The Long March Towards Joint Asymptotics: My 1st Steps...

Guang Cheng

Department of Statistics, Purdue University

Duke Data Seminar on March 28, 2013
Joint Work with Zuofeng Shang in Notre Dame

Outline

Introduction: Motivated Example

- A simple and widely used model:

$$Y = X'\theta_0 + f_0(Z) + \epsilon,$$

where ϵ has mean zero and f_0 is assumed to be smooth.

Introduction: Motivated Example

- A simple and widely used model:

$$Y = X'\theta_0 + f_0(Z) + \epsilon,$$

where ϵ has mean zero and f_0 is assumed to be smooth.

- Under the penalized least square estimation, it is shown that

$$\begin{aligned} \sqrt{n}(\widehat{\theta} - \theta_0) &\xrightarrow{d} N(0, V_{\theta_0}), \\ \|\widehat{f} - f_0\|_2 &= O_P(n^{-2/5}). \end{aligned}$$

We call the above results as Marginal Asymptotics.

Introduction: Motivated Example

- A simple and widely used model:

$$Y = X'\theta_0 + f_0(Z) + \epsilon,$$

where ϵ has mean zero and f_0 is assumed to be smooth.

- Under the penalized least square estimation, it is shown that

$$\begin{aligned}\sqrt{n}(\widehat{\theta} - \theta_0) &\xrightarrow{d} N(0, V_{\theta_0}),\\ \|\widehat{f} - f_0\|_2 &= O_P(n^{-2/5}).\end{aligned}$$

We call the above results as Marginal Asymptotics.

- $(\sqrt{n}(\widehat{\theta} - \theta_0), n^{2/5}(\widehat{f}(z_0) - f_0(z_0))) \xrightarrow{d} ?$

Introduction: Motivated Example

- A simple and widely used model:

$$Y = X'\theta_0 + f_0(Z) + \epsilon,$$

where ϵ has mean zero and f_0 is assumed to be smooth.

- Under the penalized least square estimation, it is shown that

$$\begin{aligned} \sqrt{n}(\widehat{\theta} - \theta_0) &\xrightarrow{d} N(0, V_{\theta_0}), \\ \|\widehat{f} - f_0\|_2 &= O_P(n^{-2/5}). \end{aligned}$$

We call the above results as Marginal Asymptotics.

- $(\sqrt{n}(\widehat{\theta} - \theta_0), n^{2/5}(\widehat{f}(z_0) - f_0(z_0))) \xrightarrow{d} ?$
- $H_0 : \theta = \theta_0$ and $f(z_0) = w_0$ v.s. $H_A : H_0$ does not hold?

Introduction: Motivated Example

- A simple and widely used model:

$$Y = X'\theta_0 + f_0(Z) + \epsilon,$$

where ϵ has mean zero and f_0 is assumed to be smooth.

- Under the penalized least square estimation, it is shown that

$$\begin{aligned} \sqrt{n}(\widehat{\theta} - \theta_0) &\xrightarrow{d} N(0, V_{\theta_0}), \\ \|\widehat{f} - f_0\|_2 &= O_P(n^{-2/5}). \end{aligned}$$

We call the above results as Marginal Asymptotics.

- $(\sqrt{n}(\widehat{\theta} - \theta_0), n^{2/5}(\widehat{f}(z_0) - f_0(z_0))) \xrightarrow{d} ?$
- $H_0 : \theta = \theta_0$ and $f(z_0) = w_0$ v.s. $H_A : H_0$ does not hold?
- $H_0 : \theta = \theta_0$ and $f \in \mathcal{F}_0$ v.s. $H_A : H_0$ does not hold?

Introduction: General Aim

- In general, we consider the *Semi-Nonparametric Models* in which the (finite dimensional) Euclidean parameter θ and (infinite dimensional) nonparametric parameter f are both of interest. For example, to understand the recent financial crisis, the semi-nonparametric copula models are applied to address tail dependence among shocks to different financial series and also to recover the shapes of the impact curve for individual financial series.

Introduction: General Aim

- In general, we consider the *Semi-Nonparametric Models* in which the (finite dimensional) Euclidean parameter θ and (infinite dimensional) nonparametric parameter f are both of interest. For example, to understand the recent financial crisis, the semi-nonparametric copula models are applied to address tail dependence among shocks to different financial series and also to recover the shapes of the impact curve for individual financial series.
- Define $\widehat{\theta}$ and $\widehat{f}$ as the estimate for θ_0 and f_0 under some type of regularization, e.g., penalized/local polynomial estimation.

Introduction: General Aim

- In this talk, we want to address the following questions:

Introduction: General Aim

- In this talk, we want to address the following questions:
  - What is the joint asymptotics of $(\widehat{\theta}, \widehat{f})$ especially when the two estimates have possibly different convergence rates and limiting distributions?

