Estimation for two-phase designs: semiparametric models and $Z-$theorems

Jon A. Wellner

University of Washington

- joint work with:
  - Norman E. Breslow, University of Washington
  - Takumi Saegusa, University of Washington
- Talk at ICSA- Applied Statistics Symposium, Indianapolis, Indiana, June 21, 2010
- *Email: jaw@stat.washington.edu*
  http: //www.stat.washington.edu/jaw/jaw.research.html

Outline

- Introduction and Review:
  semiparametric models and two-phase designs

Outline

- Introduction and Review:
  semiparametric models and two-phase designs
- More efficiency gains? Some approaches (and difficulties)

Outline

- Introduction and Review:
  semiparametric models and two-phase designs
- More efficiency gains? Some approaches (and difficulties)
- $Z-$theorems and beyond: GMM, MD, EL
  - Generalized Method of Moments (GMM)
  - Minimum distance estimation (MD)
  - Connections with empirical likelihood (EL)

Outline

- Introduction and Review: semiparametric models and two-phase designs
- More efficiency gains? Some approaches (and difficulties)
- $Z-$theorems and beyond: GMM, MD, EL
  - Generalized Method of Moments (GMM)
  - Minimum distance estimation (MD)
  - Connections with empirical likelihood (EL)
- Summary; problems and open questions

Outline

- Introduction and Review:
  semiparametric models and two-phase designs
- More efficiency gains? Some approaches (and difficulties)
- $Z-$theorems and beyond: GMM, MD, EL
  - Generalized Method of Moments (GMM)
  - Minimum distance estimation (MD)
  - Connections with empirical likelihood (EL)
- Summary; problems and open questions

1. Introduction and Review:

- **Model:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
  $(\theta, \eta) \in \Theta \times H$
  - parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
  - nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space.

1. Introduction and Review:

- **Model:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
  $(\theta, \eta) \in \Theta \times H$
  • parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
  • nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space.
- **Assumptions:**

1. Introduction and Review:

- **Model:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
  $(\theta, \eta) \in \Theta \times H$
  • parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
  • nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space.
- **Assumptions:**
- To guarantee $\sqrt{n}$ consistent, asymptotically Gaussian ML estimation of both θ and η under i.i.d. random sampling, i.e. with complete data:

1. Introduction and Review:

- **Model:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
  $(\theta, \eta) \in \Theta \times H$
  • parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
  • nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space.
- **Assumptions:**
- To guarantee $\sqrt{n}$ consistent, asymptotically Gaussian ML estimation of both θ and η under i.i.d. random sampling, i.e. with complete data:
  1. Scores $\dot{\ell}_{\theta,\eta}$ and $B_{\theta,\eta}h$, $h \in \mathcal{H} \subset \mathcal{B}$, in Donsker class $\mathcal{F}$

1. Introduction and Review:

- **Model:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
  $(\theta, \eta) \in \Theta \times H$
  • parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
  • nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space.
- **Assumptions:**
- To guarantee $\sqrt{n}$ consistent, asymptotically Gaussian ML estimation of both θ and η under i.i.d. random sampling, i.e. with complete data:
  1. Scores $\dot{\ell}_{\theta,\eta}$ and $B_{\theta,\eta}h$, $h \in \mathcal{H} \subset \mathcal{B}$, in Donsker class $\mathcal{F}$
  2. Scores $L_2(P_0)$-continuous at (θ_0, η_0)

1. Introduction and Review:

- **Model:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
  $(\theta, \eta) \in \Theta \times H$
  • parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
  • nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space.
- **Assumptions:**
- To guarantee $\sqrt{n}$ consistent, asymptotically Gaussian ML estimation of both θ and η under i.i.d. random sampling, i.e. with complete data:
  1. Scores $\dot{\ell}_{\theta,\eta}$ and $B_{\theta,\eta}h$, $h \in \mathcal{H} \subset \mathcal{B}$, in Donsker class $\mathcal{F}$
  2. Scores $L_2(P_0)$-continuous at (θ_0, η_0)
  3. “Information operator" $B_0^* B_0$ continuously invertible

