the Initial Estimate

- It is critical to initiate the iterations in a suitable neighborhood of the true value θ_0, i.e., $\widehat{\theta}^{(0)}$ is assumed to be n^{ψ}-consistent.
- Without any prior model information, a natural way is to do grid search of $\widehat{S}(\theta)$. Recall that $\widehat{S}(\theta) = \log lik_n(\theta, \widehat{\eta}(\theta))$.
- We next present two types of grid search algorithm:
  - Deterministic grid search;
  - Stochastic grid search.

Grid Search of the Initial Estimate

- It is critical to initiate the iterations in a suitable neighborhood of the true value θ_0, i.e., $\widehat{\theta}^{(0)}$ is assumed to be n^{ψ}-consistent.
- Without any prior model information, a natural way is to do grid search of $\widehat{S}(\theta)$. Recall that $\widehat{S}(\theta) = \log lik_n(\theta, \widehat{\eta}(\theta))$.
- We next present two types of grid search algorithm:
  - Deterministic grid search;
  - Stochastic grid search.
- We will calculate the convergence rate of the above numerical outcome. The technical challenge is that $\widehat{S}(\theta)$ usually has no explicit form and may not be continuous/smooth.

Primary Assumptions

$S1.$ Θ is compact;

Primary Assumptions

$S1.$ Θ is compact;

$S2.$ [Asymptotic Uniqueness] For any random sequence $\{\widetilde{\theta}_n\} \in \Theta$,

$$[\widehat{S}(\widetilde{\theta}) - \widehat{S}(\widehat{\theta})]/n = o_P(1) \text{ implies that } \widetilde{\theta} - \theta_0 = o_P(1).$$

Primary Assumptions

$S1$. Θ is compact;

$S2$. [Asymptotic Uniqueness] For any random sequence $\{\widetilde{\theta}_n\} \in \Theta$,

$$[\widehat{S}(\widetilde{\theta}) - \widehat{S}(\widehat{\theta})]/n = o_P(1) \text{ implies that } \widetilde{\theta} - \theta_0 = o_P(1).$$

$S3$. [Asymptotic Concavity] For any consistent $\widetilde{\theta}$, $\widehat{S}(\cdot)$ satisfies

$$\begin{aligned} \widehat{S}(\widetilde{\theta}) &= \widehat{S}(\theta_0) + n(\widetilde{\theta} - \theta_0)'\mathbb{P}_n\widetilde{\ell}_0 - \frac{n}{2}(\widetilde{\theta} - \theta_0)'\widetilde{I}_0(\widetilde{\theta} - \theta_0) \\ &\quad +\Delta_n(\widetilde{\theta}), \end{aligned}$$

where $\Delta_n(\theta) = n\|\theta - \theta_0\|^3 \vee n^{1-2\gamma}\|\theta - \theta_0\|$ and γ represents the convergence rate of $\widehat{\eta}(\theta)$ given later.

Remark

- Condition S2 is usually satisfied if the model is identifiable.

Remark

- Condition S2 is usually satisfied if the model is identifiable.
- Condition S3 is very weak since we only require that $\widehat{S}(\theta)$ has such an asymptotic expansion, but not that $\widehat{S}(\theta)$ is continuous.

Remark

- Condition S2 is usually satisfied if the model is identifiable.
- Condition S3 is very weak since we only require that $\widehat{S}(\theta)$ has such an asymptotic expansion, but not that $\widehat{S}(\theta)$ is continuous.
- Condition S3 can be implied by some smoothness and empirical processes conditions (concerning about the least favorable submodel).

Deterministic Search

- Form a grid of cubes with sides of length $sn^{-\psi}$ over $\mathbb{R}^d$ for some $s > 0$ and $0 < \psi \leq 1/2$;

Deterministic Search

- Form a grid of cubes with sides of length $sn^{-\psi}$ over $\mathbb{R}^d$ for some $s > 0$ and $0 < \psi \leq 1/2$;
- Obtain a set of points $\mathcal{D}_n = \{\theta_{iD}\}$ regularly spaced throughout Θ with cardinality $card(\mathcal{D}_n) \geq Cn^{d\psi}$ for some $C > 0$;

Deterministic Search

- Form a grid of cubes with sides of length $sn^{-\psi}$ over $\mathbb{R}^d$ for some $s > 0$ and $0 < \psi \leq 1/2$;
- Obtain a set of points $\mathcal{D}_n = \{\theta_{iD}\}$ regularly spaced throughout Θ with cardinality $card(\mathcal{D}_n) \geq Cn^{d\psi}$ for some $C > 0$;
- Define $\widehat{\theta}_D^{(0)} = \arg\max_{\mathcal{D}_n} \widehat{S}(\theta)$.