Introduction: General Aim

- In this talk, we want to address the following questions:
  - What is the joint asymptotics of $(\widehat{\theta}, \widehat{f})$ especially when the two estimates have possibly different convergence rates and limiting distributions?
  - How to make valid joint inferences for (θ_0, f_0)?

Introduction: General Aim

- In this talk, we want to address the following questions:
  - What is the joint asymptotics of $(\widehat{\theta}, \widehat{f})$ especially when the two estimates have possibly different convergence rates and limiting distributions?
  - How to make valid joint inferences for (θ_0, f_0)?
  - How to produce a valid prediction interval for the response given new data?

Introduction: General Aim

- In this talk, we want to address the following questions:
  - What is the joint asymptotics of $(\widehat{\theta}, \widehat{f})$ especially when the two estimates have possibly different convergence rates and limiting distributions?
  - How to make valid joint inferences for (θ_0, f_0)?
  - How to produce a valid prediction interval for the response given new data?
- Two important by-products

Introduction: General Aim

- In this talk, we want to address the following questions:
  - What is the joint asymptotics of $(\widehat{\theta}, \widehat{f})$ especially when the two estimates have possibly different convergence rates and limiting distributions?
  - How to make valid joint inferences for (θ_0, f_0)?
  - How to produce a valid prediction interval for the response given new data?
- Two important by-products
  - *Local and Global Asymptotic Inferences for the Smoothing Spline*, e.g., prove the inconsistency of Wahba's Bayesian C.I. and correct it; see Shang and Cheng (2012) in my homepage.

Introduction: General Aim

- In this talk, we want to address the following questions:
  - What is the joint asymptotics of $(\widehat{\theta}, \widehat{f})$ especially when the two estimates have possibly different convergence rates and limiting distributions?
  - How to make valid joint inferences for (θ_0, f_0)?
  - How to produce a valid prediction interval for the response given new data?
- Two important by-products
  - *Local and Global Asymptotic Inferences for the Smoothing Spline*, e.g., prove the inconsistency of Wahba's Bayesian C.I. and correct it; see Shang and Cheng (2012) in my homepage.
  - Rectify a common intuition in the literature that the (point-wise) marginal asymptotics/inferences for the nonparametric component remains the same if the added Euclidean parameter can be estimated at a faster rate (as if it were known), say root-n rate. This intuition is true only in the special cases.

Introduction: Literature Reivew

- Existing semiparametric literature focuses on the asymptotic behaviors of $\widehat{\theta}$, i.e., root-n asymptotic normality and semiparametric efficiency, by applying the so called “profile” method. The convergence rate of $\widehat{f}$ is derived as a by-product.

Introduction: Literature Reivew

- Existing semiparametric literature focuses on the asymptotic behaviors of $\widehat{\theta}$, i.e., root-n asymptotic normality and semiparametric efficiency, by applying the so called "profile" method. The convergence rate of $\widehat{f}$ is derived as a by-product.
- As far as I am aware, the only general theory in dealing with the joint asymptotics is Theorem 3.3.1 in van der Vaart and Wellner (1996) (widely applied to the survival models). However, they require both parameters be estimated at the same root-n rate (and thus can be treated as one parameter).

Introduction: Literature Reivew

- Existing semiparametric literature focuses on the asymptotic behaviors of $\widehat{\theta}$, i.e., root-n asymptotic normality and semiparametric efficiency, by applying the so called "profile" method. The convergence rate of $\widehat{f}$ is derived as a by-product.
- As far as I am aware, the only general theory in dealing with the joint asymptotics is Theorem 3.3.1 in van der Vaart and Wellner (1996) (widely applied to the survival models). However, they require both parameters be estimated at the same root-n rate (and thus can be treated as one parameter).
- No existing theory applies to general semi-nonparametric models where $\widehat{\theta}$ and $\widehat{f}$ may converge at different rates.

General Framework: Model Assumptions

General Framework: Model Assumptions

- Observe the response Y, linear covariate X and nonlinear covariate Z; denoted as $T = (Y, X, Z)$.

General Framework: Model Assumptions

- Observe the response Y, linear covariate X and nonlinear covariate Z; denoted as $T = (Y, X, Z)$.
- Assume $E\{Y|U\} = F(X'\theta_0 + f_0(Z))$, where $U = (X, Z)$ and $F(\cdot)$ is some known link function.