1. Introduction and Review:

- **Model:** semiparametric model, $X \sim P_{\theta,\eta} \in \mathcal{P}$
  $(\theta, \eta) \in \Theta \times H$
  • parametric part: $\theta \in \Theta \subset \mathbb{R}^d$
  • nonparametric part: $\eta \in H \subset \mathcal{B}$, a Banach space.
- **Assumptions:**
- To guarantee $\sqrt{n}$ consistent, asymptotically Gaussian ML estimation of both θ and η under i.i.d. random sampling, i.e. with complete data:
  1. Scores $\dot{\ell}_{\theta,\eta}$ and $B_{\theta,\eta}h$, $h \in \mathcal{H} \subset \mathcal{B}$, in Donsker class $\mathcal{F}$
  2. Scores $L_2(P_0)$-continuous at (θ_0, η_0)
  3. “Information operator" $B_0^* B_0$ continuously invertible
  4. Solution $(\hat{\theta}_n, \hat{\eta}_n)$ to score equations consistent for (θ_0, η_0)

Example: Cox (proportional hazards) regression

- Z – p-vector of covariates: $Z \sim \ H$
- $\tilde{T}$ – failure time: $[\tilde{T}|Z] \sim$ Cox(θ, Λ)
- C – censoring time: $[C|Z] \sim \ G$
- $X = (\Delta, T, Z)$ where
  - $T := \min(\tilde{T}, C)$ – observed time
  - $\Delta := 1\{\tilde{T} \leq C\}$ indicates failure at T
- Density for $x = (\delta, t, z)$:

$$e^{-e^{z\theta}\Lambda(t)} \left(e^{z\theta}\lambda(t)\,(1 - G(t - |z))\right)^{\delta} (g(t|z))^{1-\delta}\, h(z)$$

- Likelihood considered only for (θ, Λ) whereas $\eta = (\Lambda, G, H)$
  - (G, H) orthogonal parameters (complete data)

Two Phase Stratified Sampling

Problem: X not fully observed for all subjects

Coarsening: $\tilde{X} = \tilde{X}(X)$ observable part of X

Auxiliary: U helps predict inclusion in subsample

- U optional, to improve efficiency

Notation:

- $V = (\tilde{X}, U) \in \mathcal{V}$ observable for all
- $W = (X, V) \in \mathcal{W}$ observable only in validation sample

Phase I: $\{W_1, \ldots, W_n\}$ i.i.d. sample size n

- but observe only $\{V_1, \ldots, V_n\}$

Phase II: Generate sampling indicators $\{\xi_1, \ldots, \xi_n\}$

- observe all of X_i if $\xi_i = 1$

Finite population stratified sampling

Partition $\mathcal{V}$ into J strata $\mathcal{V}_1 \bigcup \cdots \bigcup \mathcal{V}_J$

Phase I: observe $N_j = \sum_{i=1}^{n} \underline{1}(V_i \in \mathcal{V}_j)$ subjects stratum j

Phase II: sample n_j of N_j (without replacement)

- Sampling indicators ξ_{ji} for subject i in stratum j
  - $(\xi_{j1}, \ldots, \xi_{jN_j})$ exchangeable with $\Pr(\xi_{ji} = 1) = \frac{n_j}{N_j}$
  - Vectors $(\xi_{j1}, \ldots, \xi_{jN_j})$ independent $j = 1, \ldots, J$

	Stratum				
	1	2	$\cdots$	J	Total
Phase I	N_1	N_2	$\cdots$	N_J	n
Phase II	n_1	n_2	$\cdots$	n_J	$n.$
Sampling fractions	$\frac{n_1}{N_1}$	$\frac{n_2}{N_2}$	$\cdots$	$\frac{n_J}{N_J}$	$\frac{n.}{n}$