Stochastic Search

However, the above deterministic search could be very slow if the dimension d of θ is high. This motivates the following computationally efficient stochastic search, i.e., of the order n^{ψ}.

Stochastic Search

However, the above deterministic search could be very slow if the dimension d of θ is high. This motivates the following computationally efficient stochastic search, i.e., of the order n^{ψ}.

- Assume $\bar{\theta}$ is independent of the data and admits a density having support Θ and bounded away from zero in some neighborhood of θ_0;

Stochastic Search

However, the above deterministic search could be very slow if the dimension d of θ is high. This motivates the following computationally efficient stochastic search, i.e., of the order n^{ψ}.

- Assume $\bar{\theta}$ is independent of the data and admits a density having support Θ and bounded away from zero in some neighborhood of θ_0;
- Let $\mathcal{S}_n$ be a set of realizations of $\bar{\theta}$ with $card(\mathcal{S}_n) \geq \widetilde{C} n^{\psi}$ for some $\widetilde{C} > 0$;

Stochastic Search

However, the above deterministic search could be very slow if the dimension d of θ is high. This motivates the following computationally efficient stochastic search, i.e., of the order n^{ψ}.

- Assume $\bar{\theta}$ is independent of the data and admits a density having support Θ and bounded away from zero in some neighborhood of θ_0;
- Let $\mathcal{S}_n$ be a set of realizations of $\bar{\theta}$ with $card(\mathcal{S}_n) \geq \widetilde{C} n^{\psi}$ for some $\widetilde{C} > 0$;
- Define $\widehat{\theta}_S^{(0)} = \arg\max_{\mathcal{S}_n} \widehat{S}(\theta)$.

Initial Estimate Theorem

Suppose Conditions S1-S3 hold. If $\widehat{\theta}$ is consistent and the efficient information matrix $\tilde{I}_0$ is nonsingular, then we have

$$\begin{aligned} \theta_D^{(0)} - \theta_0 &= O_P(n^{-\psi}), \qquad (1) \\ \theta_S^{(0)} - \theta_0 &= O_P(n^{-\psi}). \qquad (2) \end{aligned}$$

The above Theorem can be applied to a wide range of semiparametric models including the conditionally normal model, Cox model under survival data and semiparametric mixture model.

Semiparametric Maximum Likelihood Estimation

We first consider the maximum likelihood estimation of θ.

Semiparametric Maximum Likelihood Estimation

We first consider the maximum likelihood estimation of θ.

- For each fixed θ, η is estimated as a possibly nonsmooth NPMLE $\widehat{\eta}(\theta)$ (usually under some shape constraints);

Semiparametric Maximum Likelihood Estimation

We first consider the maximum likelihood estimation of θ.

- For each fixed θ, η is estimated as a possibly nonsmooth NPMLE $\widehat{\eta}(\theta)$ (usually under some shape constraints);
- $\widehat{S}(\theta)$ becomes the log-profile likelihood $\log pl_n(\theta)$;

Semiparametric Maximum Likelihood Estimation

We first consider the maximum likelihood estimation of θ.

- For each fixed θ, η is estimated as a possibly nonsmooth NPMLE $\widehat{\eta}(\theta)$ (usually under some shape constraints);
- $\widehat{S}(\theta)$ becomes the log-profile likelihood $\log pl_n(\theta)$;
- Challenge: the profile likelihood is defined as a supremum over an infinite dimensional parameter space, and thus has no closed form and is possibly nonsmooth (although it can be computed numerically in practice).