General Framework: Model Assumptions

- Observe the response Y, linear covariate X and nonlinear covariate Z; denoted as $T = (Y, X, Z)$.
- Assume $E\{Y|U\} = F(X'\theta_0 + f_0(Z))$, where $U = (X, Z)$ and $F(\cdot)$ is some known link function.
- Consider the following general criterion function $\ell(y; a)$:

General Framework: Model Assumptions

- Observe the response Y, linear covariate X and nonlinear covariate Z; denoted as $T = (Y, X, Z)$.
- Assume $E\{Y|U\} = F(X'\theta_0 + f_0(Z))$, where $U = (X, Z)$ and $F(\cdot)$ is some known link function.
- Consider the following general criterion function $\ell(y; a)$:
  - *Generalized Partly Linear Models*:

$$\ell(y; x'\theta + f(z)) = \log p(y|F(x'\theta + f(z))),$$

    where p is some known conditional distribution;

General Framework: Model Assumptions

- Observe the response Y, linear covariate X and nonlinear covariate Z; denoted as $T = (Y, X, Z)$.
- Assume $E\{Y|U\} = F(X'\theta_0 + f_0(Z))$, where $U = (X, Z)$ and $F(\cdot)$ is some known link function.
- Consider the following general criterion function $\ell(y; a)$:
  - *Generalized Partly Linear Models*:

    $$\ell(y; x'\theta + f(z)) = \log p(y|F(x'\theta + f(z))),$$

    where p is some known conditional distribution;
  - *Quasi-likelihood Models*:

    $$\ell(y; x'\theta + f(z)) = Q(y; F(x'\theta + f(z))),$$

    where Q is the quasi-likelihood defined via moment knowledge.

General Framework: Parameter Assumptions

- Denote $H_0 = (\theta_0, f_0)$ with $\theta_0 \in \mathbb{R}^p$ and $f_0 \in S^m([0,1])$, i.e., the m-th order Sobolev space, for $m > 1/2$.

General Framework: Parameter Assumptions

- Denote $H_0 = (\theta_0, f_0)$ with $\theta_0 \in \mathbb{R}^p$ and $f_0 \in S^m([0,1])$, i.e., the m-th order Sobolev space, for $m > 1/2$.
- Assumption A.1

General Framework: Parameter Assumptions

- Denote $H_0 = (\theta_0, f_0)$ with $\theta_0 \in \mathbb{R}^p$ and $f_0 \in S^m([0,1])$, i.e., the m-th order Sobolev space, for $m > 1/2$.
- Assumption A.1
  - Smoothness and Exponential Tail Conditions on $\ell(y; a)$;

General Framework: Parameter Assumptions

- Denote $H_0 = (\theta_0, f_0)$ with $\theta_0 \in \mathbb{R}^p$ and $f_0 \in S^m([0,1])$, i.e., the m-th order Sobolev space, for $m > 1/2$.
- Assumption A.1
  - Smoothness and Exponential Tail Conditions on $\ell(y; a)$;
  - Fisher Information Condition:

$$1/M \leq I(U) \equiv E\{\ddot{\ell}_a(Y; X'\theta_0 + f_0(Z))|U\} \leq M \quad a.s.;$$

General Framework: Parameter Assumptions

- Denote $H_0 = (\theta_0, f_0)$ with $\theta_0 \in \mathbb{R}^p$ and $f_0 \in S^m([0,1])$, i.e., the m-th order Sobolev space, for $m > 1/2$.
- Assumption A.1
  - Smoothness and Exponential Tail Conditions on $\ell(y; a)$;
  - Fisher Information Condition:

$$1/M \leq I(U) \equiv E\{\ddot{\ell}_a(Y; X'\theta_0 + f_0(Z))|U\} \leq M \quad a.s.;$$

  - $\epsilon \equiv \dot{\ell}_a(Y; X'\theta_0 + f_0(Z))$ satisfies $E(\epsilon|U) = 0$ and $E(\epsilon^2|U) = I(U)$ a.s..

General Framework: Parameter Assumptions

- Denote $H_0 = (\theta_0, f_0)$ with $\theta_0 \in \mathbb{R}^p$ and $f_0 \in S^m([0,1])$, i.e., the m-th order Sobolev space, for $m > 1/2$.
- Assumption A.1
  - Smoothness and Exponential Tail Conditions on $\ell(y; a)$;
  - Fisher Information Condition:

$$1/M \le I(U) \equiv E\{\ddot{\ell}_a(Y; X'\theta_0 + f_0(Z))|U\} \le M \quad a.s.;$$

  - $\epsilon \equiv \dot{\ell}_a(Y; X'\theta_0 + f_0(Z))$ satisfies $E(\epsilon|U) = 0$ and $E(\epsilon^2|U) = I(U)$ a.s..
- The above assumptions are mild, and are easily satisfied by (i) Partly Linear Model with Normal Error; (ii) Semiparametric Gamme Model: $Y|X, Z \sim Gamma(\alpha, \exp(X'\theta_0 + f_0(Z)))$; (iii) Semiparametric Logistic Regression.

General Framework:Penalized Estimation

Our penalized estimate $\widehat{H}_{n,\lambda}$ is defined as

$$\widehat{H}_{n,\lambda} \equiv (\widehat{\theta}_{n,\lambda}, \widehat{f}_{n,\lambda}) = \arg\max \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i; X_i^T \theta + f(Z_i)) - \lambda J(f, f) \right\},$$

where $J(f, f) = \int_0^1 |f^{(m)}(z)|^2 dz$ is the roughness penalty and $\lambda \to 0$ is the smoothing parameter.

The local polynomial estimation will be discussed later.

