M- and Z- theorems; GMM and Empirical Likelihood

Wellner; 5/13/98, 1/26/07, 5/08/09, 6/14/2010

Z-theorems: Notation and Context

Suppose that $\Theta \subset R^k$, and that

$$\begin{aligned} \Psi_n &: \Theta \to \mathbb{R}^k, \quad \text{random maps} \\ \Psi &: \Theta \to \mathbb{R}^k, \quad \text{deterministic maps.} \end{aligned}$$

Suppose that $\widehat{\theta}_n$ and θ_0 are the corresponding solutions (or approximate solutions) of

$$\Psi_n(\widehat{\theta}_n) = 0 \qquad \text{or} \qquad \Psi_n(\widehat{\theta}_n) = o_p(n^{-1/2}),$$

$$\Psi(\theta_0) = 0.$$

In the simple case of i.i.d. data $X_1, \ldots, X_n$ i.i.d. P_0 with empirical measure $\mathbb{P}_n$, and then, for the usual case of linear estimating equations, the functions Ψ_n, Ψ are given by

$$\Psi_n(\theta) = \mathbb{P}_n\psi(\cdot, \theta), \qquad \text{and} \qquad \Psi(\theta) = P_0\psi(\cdot, \theta)$$

for a vector of functions $\psi : \mathcal{X} \times \Theta \to R^k$, $\psi(x, \theta) = \underline{\psi}(x, \theta)$; often the functions ψ are score functions motivated by likelihood, pseudolikelihood, quasilikelihood, or some other "likelihood" for the data.

Here are the four basic conditions needed for Huber's $Z-$ theorem:

A1 $\Psi_n(\widehat{\theta}_n) = o_p(n^{-1/2})$ and $\Psi(\theta_0) = 0$.

A2 $\sqrt{n}(\Psi_n - \Psi)(\theta_0) \to_d \mathbb{Z}_0$.

A3 For every sequence $\delta_n \to 0$,

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

A4 The function Ψ is (Fréchet-)differentiable at θ_0 with nonsingular derivative $\dot{\Psi}(\theta_0) \equiv \dot{\Psi}_0$:

$$\Psi(\theta) - \Psi(\theta_0) - \dot{\Psi}_0(\theta - \theta_0) = o(|\theta - \theta_0|).$$

Theorem 1. (Huber (1967); Pollard (1985)). Suppose that A1 - A4 hold. Let $\widehat{\theta}_n$ be random maps into $\Theta \subset R^k$ satisfying $\widehat{\theta}_n \to_p \theta_0$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d -\dot{\Psi}_0^{-1}(\mathbb{Z}_0);$$

if $\mathbb{Z}_0 \sim N_k(0, A)$, then this yields, with $\dot{\Psi}_0 \equiv B$,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d N_k(0, B^{-1}A(B^{-1})^T)\,.$$

Proof. By definition of $\widehat{\theta}_n$ and θ_0,

$$\begin{aligned}
\sqrt{n}(\Psi(\widehat{\theta}_n) - \Psi(\theta_0)) &= \sqrt{n}(\Psi(\widehat{\theta}_n) - \Psi_n(\widehat{\theta}_n)) + o_p(1) \\
&= -\sqrt{n}(\Psi_n - \Psi)(\theta_0) \\
&\qquad - \left\{\sqrt{n}(\Psi_n - \Psi)(\widehat{\theta}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)\right\} + o_p(1) \\
&= -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_p(1 + \sqrt{n}|\widehat{\theta}_n - \theta_0|) + o_p(1)\,; \qquad (1)
\end{aligned}$$

here the last equality holds by A3 and $\widehat{\theta}_n \to_p \theta_0$. Since $\dot{\Psi}_0$ is continuously invertible, there exists a constant $c > 0$ such that

$$\|\dot{\Psi}_0(\theta - \theta_0)\| \geq c\|\theta - \theta_0\|$$

for every θ; this is just the basic property of a nonsingular matrix. By A4 (differentiability of Ψ), this yields

$$|\Psi(\theta) - \Psi(\theta_0)| \geq c|\theta - \theta_0| + o(|\theta - \theta_0|)\,.$$

By (1) it follows that

$$\sqrt{n}|\widehat{\theta}_n - \theta_0|(c + o_p(1)) \leq O_p(1) + o_p(1 + \sqrt{n}|\widehat{\theta}_n - \theta_0|)\,,$$

which implies

$$\sqrt{n}|\widehat{\theta}_n - \theta_0| = O_p(1)\,.$$

Hence from (1) again and A.4 it follows that

$$\dot{\Psi}_0(\sqrt{n}(\widehat{\theta}_n - \theta_0)) + o_p(\sqrt{n}|\widehat{\theta}_n - \theta_0|) = -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_p(1)$$

and therefore

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d -\dot{\Psi}_0^{-1}(\mathbb{Z}_0)$$

by A2 and A4. □

Example 1. Suppose that $X_1, \ldots, X_n$ are i.i.d. with distribution P_0 on $\mathbb{R}$. Suppose we assume (perhaps incorrectly) that the model is $\mathcal{P} = \{P_\theta : p_\theta(x) = f(x - \theta), \theta \in \mathbb{R}\}$ where f is the logistic density given by

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2}.$$

If $P_0 \notin \mathcal{P}$, then the model is not correctly specified and we say that $\mathcal{P}$ is *miss-specified*. If $\widehat{\theta}_n$ is the MLE of θ for the model $\mathcal{P}$, what can we say about $\widehat{\theta}_n$ when $P_0 \notin \mathcal{P}$? For the logistic model $\mathcal{P}$ the MLE is given by the solution of the score equation

$$\mathbb{P}_n \dot{\mathbf{l}}_\theta(X) = \frac{1}{n}\sum_{i=1}^{n} \dot{\mathbf{l}}_\theta(X_i) = 0$$

where

$$\dot{\mathbf{l}}_\theta(x) = -(f'/f)(x-\theta) = \frac{1-\exp(-(x-\theta))}{1+\exp(-(x-\theta))}$$

takes values in $(-1,1)$ and is continuous and strictly decreasing as a function of θ for each fixed x. It follows that

$$\Psi_n(\theta) = \mathbb{P}_n \dot{\mathbf{l}}_\theta = \frac{1}{n}\sum_{i=1}^{n} \dot{\mathbf{l}}_\theta(X_i)$$

is continuous and strictly decreasing with values in $(-1,1)$. Furthermore $\Psi_n(\theta) \to \pm 1$ as $\theta \to \mp\infty$, and hence there exists a unique solution $\widehat{\theta}_n$ of $\Psi_n(\widehat{\theta}_n) = 0$.