The Construction of $\widehat{\theta}^{(k)}$

We construct $\widehat{\theta}^{(k)}$ as the following Newton-Raphson form:

$$\widehat{\theta}^{(k)} = \widehat{\theta}^{(k-1)} + \left[\widehat{I}(\widehat{\theta}^{(k-1)}, t_n^{(k-1)})\right]^{-1} \widehat{\ell}(\widehat{\theta}^{(k-1)}, s_n^{(k-1)}), \qquad (3)$$

where $\widehat{\ell}(\theta, s_n)$ and $\widehat{I}(\theta, t_n)$ are the first and second numerical derivatives of $\log pl_n(\theta)$ with step sizes $s_n, t_n \to 0$, respectively.

The Construction of $\widehat{\theta}^{(k)}$

We construct $\widehat{\theta}^{(k)}$ as the following Newton-Raphson form:

$$\widehat{\theta}^{(k)} = \widehat{\theta}^{(k-1)} + \left[\widehat{I}(\widehat{\theta}^{(k-1)}, t_n^{(k-1)})\right]^{-1} \widehat{\ell}(\widehat{\theta}^{(k-1)}, s_n^{(k-1)}), \qquad (3)$$

where $\widehat{\ell}(\theta, s_n)$ and $\widehat{I}(\theta, t_n)$ are the first and second numerical derivatives of $\log pl_n(\theta)$ with step sizes $s_n, t_n \to 0$, respectively.

- We use the numerical derivatives of $\log pl_n(\theta)$ since its differentiability is usually unknown;

The Construction of $\widehat{\theta}^{(k)}$

We construct $\widehat{\theta}^{(k)}$ as the following Newton-Raphson form:

$$\widehat{\theta}^{(k)} = \widehat{\theta}^{(k-1)} + \left[\widehat{I}(\widehat{\theta}^{(k-1)}, t_n^{(k-1)})\right]^{-1} \widehat{\ell}(\widehat{\theta}^{(k-1)}, s_n^{(k-1)}), \qquad (3)$$

where $\widehat{\ell}(\theta, s_n)$ and $\widehat{I}(\theta, t_n)$ are the first and second numerical derivatives of $\log pl_n(\theta)$ with step sizes $s_n, t_n \to 0$, respectively.

- We use the numerical derivatives of $\log pl_n(\theta)$ since its differentiability is usually unknown;
- A close inspection of (3) reveals that we have constructed $\widehat{\theta}^{(k)}$ even without knowing the forms of efficient score function $\widetilde{\ell}_0$ and efficient information matrix $\widetilde{I}_0$.

Intuition of calculating k^*:

Intuition of calculating k^*:

- We expect that $\widehat{\theta}^{(k)}$ approaches to MLE $\widehat{\theta}$, which is exactly the maximizer of $\log pl_n(\theta)$, asymptotically as $k \to \infty$ if $\widehat{\ell}(\cdot)$ and $\widehat{I}(\cdot)$ are consistent estimators of $\widetilde{\ell}_0$ and $\widehat{I}_0$.

Intuition of calculating k^*:

- We expect that $\widehat{\theta}^{(k)}$ approaches to MLE $\widehat{\theta}$, which is exactly the maximizer of $\log pl_n(\theta)$, asymptotically as $k \to \infty$ if $\widehat{\ell}(\cdot)$ and $\widehat{I}(\cdot)$ are consistent estimators of $\widetilde{\ell}_0$ and $\widetilde{I}_0$.
  - The above consistency can be guaranteed by the concave form of $\log pl_n(\theta)$ shown in Murphy and van der Vaart (2000).

Intuition of calculating k^*:

- We expect that $\widehat{\theta}^{(k)}$ approaches to MLE $\widehat{\theta}$, which is exactly the maximizer of $\log pl_n(\theta)$, asymptotically as $k \to \infty$ if $\widehat{\ell}(\cdot)$ and $\widehat{I}(\cdot)$ are consistent estimators of $\widetilde{\ell}_0$ and $\widetilde{I}_0$.
  - The above consistency can be guaranteed by the concave form of $\log pl_n(\theta)$ shown in Murphy and van der Vaart (2000).
- We can further quantify how fast $\widehat{\theta}^{(k)}$ converges to $\widehat{\theta}$ if we know how fast $\widehat{\ell}$ $(\widehat{I})$ converges to $\widetilde{\ell}_0$ $(\widetilde{I}_0)$.