Main Results: Preliminary

As a RKHS, $S^m([0,1])$ has the reproducing kernel $K(z_1, z_2)$:

$$K_z(\cdot) \equiv K(z, \cdot) \in S^m([0,1]) \text{ and } \langle K_z, f\rangle_1 = f(z),$$

where the inner product $\langle f, \widetilde{f}\rangle_1 \equiv E\{B(Z)f(Z)\widetilde{f}(Z)\} + \lambda J(f, \widetilde{f})$ and $B(Z) = E\{I(U)|Z\}$, and induces a p.d. operator W_λ:

$$\langle W_\lambda f, \widetilde{f}\rangle_1 = \lambda J(f, \widetilde{f}).$$

Our first contribution is to construct a semiparametric extension of K_z and W_λ, i.e., R_u abd P_λ:

$$\langle R_u, H\rangle = H(u) = x'\theta + f(z) \text{ and } \langle P_\lambda H, \widetilde{H}\rangle = \lambda J(f, \widetilde{f}),$$

where the inner product $\langle H, \widetilde{H}\rangle = E\{I(U)H(U)\widetilde{H}(U)\} + \lambda J(f, \widetilde{f})$.

Main Results: Theorem 1 (Joint Asymptotics I)

Denote $H_0^* = H_0 - P_\lambda H_0 = (\theta_0^*, f_0^*)$ as the biased center. Under suitable range of λ, including $\lambda \asymp n^{-2m/(2m+1)}$, we have, for any $z_0 \in [0,1]$,

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0^*) \\ \sqrt{n\lambda^{1/2m}}(\widehat{f}_{n,\lambda}(z_0) - f_0^*(z_0)) \end{pmatrix} \xrightarrow{d} N(0, \Psi^*),$$

where

$$\Psi^* = \begin{pmatrix} \Omega^{-1} & \Omega^{-1}(\alpha_{z_0} + \beta_{z_0}) \\ (\alpha_{z_0} + \beta_{z_0})^T \Omega^{-1} & \sigma_{z_0}^2 + 2\beta_{z_0}^T \Omega^{-1} \alpha_{z_0} + \beta_{z_0}^T \Omega^{-1} \beta_{z_0}, \end{pmatrix},$$

$\alpha_{z_0}, \beta_{z_0} \in \mathbb{R}^p, \sigma_{z_0} \in \mathbb{R}^1$ are determined by some Riesz representer,

$$\Omega = E\{I(U)(X - E(X|Z))^{\otimes 2}\}.$$

Discussions on Theorem 1

- The key technical tool we develop is the Joint Bahadur Representation, which is built upon the concentration inequality for the following empirical processes

$$\mathbb{G}_n(H) \equiv \sqrt{n}(\mathbb{P}_n - P)\{\psi_n(T, H)R_U\},$$

where ψ_n is some real-valued Lipschitz continuous function.

Discussions on Theorem 1

- The key technical tool we develop is the Joint Bahadur Representation, which is built upon the concentration inequality for the following empirical processes

$$\mathbb{G}_n(H) \equiv \sqrt{n}(\mathbb{P}_n - P)\{\psi_n(T, H)R_U\},$$

  where ψ_n is some real-valued Lipschitz continuous function.
- In Theorem 1, $\widehat{H}_{n,\lambda} = (\widehat{\theta}'_{n,\lambda}, \widehat{f}_{n,\lambda})'$ is centered around the biased center $H_0^* \neq H_0 = (\theta_0', f_0)'$. A natural question to ask is how one can remove the asymptotic estimation bias, i.e., $P_\lambda H_0$, partly or completely.

Main Results: Theorem 2 (Joint Asymptotics II)

Under Conditions in Theorem 1, e.g., $\lambda \asymp n^{-2m/(2m+1)}$, and some additional smoothness on $E(X_k|Z)$'s, we have, for any $z_0 \in [0,1]$,

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \\ \sqrt{n\lambda^{1/2m}}(\widehat{f}_{n,\lambda}(z_0) - f_0(z_0)) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ b_{z_0} \end{pmatrix}, \Psi \right), \quad (1)$$

where

$$\Psi = \begin{pmatrix} \Omega^{-1} & 0 \\ 0 & \sigma_{z_0}^2 \end{pmatrix} \text{ and } \lim_{n\to\infty} \sqrt{n\lambda^{1/2m}}(W_\lambda f_0)(z_0) = -b_{z_0}.$$

The above Theorem says that

Main Results: Theorem 2 (Joint Asymptotics II)

Under Conditions in Theorem 1, e.g., $\lambda \asymp n^{-2m/(2m+1)}$, and some additional smoothness on $E(X_k|Z)$'s, we have, for any $z_0 \in [0,1]$,