Let

$$\Psi(\theta) \equiv P_0 \dot{\mathbf{l}}_\theta(X;\theta) = P_0\left(-\frac{f'}{f}(X-\theta)\right), \quad \theta \in \mathbb{R}.$$

Ψ is also a monotone strictly decreasing function of θ with a unique solution $\theta_0 \equiv \theta(P_0)$ of $\Psi(\theta) = 0$. Thus condition A1 of Huber's theorem holds.

For condition A2, note that

$$\begin{aligned}\sqrt{n}(\Psi_n - \Psi)(\theta_0) &= \sqrt{n}(\mathbb{P}_n - P_0)(\dot{\mathbf{l}}_\theta(\cdot;\theta_0) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\dot{\mathbf{l}}_\theta(X_i;\theta_0) - P_0\dot{\mathbf{l}}_\theta(\cdot;\theta_0))\end{aligned}$$

is a normalized sum of i.i.d. random variables with mean 0 which are bounded and hence

$$Var_{P_0}(\dot{\mathbf{l}}_\theta(X_1;\theta_0)) = P_0\dot{\mathbf{l}}_\theta^2(X;\theta_0) \equiv A < \infty.$$

Thus by the ordinary CLT

$$\sqrt{n}(\Psi_n - \Psi)(\theta_0) \to_d \mathbb{Z} \sim N(0, A),$$

and condition A2 holds. Now since

$$\Psi(\theta) = P_0\dot{\mathbf{l}}_\theta(X;\theta) = P_0\left(\frac{1-\exp(-(X-\theta))}{1+\exp(-(X-\theta))}\right)$$

is bounded and $\dot{\mathbf{l}}_\theta(X;\theta)$ is a bounded and differentiable function of θ for all x. Thus

$$\begin{aligned}\dot{\Psi}(\theta_0) &= P_0\ddot{\mathbf{l}}_{\theta\theta}(X;\theta_0) = -2P_0\left(\frac{e^{-(X-\theta_0)}}{(1+e^{-(X-\theta_0)})^2}\right)\\ &= -2P_0(f(X-\theta_0)) \equiv -B < 0\end{aligned}$$

for any P_0 and where f is the standard logistic density. Thus condition A4 holds.

Note that if P_0 has density $g(\cdot-\theta)$ where g is symmetric about 0, then

$$\begin{aligned}\Psi(\theta) &= P_0\dot{\mathbf{l}}_\theta(X) = \int -\frac{f'}{f}(x-\theta)g(x-\theta_0)dx\\ &= \int -\frac{f'}{f}(y-(\theta-\theta_0))g(y)dy\\ &= 0 \quad \text{at} \quad \theta=\theta_0\end{aligned}$$

since $-(f'/f)$ is odd and g is even. Thus $\theta(P_0)=\theta_0$, the center of symmetry of P_0.

It remains only to verify condition A3: now

$$\begin{aligned}\sqrt{n}(\Psi_n-\Psi)(\theta)-\sqrt{n}(\Psi_n-\Psi)(\theta_0) &= \mathbb{G}_n(\dot{\mathbf{l}}_\theta(X;\theta)-\dot{\mathbf{l}}_\theta(X;\theta_0))\\ &= \mathbb{G}_n(f_\theta - f_{\theta_0})\end{aligned}$$

where

$$\mathcal{F} = \{f_\theta \equiv \dot{\mathbf{l}}_\theta(\cdot;\theta): \quad \theta\in\mathbb{R}\} = \{-(f'/f)(\cdot-\theta): \quad \theta\in\mathbb{R}\}$$

is the class of shifts of the (one!) bounded monotone function $-(f'/f)$. Thus the collection of subgraphs of the class $\mathcal{F}$ are linearly - ordered by inclusion: if $C_\theta \equiv \{(x,y): y\le f_\theta(x), x\in\mathbb{R}\}$ then $C_\theta \subset C_{\theta'}$ if $\theta<\theta'$; that is, this class of sets is linearly ordered by inclusion, and hence a VC class of sets with VC index $V(\mathcal{C})=2$. Thus $\mathcal{F}$ is a bounded VC - subgraph class of functions, and hence is uniformly $P-$Donsker and (in particular) Donsker for every P_0. Thus for every sequence $\delta_n'\to 0$ we have

$$Pr\left(\sup_{\rho_{P_0}(f_\theta,f_{\theta_0})<\delta_n'}|\mathbb{G}_n(f_\theta-f_{\theta_0})|>\epsilon\right)\to 0$$

for every $\epsilon>0$, and this implies that

$$\sup_{|\theta-\theta_0|\le\delta_n}|\sqrt{n}(\Psi_n-\Psi)(\theta)-\sqrt{n}(\Psi_n-\Psi)(\theta_0)| = o_p(1)$$

which in turn implies that condition A3 holds. Thus by Huber's $Z-$theorem

$$\sqrt{n}(\widehat{\theta}_n-\theta_0)\to_d -\dot{\Psi}(\theta_0)^{-1}\mathbb{Z} = B^{-1}\mathbb{Z}\sim N(0,B^{-1}AB^{-1}).$$

It is interesting to consider these calculations in the more general case of an arbitrary log-concave density f replacing the standard logistic density.

Example 2. Now consider moment estimators for the two component exponential mixture model $\mathcal{P}$ defined by

$$p_\theta(x) = p\mu_1^{-1}\exp(-x/\mu_1) + (1-p)\mu_2^{-1}\exp(-x/\mu_2)$$

where $\theta = (p, \mu_1, \mu_2) \in (0,1) \times \mathbb{R}^+ \times \mathbb{R}^+$. Here we define

$$\begin{aligned}
\psi_1(x,\theta) &= x - (p\mu_1 + (1-p)\mu_2), \\
\psi_2(x,\theta) &= \frac{1}{2}x^2 - (p\mu_1^2 + (1-p)\mu_2^2), \\
\psi_3(x,\theta) &= \frac{1}{6}x^3 - (p\mu_1^3 + (1-p)\mu_2^3).
\end{aligned}$$

Thus if we set $\underline{T}(x) = (x, x^2/2, x^3/6)^T$ and $\underline{t}(\theta) = (p\mu_1 + (1-p)\mu_2, (p\mu_1^2 + (1-p)\mu_2^2, p\mu_1^3 + (1-p)\mu_2^3)^T$, we can write

$$\underline{\psi}(x,\theta) = \underline{T}(x) - \underline{t}(\theta).$$

Then we set

$$\begin{aligned}
\underline{\Psi}_n(\theta) &= \mathbb{P}_n\underline{\psi}(\cdot,\theta) = \mathbb{P}_n(\underline{T}(X)) - \underline{t}(\theta), \\
\underline{\Psi}(\theta) &= P_0\underline{\psi}(\cdot,\theta) = P_0(\underline{T}(X)) - \underline{t}(\theta).
\end{aligned}$$

Then $\Psi(\theta_0) = 0$ if $P_0 = P_{\theta_0} \in \mathcal{P}$ and $\Psi_n(\theta) = 0$ has a solution $\widetilde{\theta}_n$ given by xyz. Thus condition A1 of Huber's theorem holds.