Bernoulli sampling

- Also known as Manski-Lerman sampling
- Observe V_i and independently generate ξ_i with

$$\Pr(\xi = 1|W) = \Pr(\xi = 1|V) \equiv \pi_0(V)$$

- π_0 known sampling function (MAR)
  - Stratified Bernoulli sampling: $\pi_0(V) = p_j$ for $V \in \mathcal{V}_j$
- Preserves i.i.d. structure
- Desirable to *estimate* known π_0 using parametric model (later)

$$\Pr(\xi = 1|V;\alpha) := \pi_\alpha(V)$$

Horovitz-Thompson (or IPW Likelihood) Estimators

- Define Inverse Probability Weighted (IPW) empirical measure:

$$\mathbb{P}_n^{\pi} = \frac{1}{n}\sum_{i=1}^{n} \frac{\xi_i}{\pi_i}\delta_{X_i}, \qquad \delta_x = \text{ Dirac measure at } x$$

$$\pi_i = \begin{cases} \pi_0(V_i) & \text{if Bernoulli sampling} \\ \frac{n_j}{N_j} 1\{V_i \in \mathcal{V}_j\} & \text{if finite pop'ln stratified sampling} \end{cases}$$

Horovitz-Thompson (or IPW Likelihood) Estimators

- Define Inverse Probability Weighted (IPW) empirical measure:

$$\mathbb{P}_n^{\pi} = \frac{1}{n}\sum_{i=1}^{n} \frac{\xi_i}{\pi_i}\delta_{X_i}, \qquad \delta_x = \text{ Dirac measure at } x$$

$$\pi_i = \begin{cases} \pi_0(V_i) & \text{if Bernoulli sampling} \\ \frac{n_j}{N_j} 1\{V_i \in \mathcal{V}_j\} & \text{if finite pop'ln stratified sampling} \end{cases}$$

- Jointly solve the finite - (for θ) and infinite (for η) dimensional equations

$$\begin{aligned} \mathbb{P}_n^{\pi} \dot{l}_\theta &= 0 && \text{in } \mathbb{R}^d \\ \mathbb{P}_n^{\pi} \dot{l}_\eta h &= 0 && \text{for all } h \in \mathcal{H} \end{aligned}$$

Horovitz-Thompson (or IPW Likelihood) Estimators

- Define Inverse Probability Weighted (IPW) empirical measure:

$$\mathbb{P}_n^\pi = \frac{1}{n}\sum_{i=1}^n \frac{\xi_i}{\pi_i}\delta_{X_i}, \qquad \delta_x = \text{ Dirac measure at } x$$

$$\pi_i = \begin{cases} \pi_0(V_i) & \text{if Bernoulli sampling} \\ \frac{n_j}{N_j} 1\{V_i \in \mathcal{V}_j\} & \text{if finite pop'ln stratified sampling} \end{cases}$$

- Jointly solve the finite - (for θ) and infinite (for η) dimensional equations

$$\begin{aligned} \mathbb{P}_n^\pi \dot{l}_\theta &= 0 && \text{in } \mathbb{R}^d \\ \mathbb{P}_n^\pi \dot{l}_\eta h &= 0 && \text{for all } h \in \mathcal{H} \end{aligned}$$

- MLE for complete data solves same equations with $\mathbb{P}_n$ instead of $\mathbb{P}_n^\pi$.

First Main Result:

- $\widehat{\theta}_n$ solving the IPW estimating equations is asymptotically linear

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} \widetilde{l}_{\theta_0}(X_i) + o_p(1) \\ &= \mathbb{G}_n^{\pi}(\widetilde{l}_{\nu_0}) + o_p(1)\end{aligned}$$

where $\widetilde{l}_\theta(x)$ is the semiparametric efficient influence function for θ (complete data)

$$\mathbb{G}_n^{\pi} = \sqrt{n}(\mathbb{P}_n^{\pi} - P).$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathbb{G}_n^\pi(\tilde{\ell}_{\theta_0,\eta_0}) + o_p(1) \rightarrow_d N(0,\Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^J \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^J \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathbb{G}_n^{\pi}(\tilde{\ell}_{\theta_0,\eta_0}) + o_p(1) \rightarrow_d N(0,\Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

- Gain from stratified sampling without replacement is centering of efficient scores

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathbb{G}_n^{\pi}(\tilde{\ell}_{\theta_0,\eta_0}) + o_p(1) \rightarrow_d N(0, \Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