Intuition of calculating k^*:

- We expect that $\widehat{\theta}^{(k)}$ approaches to MLE $\widehat{\theta}$, which is exactly the maximizer of $\log pl_n(\theta)$, asymptotically as $k \to \infty$ if $\widehat{\ell}(\cdot)$ and $\widehat{I}(\cdot)$ are consistent estimators of $\widetilde{\ell}_0$ and $\widetilde{I}_0$.
  - The above consistency can be guaranteed by the concave form of $\log pl_n(\theta)$ shown in Murphy and van der Vaart (2000).
- We can further quantify how fast $\widehat{\theta}^{(k)}$ converges to $\widehat{\theta}$ if we know how fast $\widehat{\ell}$ ($\widehat{I}$) converges to $\widetilde{\ell}_0$ ($\widetilde{I}_0$).
  - The above convergence rates of $\widehat{\ell}(\cdot)$ and $\widehat{I}(\cdot)$ are derived based on a higher order quadratic expansion of $\log pl_n(\theta)$.

Primary Assumptions

I. Regularity conditions on the least favorable submodels (guarantee the valid higher order quadratic expansion of the log-profile likelihood);

Primary Assumptions

I. Regularity conditions on the least favorable submodels (guarantee the valid higher order quadratic expansion of the log-profile likelihood);

II. We assume that, for any random sequence $\tilde{\theta}_n \xrightarrow{P} \theta_0$,

$$\|\widehat{\eta}(\tilde{\theta}_n) - \eta_0\| = O_P(\|\tilde{\theta}_n - \theta_0\| \vee n^{-\gamma}), \tag{4}$$

where $\|\cdot\|$ is some norm in $\mathcal{H}$, for some $1/4 < \gamma \leq 1/2$.

Theorem 1

Suppose Conditions *I&II* hold and proper step sizes are chosen. Let $\|\widehat{\theta}^{(0)}-\theta_0\|=O_P(n^{-\psi})$ and $\|\widehat{\theta}^{(k)}-\widehat{\theta}\|=O_P(n^{-r_k})$.

We show

Theorem 1

Suppose Conditions *I&II* hold and proper step sizes are chosen. Let $\|\widehat{\theta}^{(0)} - \theta_0\| = O_P(n^{-\psi})$ and $\|\widehat{\theta}^{(k)} - \widehat{\theta}\| = O_P(n^{-r_k})$.

We show

- r_k continuously increases from ψ to $(\gamma + 1/4)$ as $k \to \infty$;

Theorem 1

Suppose Conditions *I&II* hold and proper step sizes are chosen. Let $\|\widehat{\theta}^{(0)} - \theta_0\| = O_P(n^{-\psi})$ and $\|\widehat{\theta}^{(k)} - \widehat{\theta}\| = O_P(n^{-r_k})$.

We show

- r_k continuously increases from ψ to $(\gamma + 1/4)$ as $k \to \infty$;
- Specifically,

$$\|\widehat{\theta}^{(k)} - \widehat{\theta}\| = O_P(n^{-S(\psi,\gamma,k)}),$$

where $S(\psi, \gamma, k)$ has an easy-to-calculate explicit form.

Theorem 1

Suppose Conditions *I&II* hold and proper step sizes are chosen. Let $\|\widehat{\theta}^{(0)} - \theta_0\| = O_P(n^{-\psi})$ and $\|\widehat{\theta}^{(k)} - \widehat{\theta}\| = O_P(n^{-r_k})$.

We show

- r_k continuously increases from ψ to $(\gamma + 1/4)$ as $k \to \infty$;
- Specifically,

$$\|\widehat{\theta}^{(k)} - \widehat{\theta}\| = O_P(n^{-S(\psi,\gamma,k)}),$$

  where $S(\psi, \gamma, k)$ has an easy-to-calculate explicit form.
- $\|\widehat{\theta}^{(k^*)} - \widehat{\theta}\| = o_P(n^{-1/2})$ for $k^* \geq K(\psi, \gamma)$.