Now if $P_0 X^6 < \infty$, then $X^3 \in L^2(P_0)$ and it follows from the multivariate CLT that

$$\begin{aligned}
\sqrt{n}(\underline{\Psi}_n - \underline{\Psi})(\theta_0) &= \sqrt{n}(\mathbb{P}_n - P_0)(\underline{\psi}(\cdot;\theta_0) \\
&\to_d N_3(0, A)
\end{aligned}$$

where $A = Cov_{P_0}(X, X^2/2, X^3/6)$. Thus condition A2 of Huber's theorem holds if $P_0(X^6) < \infty$.

To see that A3 holds, note that

$$\begin{aligned}
&\sqrt{n}(\underline{\Psi}_n - \underline{\Psi})(\theta) - \sqrt{n}(\underline{\Psi}_n - \underline{\Psi})(\theta_0) \\
&= \sqrt{n}(\mathbb{P}_n - P_0)(\underline{T} - \underline{t}(\theta)) - \sqrt{n}(\mathbb{P}_n - P_0)(\underline{T} - \underline{t}(\theta_0)) \\
&= \sqrt{n}(\mathbb{P}_n - P_0)(\underline{T}) - \sqrt{n}(\mathbb{P}_n - P_0)(\underline{T}) \\
&= 0.
\end{aligned}$$

Finally,

$$\underline{\Psi}(\theta) = P_0(\underline{T}) - \underline{t}(\theta),$$

so $\dot{\underline{\Psi}}(\theta_0) = -\dot{\underline{t}}(\theta_0)$ can be easily be calculated:

$$\begin{aligned}
\dot{\underline{\Psi}}(\theta_0) &= -\dot{\underline{t}}(\theta_0) \\
&= -\begin{pmatrix} \mu_1 - \mu_2 & p & 1-p \\ \mu_1^2 - \mu_2^2 & 2p\mu_1 & 2(1-p)\mu_2 \\ \mu_1^3 - \mu_2^3 & 3p\mu_1^2 & 3(1-p)\mu_2^2 \end{pmatrix} \\
&\equiv -B
\end{aligned}$$

It is not too hard to show that B is non-singular if $\mu_1 \neq \mu_2$, so condition A4 holds. Thus by Huber's theorem,

$$\sqrt{n}(\widetilde{\theta}_n - \theta_0) \to_d B^{-1}\mathbb{Z} \sim N_3(0, B^{-1}A(B^{-1})^T).$$

The interesting part of this example is that condition A3 holds very easily, and that this is apparently generally true for method of moments estimators.

Example 3. Zero-inflated Poisson distribution (from Knight (2000), page 197). Suppose that $X \sim p_\theta$ where $p_\theta(x) = \gamma\delta_0(x) + (1-\gamma)e^{-\lambda}\lambda^x/x!$ for $\theta = (\gamma, \lambda)$ with $\gamma \in [0,1]$, $\lambda > 0$, and $x \in \mathbb{N} = \{0, 1, 2, \ldots\}$. This is a mixture model involving the discrete distributions δ_0 and Poisson(λ) on the non-negative integers. We consider method of moment estimators based on the two functions $g_1(x) = 1_{\{0\}}(x)$ and $g_2(x) = x$. Then for $X \sim p_\theta$ we have

$$\begin{aligned}
E_\theta g_1(X) &= P_\theta(X = 0) = \gamma + (1-\gamma)e^{-\lambda} \equiv t_1(\theta), \\
E_\theta g_2(X) &= E_\theta X = (1-\gamma)\lambda \equiv t_2(\theta).
\end{aligned}$$

Thus we define

$$\begin{aligned}
\psi_1(x, \theta) &= g_1(x) - t_1(\theta) = 1_{\{0\}}(x) - \gamma - (1-\gamma)e^{-\lambda}, \\
\psi_2(x, \theta) &= g_2(x) - t_2(\theta) = x - (1-\gamma)\lambda.
\end{aligned}$$

Then

$$0 = \Psi_n(\theta) = \mathbb{P}_n\psi(X, \theta) = \begin{pmatrix} \mathbb{P}_n(\{0\}) - t_1(\theta) \\ \mathbb{P}_n X - t_2(\theta) \end{pmatrix}$$

defines $\widetilde{\theta} = (\widetilde{\gamma}, \widetilde{\lambda})$ and

$$0 = \Psi(\theta) = P_0\psi(X, \theta) = \begin{pmatrix} P_0(\{0\}) - t_1(\theta) \\ P_0 X - t_2(\theta) \end{pmatrix}$$

defines $\theta_0 \equiv \theta(P_0) = (\gamma(P_0), \lambda(P_0))$ via $p_0(0) = \gamma + (1-\gamma)\exp(-\lambda)$ and $\mu_0 = (1-\gamma)\lambda$ where $p_0(0) \equiv P_0(\{0\})$ and $\mu_0 = P_0 X = E_0(X)$. Thus $1 - \gamma = \mu_0/\lambda$ and this yields $(1 - p_0)\lambda = \mu_0(1 - \exp(-\lambda))$.

If $E_0 X^2 < \infty$, then

$$\sqrt{n}(\Psi_n - \Psi)(\theta_0) = \sqrt{n}(\mathbb{P}_n - P_0)(1_{\{0\}}(X), X)^T) \to_d \mathbb{Z} \sim N_2(0, A)$$

where $A = Cov_0(1_{\{0\}}(X), X)$, so condition A2 holds. Furthermore condition A3 holds trivially since

$$\sqrt{n}(\underline{\Psi}_n - \underline{\Psi})(\theta) - \sqrt{n}(\underline{\Psi}_n - \underline{\Psi})(\theta_0) = 0$$

just as in Example 2. Much as in Example 2,

$$\dot{\Psi}(\theta_0) \;=\; -\dot{t}(\theta_0) = -\begin{pmatrix} 1 - e^{-\lambda_0} & -(1-\gamma_0)e^{-\lambda_0} \\ -\lambda_0 & 1 - \gamma_0 \end{pmatrix} \equiv -B$$

where $\det(B) = (1-\gamma_0)e^{-\lambda_0}(e^{\lambda_0} - 1 - \lambda_0) > 0$ if $\gamma_0 \in (0,1)$. Thus it follows from Huber's theorem that

$$\sqrt{n}(\widetilde{\theta}_n - \theta_0) \to_d B^{-1}\mathbb{Z} \sim N_2(0, B^{-1}A(B^{-1})^T).$$