- Gain from stratified sampling without replacement is centering of efficient scores
  - Can reduce variance (considerably) via finite popl'n sampling.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathbb{G}_n^{\pi}(\tilde{\ell}_{\theta_0,\eta_0}) + o_p(1) \rightarrow_d N(0, \Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

- Gain from stratified sampling without replacement is centering of efficient scores
  - Can reduce variance (considerably) via finite popl'n sampling.
  - Select strata via covariates so that $\tilde{\ell}$ has small conditional variances on the strata

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathbb{G}_n^\pi(\tilde{\ell}_{\theta_0,\eta_0}) + o_p(1) \rightarrow_d N(0,\Sigma)$$

- Asymptotic variances under stratified sampling

$$\Sigma = \begin{cases} \tilde{I}^{-1} + \sum_{j=1}^J \nu_j \frac{1-p_j}{p_j} E_j(\tilde{\ell}^{\otimes 2}), & \text{Bernoulli sampling} \\ \tilde{I}^{-1} + \sum_{j=1}^J \nu_j \frac{1-p_j}{p_j} Var_j(\tilde{\ell}), & \text{finite popl'n sampling} \end{cases}$$

- Gain from stratified sampling without replacement is centering of efficient scores
  - Can reduce variance (considerably) via finite popl'n sampling.
  - Select strata via covariates so that $\tilde{\ell}$ has small conditional variances on the strata
  - Alternatively: Bernoulli sampling, but model the selection probabilities $\pi_\alpha(V)$ and estimate the α's
    Apply a new $Z-$theorem with estimated nuisance parameters: Breslow and W (2007,2008)

Key Result (Breslow & Wellner, *SJOS*, 2007-8)

$$
\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta_0) &= \sqrt{n}(\tilde{\theta}_n - \theta_0) + \sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n) \\
&= \sqrt{n}\mathbb{P}_n\tilde{\ell}_0 + \sqrt{n}\left(\mathbb{P}_n^\pi - \mathbb{P}_n\right)\tilde{\ell}_0 + o_p(1) \\
\sqrt{n}(\mathbb{P}_n - P_0) &\rightsquigarrow \mathbb{G} \text{ in } \ell^\infty(\mathcal{F}) \\
\sqrt{n}\left(\mathbb{P}_n^\pi - \mathbb{P}_n\right) &\rightsquigarrow \sum_{j=1}^J \sqrt{\nu_j}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_j \quad \text{a.s.} \\
\text{Var}_{\text{TOT}} &= \text{Var}_{\text{PHS-I}} + \text{Var}_{\text{PHS-II}}
\end{aligned}
$$

- $\tilde{\theta}_n$ is **unobserved** MLE based on complete data
- $\text{Var}_{\text{PHS-II}}$ is **design based**: normalized error in Horvitz-Thompson estimation of unknown finite population total $$\tilde{\ell}_{\text{TOT}} = \sum_{i=1}^n \tilde{\ell}_0(X_i)$$
- Phase I and II contributions asymptotically independent

2. More efficiency gains? Approaches and difficulties

- Information bound for two - phase design is difficult to calculate.
  Solution: Compare to excess over complete data variances.
- Approaches to improving efficiency by reducing the phase II variance: Construct q-vector of **auxiliary** variables Z from observed data $V = (\tilde{X}, U)$. Use Z to estimate or adjust the sampling probabilities π_i:
  - Estimate sampling probabilities via parametric model $\pi_i = \pi(Z_i; \alpha)$ (Robins, Rotnitzky, and Zhao, 1994)
  - Calibration: Deville and Särndal (1992), Lumley (2010).
  - Choose Z to be highly correlated with $\tilde{\ell}_\theta(X; \hat{\theta}, \hat{\eta})$ to improve estimate of $\tilde{\ell}_{\text{TOT}}$

Choice of Auxiliary Variables Z

- **Simplify:** by considering Bernouilli (i.i.d.) sampling.
- **Influence functions:** for calibrated and estimated weights take general form (RRZ, *JASA*, 1994; vdV §25.5.3)

$$\frac{\xi}{\pi_0(V)}\tilde{\ell}_0(X) - \frac{\xi - \pi_0(V)}{\pi_0(V)}\phi(V)$$

Optimal choice for ϕ is $\phi(V) = E(\tilde{\ell}_0|V)$

– which requires knowledge of $(X|V)$.