Remark 1

- It is well known that one-step estimate is efficient given that $\widehat{\theta}^{(0)}$ is $\sqrt{n}$-consistent. However, this is not enough for the semiparametric estimation since $\widehat{\theta}^{(0)}$ may have slower than $\sqrt{n}$ rate as shown in the previous Theorem.

Remark 1

- It is well known that one-step estimate is efficient given that $\widehat{\theta}^{(0)}$ is $\sqrt{n}$-consistent. However, this is not enough for the semiparametric estimation since $\widehat{\theta}^{(0)}$ may have slower than $\sqrt{n}$ rate as shown in the previous Theorem.
- More than k^* iterations, say k, will only improve the higher order asymptotic efficiency of $\widehat{\theta}^{(k)}$.

Remark 1

- It is well known that one-step estimate is efficient given that $\widehat{\theta}^{(0)}$ is $\sqrt{n}$-consistent. However, this is not enough for the semiparametric estimation since $\widehat{\theta}^{(0)}$ may have slower than $\sqrt{n}$ rate as shown in the previous Theorem.
- More than k^* iterations, say k, will only improve the higher order asymptotic efficiency of $\widehat{\theta}^{(k)}$.
- Interestingly, the lower bound of $\|\widehat{\theta}^{(k)} - \widehat{\theta}\|$, i.e. $O_P(n^{-\gamma-1/4})$, is intrinsically decided, i.e., only dependent on the convergence rate of the nuisance parameter.

Example I: Cox Model with Current Status Data

The hazard function $\lambda(t|z)$ of the survival time T given the covariate Z is modeled as, with λ as the hazard function,

$$\lambda(t)\exp(\theta' z).$$

Current status data: observe $X = (Y, I\{T \leq Y\}, Z)$, where Y is the examination time.

We are interested in θ while treating the cumulative hazard function $\eta(y) = \int_0^y \lambda(t)dt$ as the nuisance parameter. The NPMLE $\widehat{\eta}(\theta)$ and nonsmooth $\log pl_n(\theta)$ have no explicit forms, but can be calculated numerically via isotonic regression type algorithm.

The convergence rate of $\widehat{\eta}(\theta)$ is shown to be $n^{-1/3}$. Thus, Theorem 1 implies the following table with $O_P(n^{-7/12})$ as the lower bound.

Table 1. *Cox Model under Current Status Data* ($\gamma = 1/3$)

$\psi = 1/2$	$\psi = 1/3$	$\psi = 1/4$
$r_1 = 7/12$	$r_1 = 1/2, r_2 = 7/12$	$r_1 = 3/8, r_2 = 25/48, r_3 = 7/12$
$k^* = 1$	$k^* = 2$	$k^* = 2$

Recall that $\|\widehat{\theta}^{(k)} - \widehat{\theta}\| = O_P(n^{-r_k})$.

In contrast with the nonsmooth $\log pl_n(\theta)$, the regularized $\widehat{S}(\theta)$ is usually differentiable although its form may vary under different regularizations.

In contrast with the nonsmooth $\log pl_n(\theta)$, the regularized $\widehat{S}(\theta)$ is usually differentiable although its form may vary under different regularizations.

A. We first present a unified framework for studying $\widehat{\theta}^{(k)}$ when $\widehat{S}(\theta)$ is third order differentiable.

In contrast with the nonsmooth $\log pl_n(\theta)$, the regularized $\widehat{S}(\theta)$ is usually differentiable although its form may vary under different regularizations.

A. We first present a unified framework for studying $\widehat{\theta}^{(k)}$ when $\widehat{S}(\theta)$ is third order differentiable.

B. We then discuss two special cases: kernel estimation of θ and penalized estimation of θ, in two examples.

In contrast with the nonsmooth $\log pl_n(\theta)$, the regularized $\widehat{S}(\theta)$ is usually differentiable although its form may vary under different regularizations.

A. We first present a unified framework for studying $\widehat{\theta}^{(k)}$ when $\widehat{S}(\theta)$ is third order differentiable.