Example 4. Consider estimation of θ in the simple triangular density family $\mathcal{P} = \{P_\theta : \theta \in [0,1]\}$ given by the densities

$$p(x;\theta) = 2\left\{\frac{x}{\theta}1_{[0,\theta]}(x) + \frac{1-x}{1-\theta}1_{(\theta,1]}(x)\right\}, \qquad \theta \in [0,1].$$

Then the score function for θ for one observation is

$$\dot{\mathbf{l}}_\theta(x;\theta) = -\frac{1}{\theta}1_{[0,\theta]}(x) + \frac{1}{1-\theta}1_{(\theta,1]}(x),$$

at least in the sense of a Hellinger derivative:

$$\int_0^1 \{\sqrt{p(x;\theta)} - \sqrt{p(x;\theta_0)} - 2^{-1}\dot{\mathbf{l}}_\theta(x;\theta_0)\sqrt{p(x;\theta_0)}\}^2 dx = o(|\theta - \theta_0|^2).$$

Our goal is to use Huber's theorem to study the behavior of the MLE of θ in this model when (possibly) $X, X_1, \ldots, X_n$ are i.i.d. P on $[0,1]$ with $P \notin \mathcal{P}$. Let $F(x) = P(X \le x)$ be the distribution function corresponding to X. From the score calculation above, the score equation for estimation of θ is equivalent to

$$\Psi_n(\theta) = \mathbb{F}_n(\theta) - \theta = 0.$$

The corresponding population version of Ψ_n is Ψ given by

$$\Psi(\theta) = F(\theta) - \theta, \qquad \text{where } \; F(x) = \int_0^x p(y)dy.$$

It is reasonable to assume that $\Psi(\theta_0) = 0$ has a unique solution $\theta_0 = \theta_0(P)$ if P has a density p which is unimodal on $[0,1]$. (Of course it is easy to construct examples in which this has only trivial solutions 0 or 1, or for which there are many solutions: for the former, consider the "anti-triangular density" $p(x) = 2(1/2 - x)1_{[0,1/2]}(x) + 2(x - 1/2)1_{(1/2,1]}(x)$; for the latter consider $p(x) = 1 + (1/2)\sin(6\pi x)$.)

Let $\theta_0 = \theta_0(P)$ satisfy $\Psi(\theta_0) = 0 = F(\theta_0) - \theta_0$. Then $F(\theta_0) = \theta_0$, and

$$\begin{aligned}
\sqrt{n}\Psi_n(\theta_0) &= \sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0)) \\
&= \sqrt{n}(\mathbb{F}_n(\theta_0) - F(\theta_0) \\
&\to_d \mathbb{Z}_0 \sim N(0, F(\theta_0)(1 - F(\theta_0))) = N(0, \theta_0(1-\theta_0)) \equiv N(0, A)
\end{aligned}$$

Thus A2 holds. Moreover if F is differentiable at θ_0 with derivative $p(\theta_0)$, then A4 holds with

$$\dot{\Psi}(\theta_0) = p(\theta_0) - 1 \equiv B$$

Furthermore, the condition A3 holds since, for any $\delta_n \to 0$,

$$\begin{aligned}
&\sup_{\theta:|\theta-\theta_0|\le\delta_n} |\sqrt{n}(\Psi_n - \Psi)(\theta) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)| \\
&= \sup_{\theta:|\theta-\theta_0|\le\delta_n} |\sqrt{n}(\mathbb{F}_n - F)(\theta) - \sqrt{n}(\mathbb{F}_n - F)(\theta_0)| \\
&\to_p 0
\end{aligned}$$

if F is continuous at θ_0 by the asymptotic equicontinuity of the empirical process. We conclude from Huber's theorem that any solution of $\Psi_n(\hat{\theta}_n) = o_p(n^{-1/2})$ satisfies

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta_0) &\to_d -(p(\theta_0) - 1)^{-1}\mathbb{Z}_0 \\
&\sim N(0, B^{-1}AB^{-1}) = N\left(0, \frac{\theta_0(1-\theta_0)}{(p(\theta_0)-1)^2}\right).
\end{aligned}$$

When $P \in \mathcal{P}$ holds, $\theta_0(P_\theta) = \theta$ (so $\theta_0(P_{\theta_0}) = \theta_0$), and $p(\theta_0; \theta_0) = 2$. Thus the asymptotic variance in the conclusion of Huber's theorem reduces to $\theta_0(1-\theta_0)$, which agrees with the information bound calculation based on the score $\dot{\mathbf{l}}_\theta$. □

Now our goal is to extend Theorem 1 to an infinite-dimensional setting in which Θ is a Banach space. A sufficiently general Banach space is the space

$$l^\infty(H) \equiv \{z : H \to R \big| \, ||z|| = \sup_{h\in H} |z(h)| < \infty\}$$

where H is a collection of functions. We suppose that

$$\Psi_n : \Theta \to L \equiv l^\infty(H'), \quad n = 1, 2, ...$$

are random, and that

$$\Psi : \Theta \to L \equiv l^\infty(H')\,,$$

is deterministic. Suppose that either

$$\Psi_n(\widehat{\theta}_n) = 0 \qquad \text{in} \qquad L;$$

(i.e. $\Psi_n(\widehat{\theta}_n)(h') = 0$ for all $h' \in H'$), or

$$\Psi_n(\widehat{\theta}_n) = o_p(n^{-1/2}) \qquad \text{in} \qquad L;$$

(i.e. $||\Psi_n(\widehat{\theta}_n)||_{H'} = o_p(n^{-1/2})$).
Here are the four basic conditions needed for the infinite-dimensional version of Huber's $Z-$ theorem due to Van der Vaart (1995):

B1 $\Psi_n(\widehat{\theta}_n) = o_p(n^{-1/2})$ in $l^\infty(H')$ and $\Psi(\theta_0) = 0$ in $l^\infty(H')$.

B2 $\sqrt{n}(\Psi_n - \Psi)(\theta_0) \Rightarrow \mathbb{Z}_0$ in $l^\infty(H')$.