In fact $\phi_{\mathrm{C}}(V) = QZ(V)$ for calibration and $\phi_{\mathrm{E}}(V) = RZ(V)\pi_0(V)$ for estimation based on auxiliary variables $Z = Z(V)$. (For estimation these must contain the stratum indicators.)

- Breslow, Lumley *et al*, *AJE* **169**:1398-1405, 2009
- Breslow, Lumley *et al*, *SiB* **1**:32-49, 2009

3. $Z-$theorems and beyond: GMM, MD, EL

- Setting for classical Huber (1967) $Z-$theorem:
  - $\theta \in \Theta \subset \mathbb{R}^d$
  - $\Psi_n : \Theta \to \mathbb{R}^d$, random;
  - $\Psi : \Theta \to \mathbb{R}^d$, deterministic; $\Psi(\theta_0) = 0$.

3. $Z-$theorems and beyond: GMM, MD, EL

- Setting for classical Huber (1967) $Z-$theorem:
  - $\theta \in \Theta \subset \mathbb{R}^d$
  - $\Psi_n : \Theta \to \mathbb{R}^d$, random;
  - $\Psi : \Theta \to \mathbb{R}^d$, deterministic; $\Psi(\theta_0) = 0$.
- **Theorem A:** Suppose that $\hat{\theta}_n \to_p \theta_0$, and that:

  A1. $\Psi_n(\hat{\theta}_n) = o_p(n^{-1/2})$

  A2. $\sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0)) \to_d \mathbb{Z} \sim N_d(0, V)$

  A3. Ψ is differentiable at θ_0 with non-singular derivative
  $$\dot{\Psi}_0 = \dot{\Psi}(\theta_0).$$

  A4.
  $|\sqrt{n}(\Psi_n - \Psi)(\hat{\theta}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)| = o_p(1 + \sqrt{n}|\hat{\theta}_n - \theta_0|)$.

  Then
  $$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d -\dot{\Psi}_0^{-1}\mathbb{Z} \sim N_d(0, \dot{\Psi}_0^{-1}V(\dot{\Psi}_0^{-1})^T).$$

- Setting for Hansen '82; Pakes-Pollard '89 finite-dimensional GMM-theorem
  - $\theta \in \Theta \subset \mathbb{R}^d$
  - $\Psi_n : \Theta \to \mathbb{R}^p$, $p \geq d$ random; $\|h\|_2^2 \equiv \sum_{j=1}^p h_j^2$
  - $\Psi : \Theta \to \mathbb{R}^p$, deterministic; $\Psi(\nu_0, \eta_0) = 0$.
- Conditions:

  C0. $\hat{\theta}_n \to_p \theta_0$ in $\mathbb{R}^d$.

  C1. $\|\Psi_n(\hat{\theta}_n)\|_2 = \inf_\theta \|\Psi_n(\theta)\|_2 + o_p(n^{-1/2})$.

  C2. $\sqrt{n}(\Psi_n - \Psi)(\theta_0) \to_d \mathbb{Z} \sim N_d(0, V)$ in $\mathbb{R}^p$

  C3. $\theta \mapsto \Psi(\theta)$ is differentiable wrt θ at θ_0 with $\dot{\Psi}(\theta_0) \equiv \Gamma$ non-singular.

  C4. For every sequence $\delta_n \searrow 0$

$$\sup_{|\theta - \theta_0| \leq \delta_n} \frac{\|\sqrt{n}(\Psi_n - \Psi)(\theta) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)\|}{1 + \sqrt{n}\|\Psi_n(\theta)\| + \sqrt{n}\|\Psi(\theta)\|} = o_p(1).$$

  C5. θ_0 is an interior point of Θ.