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \\ \sqrt{n\lambda^{1/2m}}(\widehat{f}_{n,\lambda}(z_0) - f_0(z_0)) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ b_{z_0} \end{pmatrix}, \Psi \right), \quad (1)$$

where

$$\Psi = \begin{pmatrix} \Omega^{-1} & 0 \\ 0 & \sigma_{z_0}^2 \end{pmatrix} \text{ and } \lim_{n\to\infty} \sqrt{n\lambda^{1/2m}}(W_\lambda f_0)(z_0) = -b_{z_0}.$$

The above Theorem says that

- Under additional smoothness condition, we can completely remove the estimation bias for θ but only partly for f;

Main Results: Theorem 2 (Joint Asymptotics II)

Under Conditions in Theorem 1, e.g., $\lambda \asymp n^{-2m/(2m+1)}$, and some additional smoothness on $E(X_k|Z)$'s, we have, for any $z_0 \in [0,1]$,

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \\ \sqrt{n\lambda^{1/2m}}(\widehat{f}_{n,\lambda}(z_0) - f_0(z_0)) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ b_{z_0} \end{pmatrix}, \Psi\right), \quad (1)$$

where

$$\Psi = \begin{pmatrix} \Omega^{-1} & 0 \\ 0 & \sigma^2_{z_0} \end{pmatrix} \text{ and } \lim_{n\to\infty} \sqrt{n\lambda^{1/2m}}(W_\lambda f_0)(z_0) = -b_{z_0}.$$

The above Theorem says that

- Under additional smoothness condition, we can completely remove the estimation bias for θ but only partly for f;
- More interestingly, we have achieved the asymptotic independence between $\widehat{\theta}_{n,\lambda}$ and $\widehat{f}_{n,\lambda}(z_0)$ at the same time (existence of deeper theory?).

Discussions on Theorem 2

Theorem 2 implies that our parametric limit distribution, i.e.,

$$\sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \overset{d}{\to} N(0, \Omega^{-1})$$

is exactly the same as that obtained in Mammen and van de Geer (1997). And $\widehat{\theta}_{n,\lambda}$ is semiparametric efficient under some model assumptions;

Discussions on Theorem 2

- However, our (point-wise) nonparametric limit distribution, i.e.,

$$\sqrt{n\lambda^{1/2m}}(\widehat{f}_{n,\lambda}(z_0) - f_0(z_0)) \xrightarrow{d} N(b_{z_0}, \sigma_{z_0}^2)$$

is in general different from that obtained in the nonparametric smoothing spline setup (Shang and Cheng, 2012) in terms of different values of b_{z_0} and $\sigma_{z_0}^2$.

Discussions on Theorem 2

- However, our (point-wise) nonparametric limit distribution, i.e.,
$$\sqrt{n\lambda^{1/2m}}(\widehat{f}_{n,\lambda}(z_0) - f_0(z_0)) \xrightarrow{d} N(b_{z_0}, \sigma^2_{z_0})$$
is in general different from that obtained in the nonparametric smoothing spline setup (Shang and Cheng, 2012) in terms of different values of b_{z_0} and $\sigma^2_{z_0}$.
- This is due to the different eigen-systems in the semi-nonparametric and nonparametric contexts.

Discussions on Theorem 2

- However, our (point-wise) nonparametric limit distribution, i.e.,

$$\sqrt{n\lambda^{1/2m}}(\widehat{f}_{n,\lambda}(z_0) - f_0(z_0)) \overset{d}{\to} N(b_{z_0}, \sigma^2_{z_0})$$

  is in general different from that obtained in the nonparametric smoothing spline setup (Shang and Cheng, 2012) in terms of different values of b_{z_0} and $\sigma^2_{z_0}$.
- This is due to the different eigen-systems in the semi-nonparametric and nonparametric contexts.
- Therefore, we want to emphasize that the common intuition that "the (point-wise) asymptotic inferences for the nonparametric component is not affected by the inclusion of a faster convergent parametric estimate" is in general wrong (unless in the special least square estimation).

Motivations for Theorem 3

To further illustrate Theorem 2, we consider the penalized least square estimation in the partly linear model, i.e., partial smoothing spline model. In particular, we give the explicit expressions for the asymptotic estimation bias b_{z_0} and asymptotic covariance Ψ. In addition, we give smoothing parameter conditions under which the remaining nonparametric estimation bias is removed as well.

Main Results: Theorem 3 (L_2 regression)

Let $\ell(y;a) = -(y-a)^2/2$, $f_0 \in S^{2m}$ and $m > 1+\sqrt{3}/2 \approx 1.866$.
(i) If $\lambda/n^{-2m/(4m+1)} \to c > 0$, then we have, for any $z_0 \in [0,1]$,

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \\ n^{\frac{2m}{(4m+1)}}(\widehat{f}_{n,\lambda}(z_0) - f_0(z_0)) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ \frac{(-1)^{m-1}c^{2m}f_0^{(2m)}(z_0)}{\pi(z_0)} \end{pmatrix}, \Psi \right).$$

(ii) If $\lambda \asymp n^{-d}$ for $2m/(4m+1) < d \leq 4m^2/(10m-1)$ (under-smoothing condition), then we have, for any $z_0 \in [0,1]$,

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \\ \sqrt{n\lambda^{1/2m}}(\widehat{f}_{n,\lambda}(z_0) - f_0(z_0)) \end{pmatrix} \xrightarrow{d} N(0, \Psi).$$

(iii) The above Ψ has an explicit expression:

$$\Psi = \begin{pmatrix} \{E[X - E(X|Z)]^{\otimes 2}\}^{-1} & 0 \\ 0 & \frac{\int_0^\infty (1+x^{2m})^{-2}dx}{\pi} \end{pmatrix}.$$

Joint Inferences: Prediction/Confidence Interval

- Consider the prediction interval for the new response Y_{new} given the future data (x_0, z_0).