B3 For every sequence $\delta_n \to 0$,

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

B4 The function Ψ is (Fréchet-)differentiable at θ_0 with derivative $\dot{\Psi}(\theta_0) \equiv \dot{\Psi}_0$ having a bounded (continuous) inverse:

$$\|\Psi(\theta) - \Psi(\theta_0) - \dot{\Psi}_0(\theta - \theta_0)\| = o(\|\theta - \theta_0\|)\,.$$

Theorem. (van der Vaart, 1995). Suppose that B1 - B4 hold. Let $\widehat{\theta}_n$ be random maps into $\Theta \subset l^\infty(H')$ satisfying $\widehat{\theta}_n \to_p \theta_0$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \Rightarrow -\dot{\Psi}_0^{-1}(\mathbb{Z}_0) \qquad \text{in} \qquad l^\infty(H)\,.$$

Proof. Exactly the same as in the finite-dimensional case: see van der Vaart (1995) or van der Vaart and Wellner (1996), pages 310-312. □

M-theorems: Notation and context

Suppose that $\Theta \subset \mathbb{R}^k$ and that $m : \mathcal{X} \times \Theta \to \mathbb{R}$. We often write $m_\theta(x) = m(x, \theta)$ for $x \in \mathcal{X}$, $\theta \in \Theta$. Suppose that $\widehat{\theta}_n$ and θ_0 are the corresponding maximizers (or approximate maxmizers in the first case) of

$$\begin{aligned} \mathbb{M}_n(\theta) &\equiv \mathbb{P}_n m(X, \theta) = \mathbb{P}_n m_\theta(X), \qquad \text{and} \\ M(\theta) &\equiv P_0 m(X, \theta) = P_0 m_\theta(X), \end{aligned}$$

respectively. A common choice for $m(x, \theta)$ would be $\log p(x; \theta) \equiv \log p_\theta(x)$ where $p_\theta(\cdot)$ is the density of P_θ with respect to some dominating measure μ for a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Then $\widehat{\theta}_n$ is a Maximum Likelihood (or approximate maximum likelihood) estimator for the model $\mathcal{P}$.

Theorem. Suppose that for each θ in an open subset of $\Theta \subset \mathbb{R}^k$, $x \mapsto m_\theta(x)$ is a measurable function such that $\theta \mapsto m_\theta(x) = m(x, \theta)$ is differentiable at θ_0 for P_0- almost every x with derivative $\dot{m}_{\theta_0}(x)$ and such that, for every θ_1, θ_2 in a neighborhood of θ_0 and a measurable function $\dot{m}$ with $P_0 \dot{m}^2 < \infty$,

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \dot{m} \|\theta_1 - \theta_2\|.$$

Furthermore, suppose that $\theta \mapsto P_0 m_\theta$ has a second order Taylor expansion at a point of maximum θ_0 with nonsingular second derivative matrix V_{θ_0}: i.e.

$$P_0 m_\theta = P_0 m_{\theta_0} + \frac{1}{2}(\theta - \theta_0)^T V_{\theta_0}(\theta - \theta_0) + o(\|\theta - \theta_0\|^2).$$

If $\mathbb{P}_n m_{\widehat{\theta}_n}(X) \geq \sup_\theta \mathbb{P}_n m_\theta(X) - o_p(n^{-1})$ and $\widehat{\theta}_n \to_p \theta_0$, then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{m}_{\theta_0}(X_i) + o_p(1).$$

In particular

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d N_k(0, V_{\theta_0}^{-1} P_0(\dot{m}_{\theta_0} \dot{m}_{\theta_0}^T) V_{\theta_0}^{-1}).$$

Proof. See van der Vaart, *Asymptotic Statistics*, section 5.3, pages 51 - 60. □

Also see van der Vaart, *Asymptotic Statistics*, section 5.5 and Theorem 5.39 page 65, for a theorem of this type with $m_\theta(x) = \log p_\theta(x)$. This is the cleanest theorem I know concerning asymptotic normality of MLE's in parametric models.

Applications and Extensions of Van der Vaart's Z-theorem:

- Gamma frailty model; Murphy (1995).
- Partially censored data; Van der Vaart (1995).
- Correlated gamma-frailty model; Parner (1998).
- Semiparametric biased sampling models; Gilbert (2000).
- Two-phase sampling models with data missing by design; Breslow, McNeney and Wellner (2003), Breslow and Wellner (2007), (2008).

However, in many statistical problems the parameter usually includes both a finite-dimensional parameter (e.g. regression parameters) and an infinite dimensional (nuisance) parameter. We now suppose that $\theta = (\beta, \Lambda)$, where β is a finite-dimensional parameter, say in $\mathbb{R}^d$, and Λ an infinite dimensional parameter (a function). The M-estimators of β, $\hat{\beta}_n$, and of Λ, $\hat{\Lambda}_n$, respectively, often have different convergence rates. The convergence rate for $\hat{\Lambda}_n$ is often smaller than $n^{1/2}$, such as $n^{1/3}$, or $n^{2/5}$ in some cases. Huang (1996) established a general theorem to show that under certain hypotheses, the maximum likelihood estimator of a finite dimensional parameter has $n^{1/2}$ convergence rate and is asymptotically semiparametric efficient, even though the convergence rate for the maximum likelihood estimator of the infinite dimensional parameter is smaller than $n^{1/2}$. He also successfully applied his general theorem to the proportional hazards model with interval censored data.

The following theorem due to Zhang (1998) generalizes the theorem of Huang (1996) to the case of inefficient M-estimators; it shows that under reasonable regularity hypotheses, the M-estimator of a finite-dimensional parameter β has $n^{1/2}$ convergence rate, and that $\hat{\beta}_n$ is asymptotically normal, even though the M-estimator of the corresponding infinite dimensional parameter Λ converges perhaps more slowly than $n^{1/2}$. The resulting asymptotic covariance matrix for the M-estimator of β has the well-known "sandwich" structure.

Here is the notation and conditions needed for the theorem. Let $\theta = (\beta, \Lambda)$, where $\beta \in \mathbb{R}^d$, and Λ is an infinite dimensional parameter in a class of functions $\mathcal{F}$. Λ_η is a parametric path in $\mathcal{F}$ through Λ, i.e. $\Lambda_\eta \in \mathcal{F}$, and $\Lambda_\eta|_{\eta=0} = \Lambda$.

Let $\mathbf{H} = \left\{ h : h = \frac{\partial \Lambda_\eta}{\partial \eta}\Big|_{\eta=0} \right\}$ and define

$$m_1(\beta, \Lambda; x) = \nabla_\beta m_{(\beta,\Lambda)}(x) \equiv \left(\frac{\partial}{\partial \beta_1} m_{(\beta,\Lambda)}(x), \cdots, \frac{\partial}{\partial \beta_d} m_{(\beta,\Lambda)}(x) \right)'.$$

$$m_2(\beta, \Lambda; x)[h] = \frac{\partial}{\partial \eta} m_{(\beta,\Lambda_\eta)}(x)\Big|_{\eta=0},$$

$$m_{11}(\beta,\Lambda;x)=\nabla^2_\beta m_{(\beta,\Lambda)}(x),$$
$$m_{12}(\beta,\Lambda;x)[h]=\frac{\partial}{\partial\eta}m_1(\beta,\Lambda_\eta;x)\Big|_{\eta=0},$$
$$m_{21}(\beta,\Lambda;x)[h]=\nabla_\beta m_2(\beta,\Lambda;x)[h],$$

and

$$m_{22}(\beta,\Lambda;x)[h,h]=\frac{\partial^2}{\partial\eta^2}m(\beta,\Lambda_\eta;x)\Big|_{\eta=0}.$$