- **Theorem B:** (Hansen,1982; Pakes and Pollard, 1989) Suppose that C0 - C5 hold. Then

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta_0) &\to_d -(\Gamma^T\Gamma)^{-1}\Gamma^T\mathbb{Z} \\ &\sim N_d(0, (\Gamma^T\Gamma)^{-1}(\Gamma^T V\Gamma)(\Gamma^T\Gamma)^{-1}).\end{aligned}$$

- Suppose that $A_n(\theta)$ is a sequence of (possibly random) $p \times p$ matrices and that $\|\cdot\|_2^2$ is replaced by $\|A_n(\theta)\Psi_n(\theta)\|_2^2 = \Psi_n(\theta)^T A_n^T A_n \Psi_n(\theta)$ in the above.
- C6. Suppose that $A_n(\theta)$ converges to a nonsingular, nonrandom matrix A:

$$\sup_{|\theta-\theta_0|\le\delta_n} \|A_n(\theta) - A(\theta)\| = o_p(1)$$

for every sequence $\delta_n \to 0$.

- **Theorem C:** (GMM: Pakes and Pollard, 1989; Hansen, 1982). If C0-C6 hold, then Theorem B holds with Ψ replaced by $A\Psi(\theta)$, V replaced by AVA^T, and Γ replaced by $A\Gamma = A\dot{\Psi}_0$. Thus with $W \equiv A^T A$

$$\begin{aligned} \sqrt{n}(\widehat{\theta}_n - \theta_0) &\rightarrow_d -(\Gamma^T W \Gamma)^{-1}\Gamma^T W \mathbb{Z} \\ &\sim N_d(0, (\Gamma^T W \Gamma)^{-1}(\Gamma^T W V W \Gamma)(\Gamma^T W \Gamma)^{-1}). \end{aligned}$$

- The covariance is minimized by the choice $W = V^{-1}$ when V is non-singular and then it reduces to

$$(\Gamma^T V^{-1}\Gamma)^{-1}. \tag{1}$$

- Note that this further reduces to the asymptotic variance of Huber's Z-theorem when $p = d$ and Γ is non-singular.
- (1) is exactly the form of the covariance of Empirical Likelihood and Generalized Empirical Likelihood Estimators: Qin and Lawless (1994), Newey and Smith (2004), under stronger regularity conditions.

- Chamberlain (1987) shows that $(\Gamma^T V^{-1}\Gamma)^{-1}$ is the efficiency bound for estimation of θ in the constraint-defined model $\mathcal{P} = \{P: \ \Psi(\theta) = 0, \theta \in \mathbb{R}^p\}$. Newey (2004) treats efficiency in the case when V is singular.
- Andrews (2002) studies the GMM estimators when C.5 fails.
- P. W. Millar (1984) studies infinite-dimensional versions of GMM estimators as Minimum Distance Estimators, and gives a theorem that contains the Pakes-Pollard (1989) theorems. Millar allows $\Theta \subset \mathbb{B}$, a Banach space, and assumes that the functions Ψ_n and Ψ take values in another Banach space $\mathbb{L}$, but focuses on cases in which $\mathbb{L}$ is a Hilbert space, and in fact the theorem of Hansen (1982) and Pakes and Pollard (1989) continue to hold in this setting.
- (Connections to Empirical Likelihood): Lopez, van Keilegom, and Veraverbeke (2009) use the methods of Pakes and Pollard (1989) and Sherman (1993) to extend the results of Qin and Lawless (1994) to non-smooth functions. (Smoothness weakened; boundedness of basic functions strengthened. Can we weaken both?)