Joint Inferences: Prediction/Confidence Interval

- Consider the prediction interval for the new response Y_{new} given the future data (x_0, z_0).
- The above prediction interval makes sense mostly for the continuous response. As for the discrete response, it is more reasonable to construct the confidence interval for $\rho_0 = \rho(\theta_0, f_0(z_0)) \in \mathbb{R}^1$ (generally speaking).

Joint Inferences: Prediction/Confidence Interval

- Consider the prediction interval for the new response Y_{new} given the future data (x_0, z_0).
- The above prediction interval makes sense mostly for the continuous response. As for the discrete response, it is more reasonable to construct the confidence interval for $\rho_0 = \rho(\theta_0, f_0(z_0)) \in \mathbb{R}^1$ (generally speaking).
- For example, in the semiparametric logistic regression,

$$\rho_0 = P(Y = 1 | X = x, Z = z_0) = \frac{\exp(\theta_0' x + f_0(z_0))}{1 + \exp(\theta_0' x + f_0(z_0))}.$$

Joint Inferences: Prediction/Confidence Interval

- Consider the prediction interval for the new response Y_{new} given the future data (x_0, z_0).
- The above prediction interval makes sense mostly for the continuous response. As for the discrete response, it is more reasonable to construct the confidence interval for $\rho_0 = \rho(\theta_0, f_0(z_0)) \in \mathbb{R}^1$ (generally speaking).
- For example, in the semiparametric logistic regression,

$$\rho_0 = P(Y = 1 | X = x, Z = z_0) = \frac{\exp(\theta_0' x + f_0(z_0))}{1 + \exp(\theta_0' x + f_0(z_0))}.$$

- Based on the joint asymptotic normality result in Theorem 2, we are able to construct P.I. or C.I.. However, we have to estimate either $\sigma_{z_0}^2$ or Ω^{-1} and also the error variance. This motivates the likelihood ratio testing.

Joint Inferences: Local Likelihood Ratio Testing

- $H_0 : \theta = \theta_0$ and $f(z_0) = w_0$ v.s. $H_A : \theta \neq \theta_0$ or $f(z_0) \neq w_0$;

Joint Inferences: Local Likelihood Ratio Testing

- $H_0 : \theta = \theta_0$ and $f(z_0) = w_0$ v.s. $H_A : \theta \neq \theta_0$ or $f(z_0) \neq w_0$;
- LRT test statistic is defined as

$$LRT_{n,\lambda} = \ell_{n,\lambda}(\widehat{H}^0_{n,\lambda}) - \ell_{n,\lambda}(\widehat{H}_{n,\lambda}),$$

where $\widehat{H}^0_{n,\lambda}$ is the maximizer under H_0.

Joint Inferences: Local Likelihood Ratio Testing

- $H_0 : \theta = \theta_0$ and $f(z_0) = w_0$ v.s. $H_A : \theta \neq \theta_0$ or $f(z_0) \neq w_0$;
- LRT test statistic is defined as

$$LRT_{n,\lambda} = \ell_{n,\lambda}(\widehat{H}^0_{n,\lambda}) - \ell_{n,\lambda}(\widehat{H}_{n,\lambda}),$$

  where $\widehat{H}^0_{n,\lambda}$ is the maximizer under H_0.
- Under H_0, we have

$$2n \cdot LRT_{n,\lambda} \xrightarrow{d} D_1 + c_0 D_2,$$

  where $D_1 \sim \chi^2_p$ (parametric effect), $D_2 \sim \chi^2_1$ (nonparametric effect), and D_1 is independent of D_2. Here, c_0 is uniquely determined by the underlying eigen-system.

Global Result I: Global Likelihood Ratio testing

As for the global testing $H_0 : \theta = \theta_0$ and $f = f_0$, we summarize our interesting results below:

Global Result I: Global Likelihood Ratio testing

As for the global testing $H_0 : \theta = \theta_0$ and $f = f_0$, we summarize our interesting results below:

- Under H_0, we prove that

$$-2nr_K \cdot LRT^G_{n,\lambda} \overset{a}{\sim} \chi^2_{u_n}, \tag{2}$$

where the scaling r_K and the degree of freedom $u_n \to \infty$ are uniquely determined by the kernel function.