We also define

$$S_1(\beta,\Lambda)=Pm_1(\beta,\Lambda;X),$$
$$S_2(\beta,\Lambda)[h]=Pm_2(\beta,\Lambda;X)[h],$$
$$S_{1n}(\beta,\Lambda)=\mathbb{P}_n m_1(\beta,\Lambda;X),$$
$$S_{2n}(\beta,\Lambda)[h]=\mathbb{P}_n m_2(\beta,\Lambda;X)[h],$$
$$\dot{S}_{11}(\beta,\Lambda)=Pm_{11}(\beta,\Lambda;X),$$
$$\dot{S}_{12}(\beta,\Lambda)[h]=\dot{S}'_{21}(\beta,\Lambda)[h]=Pm_{12}(\beta,\Lambda;X)[h],$$

and

$$\dot{S}_{22}(\beta,\Lambda)[h,h]=Pm_{22}(\beta,\Lambda;X)[h,h].$$

Furthermore, for $\mathbf{h}=(h_1,\cdots,h_d)'\in\mathbf{H}^d$, where $h_j\in\mathbf{H}$ for $j=1,2,\cdots,d,$, and $\mathbf{H}^d=\underbrace{\mathbf{H}\times\mathbf{H}\times\cdots\times\mathbf{H}}_{d}$, denote

$$m_2(\beta,\Lambda;x)[\mathbf{h}]=(m_2(\beta,\Lambda;x)[h_1],\cdots,m_2(\beta,\Lambda;x)[h_d])',$$

$$m_{12}(\beta,\Lambda;x)[\mathbf{h}]=(m_{12}(\beta,\Lambda;x)[h_1],\cdots,m_{12}(\beta,\Lambda;x)[h_d]),$$
$$m_{21}(\beta,\Lambda;x)[\mathbf{h}]=(m_{21}(\beta,\Lambda;x)[h_1],\cdots,m_{21}(\beta,\Lambda;x)[h_d]),$$
$$m_{22}(\beta,\Lambda;x)[\mathbf{h},h](m_{22}(\beta,\Lambda;x)[h_1,h],\cdots,m_{22}(\beta,\Lambda;x)[h_d,h])^T,$$

and define

$$S_2(\beta,\Lambda)[\mathbf{h}]=Pm_2(\beta,\Lambda;X)[\mathbf{h}],$$
$$S_{2n}(\beta,\Lambda)[\mathbf{h}]=\mathbb{P}_n m_2(\beta,\Lambda;X)[\mathbf{h}],$$
$$\dot{S}_{12}(\beta,\Lambda)[\mathbf{h}]=Pm_{12}(\beta,\Lambda;X)[\mathbf{h}],$$
$$\dot{S}_{21}(\beta,\Lambda)[\mathbf{h}]=Pm_{21}(\beta,\Lambda;X)[\mathbf{h}],$$

and

$$\dot{S}_{22}(\beta,\Lambda)[\mathbf{h},h]=Pm_{22}(\beta,\Lambda;X)[\mathbf{h},h].$$

The following Assumptions will be used to formulate our general theorem:

A1. **(Consistency and rate of convergence):**

$$|\hat{\beta}_n - \beta_0| = o_p(1) \quad \text{and} \quad \|\hat{\Lambda}_n - \Lambda_0\| = O_p(n^{-\gamma})$$

for some $\gamma > 0$.

A2. **(Zero-mean structure):**

$$S_1(\beta_0, \Lambda_0) = 0, \quad \text{and} \quad S_2(\beta_0, \Lambda_0)[h] = 0, \quad \text{for all } h \in \mathbf{H}.$$

A3. **(Positive "pseudo-information"):** There exists an $\mathbf{h}^* = (h_1^*, \cdots, h_d^*)^T$, $h_j^* \in \mathbf{H}$ $j = 1, \cdots, d$, such that

$$\dot{S}_{12}(\beta_0, \Lambda_0)[h] - \dot{S}_{22}(\beta_0, \Lambda_0)[\mathbf{h}^*, h] = 0, \tag{2}$$

for all $h \in \mathbf{H}$. Moreover, the matrix

$$A = -\dot{S}_{11}(\beta_0, \Lambda_0) + \dot{S}_{21}(\beta_0, \Lambda_0)[\mathbf{h}^*] = -P\left(m_{11}(\beta_0, \Lambda_0; X) - m_{21}(\beta_0, \Lambda_0; X)[\mathbf{h}^*]\right)$$

is nonsingular.

A4. **(Approximate solution of pseudo-score equations):** The estimator $(\hat{\beta}_n, \hat{\Lambda}_n)$ satisfies

$$S_{1n}(\hat{\beta}_n, \hat{\Lambda}_n) = o_{p^*}(n^{-1/2}),$$

and

$$S_{2n}(\hat{\beta}_n, \hat{\Lambda}_n)[\mathbf{h}^*] = o_{p^*}(n^{-1/2}).$$

A5. **(Stochastic equicontinuity):** For any $\delta_n \downarrow 0$ and $C > 0$,

$$\sup_{|\beta - \beta_0| \le \delta_n, \|\Lambda - \Lambda_0\| \le Cn^{-\gamma}} \left|\sqrt{n}(S_{1n} - S_1)(\beta, \Lambda) - \sqrt{n}(S_{1n} - S_1)(\beta_0, \Lambda_0)\right| = o_{p^*}(1),$$

and

$$\sup_{|\beta - \beta_0| \le \delta_n, \|\Lambda - \Lambda_0\| \le Cn^{-\gamma}} \left|\sqrt{n}(S_{2n} - S_2)(\beta, \Lambda)[\mathbf{h}^*] - \sqrt{n}(S_{2n} - S_2)(\beta_0, \Lambda_0)[\mathbf{h}^*]\right| = o_{p^*}(1).$$

A6. **(Smoothness of the model):** For some $\alpha > 1$ satisfying $\alpha\gamma > 1/2$, and for (β, Λ) in the neighborhood $\{(\beta, \Lambda) : |\beta - \beta_0| \le \delta_n, \|\Lambda - \Lambda_0\| \le Cn^{-\gamma}\}$,

$$\begin{aligned}\Big|S_1(\beta, \Lambda) - S_1(\beta_0, \Lambda_0) - \dot{S}_{11}(\beta_0, \Lambda_0)(\beta - \beta_0) - \dot{S}_{12}(\beta_0, \Lambda_0)[\Lambda - \Lambda_0]\Big| \\ = o(|\beta - \beta_0|) + O(\|\Lambda - \Lambda_0\|^\alpha),\end{aligned}$$

$$\begin{aligned}\Big|S_2(\beta, \Lambda)[\mathbf{h}^*] - S_2(\beta_0, \Lambda_0)[\mathbf{h}^*] - \dot{S}_{21}(\beta_0, \Lambda_0)[\mathbf{h}^*])(\beta - \beta_0) - (\dot{S}_{22}(\beta_0, \Lambda_0)[\mathbf{h}^*, \Lambda - \Lambda_0]\Big| \\ = o(|\beta - \beta_0|) + O(\|\Lambda - \Lambda_0\|^\alpha).\end{aligned}$$