- Chamberlain (1987) shows that $(\Gamma^T V^{-1}\Gamma)^{-1}$ is the efficiency bound for estimation of θ in the constraint-defined model $\mathcal{P} = \{P : \ \Psi(\theta) = 0, \theta \in \mathbb{R}^p\}$. Newey (2004) treats efficiency in the case when V is singular.
- Andrews (2002) studies the GMM estimators when C.5 fails.
- P. W. Millar (1984) studies infinite-dimensional versions of GMM estimators as Minimum Distance Estimators, and gives a theorem that contains the Pakes-Pollard (1989) theorems. Millar allows $\Theta \subset \mathbb{B}$, a Banach space, and assumes that the functions Ψ_n and Ψ take values in another Banach space $\mathbb{L}$, but focuses on cases in which $\mathbb{L}$ is a Hilbert space, and in fact the theorem of Hansen (1982) and Pakes and Pollard (1989) continue to hold in this setting.
- (Connections to Empirical Likelihood): Lopez, van Keilegom, and Veraverbeke (2009) use the methods of Pakes and Pollard (1989) and Sherman (1993) to extend the results of Qin and Lawless (1994) to non-smooth functions. (Smoothness weakened; boundedness of basic functions strengthened. Can we weaken both?)

- Setting for BKRW (1993), van der Vaart (1995) infinite-dimensional $Z-$theorem: see van der Vaart and Wellner (1996)
  - $\theta \in \Theta \subset B$, a Banach space
  - $\Psi_n : \Theta \to \mathbb{L}$, random;
  - $\Psi : \Theta \to \mathbb{L}$, deterministic; $\Psi(\theta_0) = 0$.

- **Theorem B:** Suppose that: $\hat{\theta}_n \to_p \theta_0$ in B, and that:

  B1. $\Psi_n(\hat{\theta}_n) = o_p(n^{-1/2})$ in $\mathbb{L}$

  B2. $\sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0)) \Rightarrow \mathbb{Z}$ in $\mathbb{L}$

  B3. Ψ is Fréchet differentiable at θ_0 with (continuously) invertible derivative $\dot{\Psi}_0 = \dot{\Psi}(\theta_0)$.

  B4. For every $\delta_n \to 0$

$$\sup_{\|\theta - \theta_0\| \le \delta_n} \frac{\|\sqrt{n}((\Psi_n - \Psi)(\theta) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)\|_{\mathbb{L}}}{1 + \sqrt{n}\|\theta - \theta_0\|_B} = o_p(1).$$

  Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -\dot{\Psi}_0^{-1}\mathbb{Z} \quad \text{in} \quad B.$$

- Setting for Millar's infinite-dimensional GMM (or-MDE) theorem.
  - $\theta \in \Theta \subset B$, a Banach space
  - $\Psi_n : \Theta \to \mathbb{L}$, random; $\mathbb{L}$ Hilbert
  - $\Psi : \Theta \to \mathbb{L}$, deterministic; $\Psi(\theta_0) = 0$.
- **Theorem B:** Assume that
  B0. $\hat{\theta}_n \to_p \theta_0$ in B
  B1. $\|\Psi_n(\hat{\theta}_n)\|_{\mathbb{L}} = o_p(n^{-1/2}) + \inf_{\theta\in\Theta} \|\Psi_n(\theta)\|_{\mathbb{L}}$
  B2. $\sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0)) \Rightarrow \mathbb{Z}$ in $\mathbb{L}$
  B3. Ψ is differentiable at θ_0 with invertible derivative $\dot{\Psi}_0 = \Gamma$ satisfying $\Gamma^T\Gamma : B \to B$ invertible.
  B4. For every $\delta_n \to 0$

$$\sup_{\|\theta-\theta_0\|\le\delta_n} \frac{\|\sqrt{n}((\Psi_n - \Psi)(\theta) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)\|_{\mathbb{L}}}{1 + \sqrt{n}\|\Psi_n(\theta)\|_{\mathbb{L}} + \sqrt{n}\|\Psi(\theta)\|_{\mathbb{L}}} = o_p(1).$$

  Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -(\Gamma^T\Gamma)^{-1}\Gamma^T\mathbb{Z} \quad \text{in} \quad B.$$

4. Summary; problems and open questions

- Z-theorems
  - classical Huber $Z-$theorem
  - van der Vaart (1995): infinite dimensional $Z-$theorem; see also vdV-W (1996).
  - Breslow - Wellner (2008) infinite dimensional $Z-$theorem with (possibly) infinite-dimensional nuisance parameter
- GMM or MD theorems
  - Hansen (1982)
  - Pakes-Pollard (1989): further restrictions $Z-$theorem or GMM; related to EL
  - Millar (1984) infinite-dimensional GMM or Mininum Distance theorem.
  - Newey (1994), Chen-Linton-van Keilegom (2004) finite-dimensional $Z-$theorem with infinite-dimensional nuisance parameter.