B. We then discuss two special cases: kernel estimation of θ and penalized estimation of θ, in two examples.

C. In the end, we consider the semiparametric model selection as an extension of the penalized estimation.

Construction of $\widehat{\theta}^{(k)}$

Let $\widehat{\ell}(\cdot) = \widehat{S}^{(1)}(\cdot)/n$ and $\widehat{I}(\cdot) = -\widehat{S}^{(2)}(\cdot)/n$. We construct $\widehat{\theta}^{(k)}$ as

$$\widehat{\theta}^{(k)} = \widehat{\theta}^{(k-1)} + \left[\widehat{I}(\widehat{\theta}^{(k-1)})\right]^{-1} \widehat{\ell}(\widehat{\theta}^{(k-1)}). \tag{5}$$

However, $\widehat{I}(\cdot)$ can also be constructed as the negative numerical derivative of $\widehat{\ell}(\cdot)$ when $\widehat{S}^{(2)}(\cdot)$ has no explicit form or is hard to compute.

Primary Conditions

The higher order quadratic expansion of $\widehat{S}(\theta)$ is valid under the following Condition G:

$$\frac{1}{n}\widehat{S}^{(1)}(\theta_0) - \frac{1}{n}S^{(1)}(\theta_0) = O_P(n^{-2g}), \tag{6}$$

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{1}{n}\widehat{S}^{(2)}(\theta) - \frac{1}{n}S^{(2)}(\theta)\right| = O_P(n^{-g}), \tag{7}$$

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{1}{n}\widehat{S}^{(3)}(\theta)\right| = O_P(1), \tag{8}$$

where $S(\theta) = \sup_{\eta\in\mathcal{H}} E\log lik(\theta,\eta)$ and $1/4 < g \leq 1/2$.

Theorem 2

Suppose Condition G holds and define $\widehat{\theta}$ as the maximizer of $\widehat{S}(\theta)$. Recall that $\|\widehat{\theta}^{(0)} - \theta_0\| = O_P(n^{-\psi})$ and $\|\widehat{\theta}^{(k)} - \widehat{\theta}\| = O_P(n^{-r_k})$.

Theorem 2

Suppose Condition G holds and define $\widehat{\theta}$ as the maximizer of $\widehat{S}(\theta)$. Recall that $\|\widehat{\theta}^{(0)} - \theta_0\| = O_P(n^{-\psi})$ and $\|\widehat{\theta}^{(k)} - \widehat{\theta}\| = O_P(n^{-r_k})$.

- $\widehat{\theta}$ is semiparametric efficient;

Theorem 2

Suppose Condition G holds and define $\widehat{\theta}$ as the maximizer of $\widehat{S}(\theta)$. Recall that $\|\widehat{\theta}^{(0)} - \theta_0\| = O_P(n^{-\psi})$ and $\|\widehat{\theta}^{(k)} - \widehat{\theta}\| = O_P(n^{-r_k})$.

- $\widehat{\theta}$ is semiparametric efficient;
- r_k continuously increases from ψ to $+\infty$ as $k \to \infty$;

Theorem 2

Suppose Condition G holds and define $\widehat{\theta}$ as the maximizer of $\widehat{S}(\theta)$. Recall that $\|\widehat{\theta}^{(0)} - \theta_0\| = O_P(n^{-\psi})$ and $\|\widehat{\theta}^{(k)} - \widehat{\theta}\| = O_P(n^{-r_k})$.

- $\widehat{\theta}$ is semiparametric efficient;
- r_k continuously increases from ψ to $+\infty$ as $k \to \infty$;
- If $\widehat{I}$ is constructed analytically, then

$$\begin{aligned} \|\widehat{\theta}^{(k)} - \widehat{\theta}\| &= O_P(n^{-2^k \psi}), \\ k^* &= L_1(\psi). \end{aligned}$$

Theorem 2

Suppose Condition G holds and define $\widehat{\theta}$ as the maximizer of $\widehat{S}(\theta)$. Recall that $\|\widehat{\theta}^{(0)} - \theta_0\| = O_P(n^{-\psi})$ and $\|\widehat{\theta}^{(k)} - \widehat{\theta}\| = O_P(n^{-r_k})$.