A7. **(Asymptotic normality of projected pseudo-score):** With

$$m^*(\beta_0, \Lambda_0; x) \equiv m_1(\beta_0, \Lambda_0; x) - m_2(\beta_0, \Lambda_0; x)[\mathbf{h}^*],$$

we have

$$\sqrt{n}\mathbb{P}_n m^*(\beta_0, \Lambda_0; X) \longrightarrow_d N(0, B),$$

where $B = Em^*(\beta_0, \Lambda_0; X)^{\otimes 2} = Em^*(\beta_0, \Lambda_0; X)m^*(\beta_0, \Lambda_0; X)'$.

Theorem 2.3.5. (Asymptotic Normality) Suppose that: assumptions A1-A7 hold. Then

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = A^{-1}\sqrt{n}\mathbb{P}_n m^*(\beta_0, \Lambda_0; X) + o_{p^*}(1) \longrightarrow_d N\left(0, A^{-1}B\left(A^{-1}\right)'\right).$$

Proof : A1 and A5 yield

$$\sqrt{n}(S_{1n} - S_1)(\hat{\beta}_n, \hat{\Lambda}_n) - \sqrt{n}(S_{1n} - S_1)(\beta_0, \Lambda_0) = o_{p^*}(1).$$

Since $S_{1n}(\hat{\beta}_n, \hat{\Lambda}_n) = o_{p^*}(n^{-1/2})$ by A4 and $S_1(\beta_0, \Lambda_0) = 0$ by A2, it follows that

$$\sqrt{n}S_1(\hat{\beta}_n, \hat{\Lambda}_n) + \sqrt{n}S_{1n}(\beta_0, \Lambda_0) = o_{p^*}(1).$$

Similarly, we have that

$$\sqrt{n}S_2(\hat{\beta}_n, \hat{\Lambda}_n)[\mathbf{h}^*] + \sqrt{n}S_{2n}(\beta_0, \Lambda_0)[\mathbf{h}^*] = o_{p^*}(1).$$

Combining these equalities and A6 yields

$$\begin{aligned}\dot{S}_{11}(\beta_0, \Lambda_0)[\hat{\beta}_n - \beta_0] + \dot{S}_{12}(\beta_0, \Lambda_0)[\hat{\Lambda}_n - \Lambda_0] + S_{1n}(\beta_0, \Lambda_0) \\ +o(|\hat{\beta}_n - \beta_0|) + O(\|\hat{\Lambda}_n - \Lambda_0\|^\alpha) = o_{p^*}(n^{-1/2}),\end{aligned} \tag{3}$$

$$\begin{aligned}\dot{S}_{21}(\beta_0, \Lambda_0)[\mathbf{h}^*][\hat{\beta}_n - \beta_0] + \dot{S}_{22}(\beta_0, \Lambda_0)[\mathbf{h}^*][\hat{\Lambda}_n - \Lambda_0] + S_{2n}(\beta_0, \Lambda_0)[\mathbf{h}^*] \\ +o(|\hat{\beta}_n - \beta_0|) + O(\|\hat{\Lambda}_n - \Lambda_0\|^\alpha) = o_{p^*}(n^{-1/2}).\end{aligned} \tag{4}$$

Because $\alpha\gamma > 1/2$, then the rate of convergence assumption 1 implies

$$\sqrt{n}O(\|\hat{\Lambda}_n - \Lambda_0\|^\alpha) = o_{p^*}(1).$$

Thus by A4 and (2.3.4) minus (2.3.5), it follows that

$$\begin{aligned}(\dot{S}_{11}(\beta_0, \Lambda_0) - \dot{S}_{21}(\beta_0, \Lambda_0)[\mathbf{h}^*](\hat{\beta}_n - \beta_0) + o(|\hat{\beta}_n - \beta_0|) \\ = -(S_{1n}(\beta_0, \Lambda_0) - S_{2n}(\beta_0, \Lambda_0)[\mathbf{h}^*]) + o_{p^*}(n^{-1/2}),\end{aligned}$$

i.e.

$$-(A+o(1))(\hat{\beta}_n-\beta_0)=-\mathbb{P}_n m^*(\beta_0,\Lambda_0;X)+o_{p^*}(n^{-1/2}).$$

Hence

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n-\beta_0) &= (A+o(1))^{-1}\sqrt{n}\mathbb{P}_n m^*(\beta_0,\Lambda_0;X)+o_{p^*}(1)\\ &\to_d N\left(0,A^{-1}B\left(A^{-1}\right)'\right).\end{aligned}$$

□

GMM-theorems: Hansen, Pakes and Pollard, Newey

Suppose that $\Theta\subset\mathbb{R}^p$ and that $\mathbb{G}_n:\Theta\to\mathbb{R}^q$ is a vector of random functions with expected or limiting values $G:\Theta\to\mathbb{R}^q$ satisfying $G(\theta_0)=0$. Pakes and Pollard (1989) study estimators $\hat{\theta}_n$ of θ defined by

$$\hat{\theta}_n=\text{argmin}_{\theta\in\Theta}\|\mathbb{G}_n(\theta)\|$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^q$. They also study estimators based on minimizing quadratic forms; i.e. with the Euclidean norm replaced by

$$\|x\|_A^2\equiv\|Ax\|^2=\langle Ax,Ax\rangle=\langle x,A^TAx\rangle=x^TWx$$

where $W\equiv A^TA$ and where the nonsingular matrix A may be random and may depend on θ. This case is treated in a separate step. After studying consistency of such estimators separately, Pakes and Pollard (1989) prove the following theorem concerning their "Generalized Method of Moments" estimator $\hat{\theta}_n$.

Theorem 2.4.1. (Asymptotic Normality of GMM estimators) Suppose that:

A1. $\hat{\theta}_n\to_p\theta_0$ with $G(\theta_0)=0$ in $\mathbb{R}^q$. Also assume that

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

A2. G is differentiable at θ_0 with derivative matrix $\Gamma=\dot{G}$ $(p\times q)$.

A3. For every $\delta_n\to 0$

$$\sup_{\|\theta-\theta_0\|\le\delta_n}\frac{\sqrt{n}\|\mathbb{G}_n(\theta)-G(\theta)-(\mathbb{G}_n(\theta_0)-G(\theta_0))\|}{1+\sqrt{n}(\|\mathbb{G}_n(\theta)\|+\|G(\theta)\|)}=o_p(1).$$

A4. $\sqrt{n}\mathbb{G}_n(\theta_0)\to_d\mathbb{Z}\sim N_q(0,V)$.