- Application to semiparametric missing data models

- Application to semiparametric missing data models
  - Basic idea: separate calculations for sampling design and for model.

- Application to semiparametric missing data models
  - Basic idea: separate calculations for sampling design and for model.
  - Sampling assumptions give properties of IPW empirical process $\mathbb{G}_N^\pi$

- Application to semiparametric missing data models
  - Basic idea: separate calculations for sampling design and for model.
  - Sampling assumptions give properties of IPW empirical process $\mathbb{G}_N^{\pi}$
  - Likelihood calculations for complete data problem give efficient influence function $\tilde{\ell}_{\nu}$ for ν.

- Application to semiparametric missing data models
  - Basic idea: separate calculations for sampling design and for model.
  - Sampling assumptions give properties of IPW empirical process $\mathbb{G}_N^{\pi}$
  - Likelihood calculations for complete data problem give efficient influence function $\tilde{\ell}_{\nu}$ for ν.
  - Basic Issue: estimating the π's can lead to increased efficiency.
    - Regression on $Z = Z(V)$?
    - Calibration?

- Further problems and possible approaches:

- Further problems and possible approaches:
  - Improved methods of calibration: connection between survey sampling and Empirical Likelihood?

- Further problems and possible approaches:
  - Improved methods of calibration: connection between survey sampling and Empirical Likelihood?
  - Infinite-dimensional version of Pakes-Pollard GMM theorem (Millar)?

- Further problems and possible approaches:
  - Improved methods of calibration: connection between survey sampling and Empirical Likelihood?
  - Infinite-dimensional version of Pakes-Pollard GMM theorem (Millar)?
  - Infinite-dimensional constraint version of EL?

- Further problems and possible approaches:
  - Improved methods of calibration: connection between survey sampling and Empirical Likelihood?
  - Infinite-dimensional version of Pakes-Pollard GMM theorem (Millar)?
  - Infinite-dimensional constraint version of EL?
  - Can we handle both estimating the π's and and finite popl'n sampling? Saegusa (2010).

- Further problems and possible approaches:
  - Improved methods of calibration: connection between survey sampling and Empirical Likelihood?
  - Infinite-dimensional version of Pakes-Pollard GMM theorem (Millar)?
  - Infinite-dimensional constraint version of EL?
  - Can we handle both estimating the π's and and finite popl'n sampling? Saegusa (2010).
  - Efficiency gains via finite (without replacement) sampling. Further gains possible via other sampling designs?
    - Hájek (1964), Rosen (1972a,b), Isaki and Fuller (1982)
    - Lin (2000)

- Further problems and possible approaches, continued

- Further problems and possible approaches, continued
  - Can we handle problems with nuisance parameter estimators not converging at rate $\sqrt{n}$ together with finite-sampling or more complex designs?
    - $Z-$theorems of Huang (1995), Wellner and Zhang (2006); GMM-theorem with nuisance parameters: Newey (1994).
    - Empirical likelihood with nuisance parameters: Hjort, McKeague, van Keilegom (2009).
    - More to learn from the econometricians? Newey and Smith (2004), Schennach (2007)

References:

- Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Statist.* **34**, 86 - 102.
- Breslow, N. E. and Wellner, J. A. (2008). A $Z-$theorem with estimated parameters and correction note for “Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression”. *Scand. J. Statist.* **35**, 186 - 192.
- van der Vaart, A. W. and Wellner, J. A. (2007). Empirical processes indexed by estimated functions. In Asymptotics: particles, processes and inverse problems, *IMS Lecture Notes - Monograph Series* **55**, 234-252.

- Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009). Using the whole cohort in the anlysis of case-cohort data. *Amer. J. Epidem. 169*, 1398 - 1405.
- Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009). Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statist. Biosc. 1*, 32-49.