Semiparametric Efficiency and Optimal Estimation, with Application to Auxiliary Outcome Problems

Jinbo Chen, Ph.D.
Biostatistics Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute, Rockville, MD, USA

Norman Breslow, Ph.D.
Department of Biostatistics
University of Washington, Seattle, WA, USA

Outline

- Semiparametric efficiency and Godambe's optimality: Connection
- Application: the auxiliary outcome problem with the conditional mean model
  - MRC Cognitive Function and Aging Study
  - The efficient estimation
- Summary and Conclusions

Semiparametric Model and Estimation

Model: $P_{\theta,\eta}$ θ: finite dimensional

η: infinite dimensional.

- The conditional mean model:

$$\mathrm{E}(Y \mid X) = \mu(X;\theta)$$

- The Likelihood function:

$$f[\varepsilon \mid X]g(X), \quad \varepsilon = Y - \mu(X;\theta)$$

- The nuisance parameters η=(f, g):

$$\{ f(\varepsilon \mid X): \int \varepsilon f(\varepsilon \mid x) d\varepsilon = 0,\ x \in X \}$$

Estimation for the Semiparametric Model

Goal: estimation of θ

- Semiparametric efficient estimation (SEE): finding a $\sqrt{n}$ consistent estimator achieving the efficiency bound

- Heuristic approach: solve an (optimal) estimating equation

  - Optimal estimating function theory (Godambe, 1960)

Motivation: Optimal Estimation and Semiparametric Efficiency

The conditional mean model:

- A Class of linear estimating functions

$$\{\sum_i h(X_i;\theta)[Y_i - \mu(X_i;\theta)], h \in \mathrm{H}\}$$

- Optimal member /quasi-score (McCullagh and Nelder, 1993)

$$h^*(X)\varepsilon = \frac{\dot{\mu}(X;\theta)}{\mathrm{Var}(Y \mid X)}\varepsilon$$

- $h^*(X)\varepsilon$ is the efficient score (Chamberlain, 1987)

Optimal Estimating Function Theory of Godambe and Heyde

- A class of estimating functions indexed by H={*h(Z; θ)*}:

$$G = \{G(h;Z,\theta) : h \in \mathrm{H}\}$$

- Regular unbiased estimating function G(Z;θ)

$$\mathrm{E}\ G(Z;\theta) = 0;$$

$$\mathrm{E}\ [\partial G(Z;\theta)/\partial\theta] \text{ is nonsingular;}$$

$$\mathrm{E}\ [G(Z;\theta)G'(Z;\theta)] \text{ is nonsingular.}$$

Criterion for Finding the Optimal Member G*

- Optimality criterion:

  Let $\varepsilon(G) = (\mathrm{E}\,\dot{G})^{-1}\mathrm{E}GG'(\mathrm{E}\,\dot{G}')^{-1}$. G^* is optimal if $\varepsilon(\mathrm{G}^*)$-$\varepsilon(\mathrm{G})$ is nonpositive definite $\forall G \in \mathrm{G}$, $\theta \in \Theta$ and $\eta \in \Gamma$.

- Solve G=0 for $\hat{\theta}$:

  $$\hat{\theta}\text{-}\theta \xrightarrow{\mathrm{D}} [0, (\mathrm{E}\,\dot{G})^{-1}\mathrm{E}GG'(\mathrm{E}\,\dot{G}')^{-1}]$$

  where $\dot{G} = \partial G / \partial \theta$

Optimality Criterion: Continued

- Theorem (Corollary to Theorem 2.1 in Heyde, 1997)

  $G^* \in \mathcal{G}$ is a quasi-score estimating function if and only if

$$-\mathrm{E}\dot{G} = \mathrm{E}GG^{*'} \quad \forall G \in \mathcal{G}.$$

- Geometric Interpretation:

$$-\mathrm{E}\dot{G} = \mathrm{E}GG^{*'} \iff \mathrm{E}[G(\dot{l}_\theta - G^{*'})] = 0$$

Semiparametric Efficient Estimation
(BKRW, 1993)

- Notation:

  l_θ^* : efficient score function

  $[El_\theta^* l_\theta^*]^{-1}$: semiparametric efficiency bound

  $\tilde{l}_\theta = [El_\theta^* l_\theta^*]^{-1} l_\theta^*$: efficient influence function

  Data: $\{Z_i, i = 1, \ldots n\}$ $i.i.d$

- Regular and asymptotically linear (RAL) estimator:

  regular and $\hat{\theta} = \theta + \frac{1}{n}\sum \varphi(Z_i;\theta,\eta) + o_p(n^{1/2})$

  with $E\varphi(Z_i;\theta,\eta) = 0$, $\mathrm{var}\,\varphi(Z_i;\theta,\eta) < \infty$.