Global Result I: Global Likelihood Ratio testing

As for the global testing $H_0 : \theta = \theta_0$ and $f = f_0$, we summarize our interesting results below:

- Under H_0, we prove that

$$-2nr_K \cdot LRT_{n,\lambda}^{G} \overset{a}{\sim} \chi^2_{u_n}, \tag{2}$$

  where the scaling r_K and the degree of freedom $u_n \to \infty$ are uniquely determined by the kernel function.
- The above limit distribution is not jointly determined by θ and f as in the local likelihood ratio theorem, but only by the effect from f since we are considering a global testing now.

Global Result I: Global Likelihood Ratio testing

As for the global testing $H_0 : \theta = \theta_0$ and $f = f_0$, we summarize our interesting results below:

- Under H_0, we prove that

$$-2nr_K \cdot LRT^G_{n,\lambda} \overset{a}{\sim} \chi^2_{u_n}, \tag{2}$$

  where the scaling r_K and the degree of freedom $u_n \to \infty$ are uniquely determined by the kernel function.
- The above limit distribution is not jointly determined by θ and f as in the local likelihood ratio theorem, but only by the effect from f since we are considering a global testing now.
- The above results can be generalized to the more useful composite hypothesis such as $H_0 : \theta = \theta_0$ and f belongs to the class of q-degree polynomials

Global Result II: Simultaneous Confidence Band

- We derived a simultaneous confidence band (w.r.t. uniform norm) for the estimate $\widehat{f}_{n,\lambda}$, in the presence of the parametric component $\widehat{\theta}_{n,\lambda}$. We find that our simultaneous confidence band is in the same fashion of the one derived by Bickel and Rosenblatt (1973) in the purely nonparametric setup.

Global Result II: Simultaneous Confidence Band

- We derived a simultaneous confidence band (w.r.t. uniform norm) for the estimate $\widehat{f}_{n,\lambda}$, in the presence of the parametric component $\widehat{\theta}_{n,\lambda}$. We find that our simultaneous confidence band is in the same fashion of the one derived by Bickel and Rosenblatt (1973) in the purely nonparametric setup.
- Our confidence band (i) does not require the symmetric error distribution as in the volume of tube method (Sun and Loader 1994); and (ii) applies to the general quasi-likelihood models.

Examples: Partial Smoothing Spline

- Consider the penalized least square estimation of the partly linear models: $Y = X^T\theta_0 + f_0(Z) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.

Examples: Partial Smoothing Spline

- Consider the penalized least square estimation of the partly linear models: $Y = X^T\theta_0 + f_0(Z) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.
- We choose the eigenfunctions as

$$h_\mu(z) = \left\{ \begin{array}{cc} \sigma, & \mu = 0, \\ \sqrt{2}\sigma\cos(2\pi kz), & \mu = 2k, k = 1, 2, \ldots, \\ \sqrt{2}\sigma\sin(2\pi kz), & \mu = 2k-1, k = 1, 2, \ldots, \end{array} \right\}$$

and the eigenvalue: $\gamma_{2k} = \gamma_{2k-1} = \sigma^2(2\pi k)^{2m}$ for $k \geq 1$ and $\gamma_0 = 0$, i.e., trigonometric eigen-system.

Examples: Partial Smoothing Spline

- Consider the penalized least square estimation of the partly linear models: $Y = X^T\theta_0 + f_0(Z) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.
- We choose the eigenfunctions as

$$h_\mu(z) = \left\{ \begin{array}{cc} \sigma, & \mu = 0, \\ \sqrt{2}\sigma\cos(2\pi kz), & \mu = 2k, k = 1, 2, \ldots, \\ \sqrt{2}\sigma\sin(2\pi kz), & \mu = 2k-1, k = 1, 2, \ldots, \end{array} \right\}$$

  and the eigenvalue: $\gamma_{2k} = \gamma_{2k-1} = \sigma^2(2\pi k)^{2m}$ for $k \geq 1$ and $\gamma_0 = 0$, i.e., trigonometric eigen-system.
- We have the explicit value: $c_0 = 0.75$ (0.83) for $m = 2$ (3).

Examples: Partly Linear Logistic Regression

- Consider the binary response $Y \in \{0, 1\}$ modelled by

$$\Pr(Y = 1|X = x, Z = z) = \frac{\exp(x^T\theta_0 + f_0(z))}{1 + \exp(x^T\theta_0 + f_0(z))}.$$

Examples: Partly Linear Logistic Regression

- Consider the binary response $Y \in \{0, 1\}$ modelled by

$$\Pr(Y = 1 | X = x, Z = z) = \frac{\exp(x^T \theta_0 + f_0(z))}{1 + \exp(x^T \theta_0 + f_0(z))}.$$

- In this example, c_0 has no explicit form, and needs to be estimated as follows

$$\widehat{c}_0 = \lim_{\lambda \to 0} \frac{\sum_\nu \frac{|\widehat{h}_\nu(z)|^2}{(1+\lambda\widehat{\gamma}_\nu)^2}}{\sum_\nu \frac{|\widehat{h}_\nu(z)|^2}{1+\lambda\widehat{\gamma}_\nu}}.$$