A5. θ_0 is an interior point of θ.

Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d (\Gamma^T\Gamma)^{-1}\Gamma^T\mathbb{Z} \sim N_p(0, (\Gamma^T\Gamma)^{-1}(\Gamma^T V\Gamma)(\Gamma^T\Gamma)^{-1}).$$

Now suppose that $\{A_n(\theta) : \theta \in \Theta\}$ is a family of nonsingular random matrices for which there is a nonsingular, nonrandom matrix A such that

$$\sup_{\|\theta-\theta_0\|\leq\delta_n} \|A_n(\theta) - A\| = o_p(1). \tag{5}$$

whenever δ_n is a sequence of positive numbers that converges to zero. If conditions A2, A3, and A4 of Theorem 2.4.1 are satisfied by $\mathbb{G}_n$ and G, then they are also satisfied if $\mathbb{G}_n$ is replaced by $A_n(\theta)\mathbb{G}_n(\theta)$, $G(\theta)$ is replaced by $AG(\theta)$, V is replaced by AVA^T, and Γ is replaced by $A\Gamma = A\dot{G}$. Thus we have the following corollary of Theorem 2.4.1:

Corollary: Suppose that the hypothesis A1 of Theorem 2.4.1 is replaced by: $\hat{\theta}_n \to_p \theta_0$ where $AG(\theta_0) = 0$ and $\hat{\theta}_n$ satisfies

$$\|\tilde{\mathbb{G}}_n(\hat{\theta}_n)\| = o_p(n^{-1/2}) + \inf_\theta \|\tilde{\mathbb{G}}_n(\theta)\|$$

with $\tilde{\mathbb{G}}_n(\theta) \equiv A_n(\theta)\mathbb{G}_n(\theta)$. Suppose further that A2-A5 of Theorem 2.4.1 hold and the matrices $A_n(\theta)$ satisfy (5) with A nonsingular. Then, with $W \equiv A^TA$,

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) &\to_d ((A\Gamma)^T(A\Gamma))^{-1}(A\Gamma)^TA\mathbb{Z} = (\Gamma^TW\Gamma)^{-1}\Gamma^TW\mathbb{Z} \\ &\sim N_p(0, (\Gamma^TW\Gamma)^{-1}(\Gamma^TWVW\Gamma)(\Gamma^TW\Gamma)^{-1}).\end{aligned}$$

Remark 1: The asymptotic covariance appearing in the corollary agree with the asymptotic covariance of the GMM estimators studied by Hansen (1982) and generalized by Newey (1994) to handle nuisance parameters. Note that it is of the "sandwich" form $C^{-1}BC^{-1}$ for matrices B and C.

Remark 2: Note that if $W = V^{-1}$, then the covariance matrix in the corollary becomes simply $(\Gamma^TV^{-1}\Gamma)^{-1}$, and in fact this is the minimal value over choices of A (or W) as has been noted by many authors. This is also exactly the form of the covariance of Empirical Likelihood and Generalized Empirical Likelihood estimators as shown by Qin and Lawless (1994), Newey and Smith (2004), and others under stronger regularity conditions.

Remark 3: Chamberlain (1987) shows that $(\Gamma^TW\Gamma)^{-1}$ is the efficiency bound for estimation of θ in the constraint-defined model $\mathcal{P} = \{P : G(\theta) = 0, \theta \in \mathbb{R}^p\}$. (Alternative proof via BKRW methods?) Note that since the dimension p of θ is

smaller than q, the dimension of the vector G of constraints, $\mathcal{P}$ is a proper subset of the family of all distributions P on $\mathcal{X}$ (at least under a non-degeneracy assumption on $\{G(\theta) : \theta \in \Theta\}$).

Now our goal is to develop an analogue of the theorem of Pakes and Pollard (1989) for the empirical likelihood estimators of Qin and Lawless (1994). A start in this direction has been given by Lopez et al. (2009); these authors establish a likelihood ratio type limit theorem without imposing smoothness conditions on the functions g involved in the constraints. To do this they use methods due to Sherman (1993) and Pollard (1989).

Further questions:

1. Behavior of the (generalized) empirical likelihood ratio statistics and estimators under local alternatives?

2. Behavior of the (generalized) empirical likelihood statistics and estimators under fixed alternatives?

3. Infinite-dimensional version of empirical likelihood (analogous to infinite-dimensional $Z-$theorem? Can we handle infinite-dimensional constraints of the type "known marginal(s)"?

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag, New York. Reprint of the 1993 original.

Breslow, N., McNeney, B. and Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Ann. Statist.* **31** 1110–1139.

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 nuisance parameters and correction note for: "Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression" [Scand. J. Statist. **34** (2007), no. 1, 86–102; mr2325244]. *Scand. J. Statist.* **35** 186–192.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics* **34** 305–334.

Gilbert, P. B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Ann. Statist.* **28** 151–194.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054.

Hjort, N. L., McKeague, I. W. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.* **37** 1079–1111.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24** 540–568.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*. Univ. California Press, Berkeley, Calif., 221–233.

Knight, K. (2000). *Mathematical statistics*. Chapman & Hall/CRC Texts in Statistical Science Series, Chapman & Hall/CRC, Boca Raton, FL.

Lopez, E. M. M., van Kelegom, I. and Veraverbeke, N. (2009). Empirical likelihood for non-smooth criterion functions. *Scand. J. Statist.* **36** 413–432.

Molanes Lopez, E. M., Van Keilegom, I. and Veraverbeke, N. (2009). Empirical likelihood for non-smooth criterion functions. *Scand. J. Stat.* **36** 413–432.

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *Ann. Statist.* **23** 182–198.

Murphy, S. A. and van der Vaart, A. W. (1997). Semiparametric likelihood ratio inference. *Ann. Statist.* **25** 1471–1509.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62** 1349–1382.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72** 219–255.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57** 1027–1057.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist.* **26** 183–214.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300–325.

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *Ann. Statist.* **35** 634–672.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61** 123–137.

van der Vaart, A. W. (1995). Efficiency of infinite-dimensional M-estimators. *Statist. Neerlandica* **49** 9–30.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics, Springer-Verlag, New York. With applications to statistics.

van der Vaart, A. W. and Wellner, J. A. (2007). Empirical processes indexed by estimated functions. In *Asymptotics: particles, processes and inverse problems*, vol. 55 of *IMS Lecture Notes Monogr. Ser.* Inst. Math. Statist., Beachwood, OH, 234–252.

Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* **28** 779–814.

Wellner, J. A. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Ann. Statist.* **35** 2106–2142.