SEE: Continued

- Semiparametric efficient estimator:

$$\hat{\theta} = \theta + \frac{1}{n}\sum \tilde{l}_\theta(Z_i) + o_p(n^{1/2})$$

- $\operatorname{var}(\tilde{l}_\theta) \le \operatorname{var}(\varphi)$

  $\Rightarrow$ The efficient estimator is the most precise RAL.

  $\Rightarrow$ Sufficient to work with the class of RAL estimators to find the efficient estimator.

  $\Rightarrow$ find $\tilde{l}_\theta$ (l_θ^*)

  $\Rightarrow$ construct efficient estimators

- Influence function $\Leftrightarrow$ Regular estimating function

SEE and Optimal Estimation: Connection (Chen 2002, Ph.D dissertation)

- A corollary to Theorem 3.3.1, BKRW

  Let G $= \{G\}$ be the closed linear span of the influence functions for all RAL estimators. Then l_θ^* is uniquely identified by $l_\theta^* \in$ G and

  $$\mathrm{E}[G(\dot{l}_\theta - l_\theta^*)] = 0 \ \forall G \in \mathrm{G}.$$

- Recall Godambe's optimality criterion

  $$-\mathrm{E}\dot{G} = \mathrm{E}GG^{*'} \iff \mathrm{E}[G(\dot{l}_\theta - G^{*'})] = 0$$

Connection: Continued

- The efficient score/influence function is the Godambe's optimal member in the closed linear span of influence functions for all RAL estimators!

- Useful to calculating the efficient score functions for the missing data problem (Robins, Rotnitzky and Zhao, 1994; RRZ)

- Maybe useful to obtain the efficient score function for non-iid data (Welfelmayer 1996)
  - Trial and verification approach

An Example

- The quasi-likelihood model:

$$E(Y \mid X) = \mu(X;\theta); \ \mathrm{var}(Y \mid X) = \phi v(\mu).$$

ϕ and v are known.

- The Likelihood function:

$$f[\varepsilon_s \mid x] g(X) \text{ where } \varepsilon_s = (Y-\mu)/(\phi v)^{1/2}$$

- The nuisance parameters η=(f, g):

$$\{ f(\varepsilon_s \mid X): \ \varepsilon_s f(\varepsilon_s \mid x) d\varepsilon_s = 0,$$
$$\varepsilon_s{}^2 f(\varepsilon_s \mid x) d\varepsilon_s = 1; x \in X \}$$

Example: Continued

- A class of linear estimating functions

$$\{G_1(h;\theta):\ G_1(h;\theta) = \sum_i h(X_i;\theta)[Y_i - \mu(X_i;\theta)], h \in \mathrm{H}\}$$

$$\text{Optimal member: } \mathrm{G}_1^* = \sum_i \dot{\mu}_i \ v^{-1}(\mu_i)\ [Y_i - \mu(X_i;\theta)]$$

- A class of quadratic estimating functions

$$\{G_2(h_1,h_2;\theta):\ G_2 = \sum_i h_1(X_i;\theta)\varepsilon_i + h_2(X_i;\theta)(\varepsilon_i^2 - \phi v_i)\}$$

Optimal member G_2^*: involves $\mathrm{E}(\varepsilon^3 \mid X)$ and $\mathrm{E}(\varepsilon^4 \mid X)$.

An Example: Continued

- $[\mathrm{var}(G_1^{\,*})]^{-1} \geq [\mathrm{var}(G_2^{\,*})]^{-1}$.
- $\{G_2\}$ is the complete class of influence functions for all RAL estimators
- $G_2^{\,*}$ is the efficient score function
- The same efficient score was obtained by Rotnitzky and Robins (1995) using a different approach

Outline

- Semiparametric efficiency and Godambe's optimality: Connection

- Application: the auxiliary outcome problem with the conditional mean model (Chen and Breslow, Canadian J. Statistics, to appear)
  - MRC Cognitive Function and Aging Study
  - The efficient estimation

- Summary and Conclusions

MRC Cognitive Function and Aging Study (Clayton *et al.*, JRSS(B), 1999)

- Goal: Estimate prevalence of dementia
  - by sex and age (X = covariates)
- Outcome: dementia assessed by
  - Geriatric mental state exam (Y) – gold standard
  - Mini-mental state exam (S) – auxiliary outcome
- Design: two phase sampling
  - 10,000 main sample known (X,S)
  - 1,780 validation sample known (Y,X,S)
  - Data simulated by Clayton *et al.*