- $\widehat{\theta}$ is semiparametric efficient;
- r_k continuously increases from ψ to $+\infty$ as $k \to \infty$;
- If $\widehat{I}$ is constructed analytically, then

$$\begin{aligned} \|\widehat{\theta}^{(k)} - \widehat{\theta}\| &= O_P(n^{-2^k\psi}), \\ k^* &= L_1(\psi). \end{aligned}$$

- If $\widehat{I}$ is constructed numerically, then

$$\begin{aligned} \|\widehat{\theta}^{(k)} - \widehat{\theta}\| &= O_P(n^{-R(\psi,g,k)}), \\ k^* &= L_2(\psi, g). \end{aligned}$$

Four interesting applications of Theorem 2:

Four interesting applications of Theorem 2:

1. Parametric models where $\widehat{S}(\theta) = \log lik_n(\theta)$;

Four interesting applications of Theorem 2:

1. Parametric models where $\widehat{S}(\theta) = \log lik_n(\theta)$;
2. Kernel estimation of conditionally parametric models, i.e.,

$$\widehat{\eta}(\theta)(z) = \arg \sup_{\eta \in C^2(\mathcal{Z})} \sum_{i=1}^{n} \log lik(X_i; \theta, \eta(Z_i)) K\left(\frac{z - Z_i}{b_n}\right).$$

Condition G can be translated to the kernel conditions on $K(\cdot)$ and b_n. Therefore, k^* is related to the order of b_n in this case.

3. Penalized estimation of semiparametric models, i.e.,

$$\widehat{\eta}_{\lambda_n}(\theta) = \arg \sup_{\eta \in \mathcal{H}_k} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log lik(X_i; \theta, \eta) - \lambda_n^2 J^2(\eta) \right\}.$$

In this case, Condition G needs to be modified to take into account of λ_n. Therefore, k^* is related to the order of the smoothing parameter λ_n.

3. Penalized estimation of semiparametric models, i.e.,

$$\widehat{\eta}_{\lambda_n}(\theta) = \arg \sup_{\eta \in \mathcal{H}_k} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log lik(X_i; \theta, \eta) - \lambda_n^2 J^2(\eta) \right\}.$$

   In this case, Condition G needs to be modified to take into account of λ_n. Therefore, k^* is related to the order of the smoothing parameter λ_n.

4. Semiparametric model selection in high dimensional data. We will show that k^* iterations are also sufficient to recover the estimation sparsity.

Example II: Conditionally Normal Model

We assume that $Y|(W=w, Z=z) \sim N(\theta' w, \eta(z))$ and thus

$$\widehat{\eta}_\theta(z) = \frac{\sum_{i=1}^n (Y-\theta' W)^2 K((z-Z_i)/b_n)}{\sum_{i=1}^n K((z-Z_i)/b_n)}.$$

Given that $b_n \asymp n^{-1/5}$, Theorem 2 gives the following table.

Table 2. *Conditional Normal Model*

$\psi = 1/4$
$r_1 = 151/600, r_2 = 153/600, r_3 = 157/600, r_4 = 165/600$
$r_5 = 181/600, r_6 = 213/600, r_7 = 277/600, r_8 = 405/600$
$k^* = 8$

Example III: Sparse and Efficient Est. of Partial Spline Model

We consider the partial smoothing spline model:

$$Y = W'\theta + \eta(Z) + \epsilon, \tag{9}$$

where η belongs to the k-th order Sobolev space.

Under high dimensional data, it is common to assume that some components of θ_0 are exactly zero.

To achieve the estimation efficiency and sparsity of θ, Cheng and Zhang (2010) proposed the below regularization method

$$\min\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - W_i'\theta - \eta(Z_i))^2 + \lambda_n^2 J^2(\eta) + \tau_n^2 \sum_{j=1}^{d}\frac{|\theta_j|}{|\widetilde{\theta}_j|}\right\},$$

where $\widetilde{\theta} = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_d)'$ is the consistent initial estimate. By incorporating LARS algorithm, the general iterative estimation procedure can also be adapted to this scenario.

Given that $\widetilde{\theta}$ is $\sqrt{n}$-consistent, e.g., partial smoothing spline estimate, $\lambda_n \asymp n^{-k/(2k+1)}$ and $\tau_n \asymp n^{-k/(2k+1)}$, Theorem 2 (after adaptations) shows that $k^* = 1$ is sufficient to produce the efficient and sparse estimate of θ.

Thank you for your attention....

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