SOME SIMPLE EXAMPLES

Semiparametric RIT

Spring 2009

Paul J. Smith

OUTLINE

- Nonparametric model.
- Symmetric location problem.
- Regression.

NONPARAMETRIC MODELS

Let $X_1, \ldots, X_n$ be i.i.d. real r.v.'s with distribution $P \in \mathcal{P} =$ the set of all distributions on R, and let $\widehat{F}_n(x) = (1/n)\sum_i I\{X_i \leq x\}$ be the empirical cdf.

The tangent space $\dot{\mathcal{P}}_P$ is

$$\left\{ g \in L_2(P) \,\middle|\, \int g(x)dP(x) = 0 \right\}.$$

Proof. If g is bounded, consider submodels $p_t(x) = (1 + tg(x))p_0(x)$ or the exponential family $p_t(x) = c(t)\exp[tg(x)]p_0(x)$. If g is unbounded, consider $p_t(x) = c(t)k(tg(x))p_0(x)$ with $k(u) = 2/(1 + e^{-2u})$.

In each submodel $g(x) = (\nabla/\nabla t)|_{t=0} \log p_t(x)$.

EFFICIENCY OF EMPIRICAL DISTRIBUTION

Let $\psi(P) = \int h(x)dP$ where $E[h(X)^2] < \infty$. Then $\widehat{\psi}_n = (1/n)\sum_i h(X_i)$ is an asymptotically efficient estimator of $\psi(P)$.

We need $\dot{\psi}_P$. If $p_t = (1+tg)p$ and g is bounded, then $\psi(P_t) = E[h(X)] + tE[h(X)g(X)]$. Hence $\dot{\psi}_P g = E[h(X)g(X)]$ and the efficient influence function is $\tilde{\psi}(x) = h(x) - \int h(x)dP$. Unbounded g are handled by limiting arguments.

The optimal asymptotic variance of any regular estimator of $\psi(P)$ is $E[\tilde{\psi}(X)^2]$. But this is Var $[\sqrt{n}(\widehat{\psi}_n - \psi(P))]$.

SYMMETRIC LOCATION MODEL

Let $X_1, \ldots, X_n$ be i.i.d. with density $\eta(x-\theta)$ where η is an even function and one wishes to estimate $\theta \in R$. We consider submodels (θ_t, η_t) where $\theta_t = \theta + at$ and each θ_t is even.

Since η is even, $\eta_t(x-\theta) = \eta_t(|x-\theta|)$. Hence score functions have the form

$$g(x) = a\eta'(x-\theta)/\eta(x-\theta) + b(|x-\theta|).$$

Proof. Compute $(\nabla/\nabla t) \log \eta_t(x-\theta)|_{t=0}$. Bickel et al. (1998) prove all scores take this form. Note that b is a nuisance score, i.e., a score for η if θ is fixed.

The efficient score for estimating θ is $\dot{\ell} = \eta'/\eta$ when η is known. If η is unknown, efficient score is $\tilde{\ell} = \dot{\ell} - \Pi_{\theta,\eta}\dot{\ell}$ where $\Pi_{\theta,\eta}\dot{\ell}$ is projection of $\dot{\ell}$ into space of nuisance scores.

EFFICIENT LOCATION ESTIMATES

If η is even, η' is odd. Hence

$$E[(\eta'/\eta)(X-\theta)b(|X-\theta|)] = 0$$

so that $\dot{\ell} = \tilde{\ell}$. Thus information about θ is the same whether η is known or unknown!

Find $\tilde{\theta}_n$, a consistent estimator of θ. Let $g(s) = 2\eta(s)I\{s > 0\}$. Construct $T_i = |X_i - \tilde{\theta}_n|$, $i = 1, \ldots, n$, and let $\hat{g}_n$ be a kernel density estimator based on a smooth density ω and $T_1, \ldots, T_n$:

$$\begin{aligned}\hat{g}_n(s) &= (1/(nb_n))\sum_i \omega[(s-T_i)/b_n], \\ \hat{k}_n(s) &= (\hat{g}_n'(s)/\hat{g}_n(s))I\{s \in C_n\}.\end{aligned}$$

If $s \in C_n$, $\hat{g}_n'(s)/\hat{g}_n(s)$ is "well behaved." Then $\hat{\theta}_n$, the solution of $\sum_i \hat{k}_n(X_i - \theta) = 0$ is efficient.