Study Design

Age (yrs)	MMSE	No. of subjects Main	Validation	Sampling fraction
	0-21	291	291	1.000
65-74	22-25	950	220	0.232
	26-30	3,759	386	0.103
	0-21	1,037	496	0.478
75+	22-25	1,486	208	0.140
	26-30	2,477	179	0.072
Total		10,000	1,780	0.178

Notation

- Y = true outcome
- S = auxiliary (surrogate) outcome
- X = covariates
- R = sampling indicator (R=1→validation)
- Observed data

$$Z = (S, X, RY, R)\begin{cases} R{=}1\text{: } (S, X, Y) \\ R{=}0\text{: } (S, X) \end{cases}$$

- $\{Z_i,\ i = 1, \ldots, n\}$ i.i.d.

Conditional Mean Model

- Conditional mean (parametric)

$$\mathrm{E}(Y \mid X = x) = \mu(x;\theta), \quad \theta \in R^p$$

- Joint distribution

$$p(y,s,x) = q(s \mid y,x)\, f\left(y - \mu(x;\theta)\right) g(x)$$

  - $\eta = (q, f, g)$ (infinite dim.) "nuisance" parm

    finite variance, $\int f(\varepsilon)\mathrm{d}\varepsilon = 0$

Validation Sampling: MAR

- Missing at random:

$$\Pr(R=1 \mid y,s,x)=\Pr(R=1 \mid s,x)=\pi(s,x)$$

- Positive validation probability:

$$\pi(s,x) \geq \sigma > 0$$

- Parametric missingness model:

$$\pi(s,x)=\pi(s,x;\alpha),\ \alpha \in R^q$$

Previous Work

- Semiparametric efficient estimators of θ
  - Robins, Rotnitzky, Zhao (*JASA*, 1994)
  - Rotnitzky, Robins (*Scand J Statist*, 1995)
  - Holcroft, Rotnitzky, Robins (*J Stat Plan Inf,*1997)
  - …..

- Inefficient estimators of θ
  - Pepe, Reilly, Fleming (*J Stat Plan Inf,*1994)
    - "Mean score" (Horwitz-Thompson) estimator
  - Y. Chen (*Biometrika*, 2000)
    - "Robust imputation" estimator

Rationale for More Work

- "Simple" derivation of SEE via connection to Godambe optimality
- Method may generalize to more complex problems
- Explicit expression for efficient estimating equation
- Useful pedagogical example

Missing Data Problem (RRZ, 1994)

- Space Γ of influence functions for RAL est

$$G(h;Z) = \frac{R}{\pi} h(X)\varepsilon - \frac{R-\pi}{\pi} h(X)\mathrm{E}(\varepsilon \mid S, X)$$

- Godambe optimal quasi score is $G^* = h^*(X)\varepsilon^*$

$$\varepsilon^* = \frac{R}{\pi} Y - \frac{R-\pi}{\pi} \mathrm{E}(Y \mid S, X) - \mu(X;\theta)$$

$$h^* = \dot{\mu}(X;\theta)\mathrm{Var}^{-1}(\varepsilon^* \mid X)$$

- Estimate $\mathrm{E}(Y \mid S, X)$, $\mathrm{Var}(\varepsilon^* \mid X)$ and π from data

Semiparametric Efficiency Bound

- Efficiency bound is $\mathrm{Var}^{-1}(G^*)$ where

$$\mathrm{Var}(G^*) = \mathrm{E}\left\{\dot{\mu}(X;\theta)\left[\mathrm{Var}(\varepsilon \mid X) + A(X)\right]^{-1}\dot{\mu}^{\mathrm{T}}(X;\theta)\right\}$$

$$A(X) = \mathrm{E}\left(\frac{1-\pi}{\pi}\mathrm{Var}(\varepsilon \mid S, X)\middle| X\right)$$

  - Perfect information: $S = Y \Rightarrow A(X) = 0$

    -Usual quasi likelihood bound

  - No information: $S \perp Y \mid X \Rightarrow$ Validation only

$$\mathrm{Var}\, G^* = \mathrm{E}\left[\dot{\mu}(X;\theta)\mathrm{E}^{-1}\left(\frac{1}{\pi}\middle| X\right)\mathrm{Var}^{-1}(\varepsilon \mid X)\dot{\mu}^{\mathrm{T}}(X;\theta)\right]$$

Summary and Conclusions

- SEE and optimal estimating equation theory
  - Connection (Chen 2002, Dissertation)
- The Auxiliary Outcome Problem (Chen and Breslow, Canadian J. Statistics):
  - Efficient score has a simple closed form
- Extensions of Chen's "robust imputation" method in progress for continuous S and X