Examples: Partly Linear Logistic Regression

- Consider the binary response $Y \in \{0,1\}$ modelled by

$$\Pr(Y=1|X=x,Z=z) = \frac{\exp(x^T\theta_0 + f_0(z))}{1+\exp(x^T\theta_0 + f_0(z))}.$$

- In this example, c_0 has no explicit form, and needs to be estimated as follows

$$\widehat{c}_0 = \lim_{\lambda\to 0} \frac{\sum_\nu \frac{|\widehat{h}_\nu(z)|^2}{(1+\lambda\widehat{\gamma}_\nu)^2}}{\sum_\nu \frac{|\widehat{h}_\nu(z)|^2}{1+\lambda\widehat{\gamma}_\nu}}.$$

- The estimated $\widehat{\lambda}_\nu$ and $\widehat{h}_\nu$ solves from some estimated ODE based on $\widehat{\pi}$ and $E\{\widehat{I(U)|Z}\}$.

Local Polynomial Estimation

Different iterative estimation procedure:

- Estimating θ:

$$\widehat{\theta}_L = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i; X_i^T \theta + \widehat{f}_\theta(Z_i)),$$

where $\widehat{f}_\theta$ is defined below;

Local Polynomial Estimation

Different iterative estimation procedure:

- Estimating θ:

$$\widehat{\theta}_L = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i; X_i^T\theta + \widehat{f}_\theta(Z_i)),$$

  where $\widehat{f}_\theta$ is defined below;

- Estimating f: given any fixed θ

$$\widehat{\eta}_\theta = arg\min_{\eta} \sum_{i=1}^{n} \ell(Y_i; X_i^T\theta + Z_i^T(z)\eta(z))K_h(Z_i - z),$$

  where $Z^T(z)\eta(z)$ is the local polynomial approximation of $f(z)$ and $K_h(\cdot) = K(\cdot/h)/h$. The first element of $\widehat{\eta}_\theta$ is $\widehat{f}_\theta$.

Local Polynomial Estimation

Different iterative estimation procedure:

- Estimating θ:

$$\widehat{\theta}_L = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i; X_i^T \theta + \widehat{f}_\theta(Z_i)),$$

  where $\widehat{f}_\theta$ is defined below;
- Estimating f: given any fixed θ

$$\widehat{\eta}_\theta = arg \min_{\eta} \sum_{i=1}^{n} \ell(Y_i; X_i^T \theta + Z_i^T(z)\eta(z)) K_h(Z_i - z),$$

  where $Z^T(z)\eta(z)$ is the local polynomial approximation of $f(z)$ and $K_h(\cdot) = K(\cdot/h)/h$. The first element of $\widehat{\eta}_\theta$ is $\widehat{f}_\theta$.
- All our results for the penalized estimate carry over the joint estimate $(\widehat{\theta}_L, \widehat{f}_L)$, where $\widehat{f}_L = \widehat{f}_{\widehat{\theta}_L}$.

My 2nd Steps...

Semi-Nonparametric estimation without regularization.

My 2nd Steps...

Semi-Nonparametric estimation without regularization.

- For example, we consider the *partly linear isotonic regression*:

$$(\widehat{\theta}, \widehat{\eta}) = \arg\min \sum_{i=1}^{n} (Y_i - X_i'\theta - f(Z_i))^2,$$

where η is assumed to be monotone.

My 2nd Steps...

Semi-Nonparametric estimation without regularization.

- For example, we consider the *partly linear isotonic regression*:

$$(\widehat{\theta}, \widehat{\eta}) = \arg\min \sum_{i=1}^{n} (Y_i - X_i'\theta - f(Z_i))^2,$$

  where η is assumed to be monotone.
- It is well known that (i) $\widehat{\theta}$ is root-n asymptotic normality; (ii) $\widehat{f}(z_0)$ is cubic rate convergent with Chernoff limiting distribution for any fixed $z_0 \in (0,1)$.

My 2nd Steps...

Semi-Nonparametric estimation without regularization.

- For example, we consider the *partly linear isotonic regression*:

$$(\widehat{\theta}, \widehat{\eta}) = \arg\min \sum_{i=1}^{n} (Y_i - X_i'\theta - f(Z_i))^2,$$

  where η is assumed to be monotone.
- It is well known that (i) $\widehat{\theta}$ is root-n asymptotic normality; (ii) $\widehat{f}(z_0)$ is cubic rate convergent with Chernoff limiting distribution for any fixed $z_0 \in (0, 1)$.
- Interesting question: What will be the joint limit distribution for $(\sqrt{n}(\widehat{\theta} - \theta_0), n^{1/3}(\widehat{f}(z_0) - f_0(z_0)))$? (mixture of regular and irregular asymptotics...)

Please join me in the long march....

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