Profile Likelihood

Wenyi Xie

University of North Carolina at Chapel Hill

November 11, 2021

The main purpose of this chapter is to establish efficient semi-parametric inference for finite-dimensional parameters.

From previous talk, we have introduced efficient score functions and estimating equations and their connection to the efficient estimation.

Next we move on to introduce the main tool for constructing efficient estimators, which is based on modifications of maximum likelihood estimation.

Semi-parametric Model and Likelihood Modification

We focus on semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, where Θ is an open subset of $\mathcal{R}^k$ and H is an arbitrary, possibly infinite-dimensional set.

The parameter of interest is $\psi(P_{\theta,\eta}) = \theta$

The most common approach to efficient estimation is based on modifications of maximum likelihood estimation which lead to efficient estimation.

Such modified likelihood for semi-parametric model are generally not really likelihoods (products of densities) due to the presence of an infinite dimensional nuisance parameter η

Recall the setting of estimation of an unknown real density $f(x)$ from an i.i.d samples $X_1, \ldots, X_n$

The likelihood is $\prod_{i=1}^n f(X_i)$ and the maximizer over all densities has arbituarily high peaks at the observations, and zeros at other values, and therefore is not a density.

This can be fixed by using an empirical likelihood $\prod_{i=1}^n p_i$, where $p_1, \ldots, p_n$ are the masses assigned to the observations, and $\sum_{i=1}^n p_i = 1 \Rightarrow$ empirical distribution function estimator, which is known to be fully efficient.

Consider again the Cox model for the right-censored data.

We observe a sample of n realizations of $X = (V, d, Z)$, where $V = T \wedge C,\ d = 1\{V = T\},\ Z \in \mathcal{R}^k$ is covariate vector.

T is the failure time, and C is a censoring time.

We assume that T and C are independent given Z, and T given Z has integrated hazard function $\exp(Z^T\beta)\Lambda(t)$ for β in an open subset $B \subset \mathbb{R}^k$ and Λ is continuous and monotone increasing with $\Lambda(0) = 0$.

Censoring is not informative.

The density for single observation is proportional to

$$\left[e^{\beta' Z}\lambda(V)\right]^d \exp\left[-e^{\beta' Z}\Lambda(V)\right]$$

.

Maximizing such likelihood based on this density will results in the same issue as previous slide.

A likelihood works assigns mass only at observed failure time

$$L_n(\beta, \Lambda) = \prod_{i=1}^{n} \left[e^{\beta' Z_i} \Delta\Lambda\left(V_i\right) \right]^{d_i} \exp\left[-e^{\beta' Z_i} \Lambda\left(V_i\right) \right]$$

where $\Delta\Lambda(t)$ is the jump size of Λ at t

For each value of β, one can maximize or profile $L_n(\beta, \Lambda)$ over the nuisance parameter Λ to obtain profile likelihood $pL_n(\beta)$, which is proportional to partial likelihood.

Let $\hat{\beta}$ be the maximizer of $pL_n(\beta)$, then the maximizer $\hat{\Lambda}$ of $L_n(\hat{\beta}, \Lambda)$ is the Breslow estimator

$$\hat{\Lambda}(t) = \int_0^t \frac{\mathbb{P}_n dN(s)}{\mathbb{P}_n \left[Y(s) e^{\hat{\beta}' Z} \right]}$$

It can be shown that both $\hat{\beta}$ and $\hat{\Lambda}$ are efficient

Other Modifications to Maximum Likelihood Estimation

Another useful class of likelihood variants are penalized likelihoods. Penalty term (or terms) adds to likelihoods in order to maintain an appropriate level of smoothness for one or more of the nuisance parameters.

Other methods of generating likelihood variants are possible.

The basic idea is that using the likelihood principle to guide estimation of semiparametric models often lead to efficient estimators for the model components that are $\sqrt{n}$ consistent.

Because of the richness of this approach to estimation, one needs to verify for each new situation that a likelihood-inspired estimator is consistent, efficient and well-behaved for moderate sample sizes.

Verifying efficiency usually entails demonstrating that the estimator satisfies the efficient score equation from presentation last week.

Profile Likelihood

The Cox profile likelihood can be found in closed form by analytic methods.

However, it is not clear that a general profile likelihood is differentiable, because the supremum of differentiable functions is not necessarily differentiable itself.

Therefore, we want to find some submodels such that the efficient score is a derivative of the log likelihood along those parametric submodels.

Least Favorable Submodel

Among all parametric submodels, one can find the minimum of the information over all efficient estimators.

For semiparametric models, this minimum information is the best possible because a nonparametric problem is at least as difficult as any finite-dimensional subproblem.

In other words, the Fisher information for estimating parameters of interest in semiparametric problem should be no greater than the Fisher information for estimating those parameters in any finite-dimensional problem.

The least favorable submodel is a parametric model that achieves this minimum.

By looking at the least favorable submodel, we may obtain a best possible asymptotic variance of the estimator in the original semiparametric problem.

Least favorable submodels

The basic idea of using least favorable submodels is to connect semiparametric model to a fully parametric problem and reduce the infinite-dimensional problem to a problem involving the finite-dimensional fully-parametric "least favorable submodel".

Those submodels are of the same dimension as θ. Once such reduction is made, classical Cramer conditions on the low-dimensional model can be used.

Approximately Least Favorable submodels

Recall that given the set of scored for the nuisance parameters, the efficient score function for θ at the truth (θ_0, η_0) is defined as

$$\tilde{l}_{\theta_0,\eta_0} = l_{\theta_0,\eta_0} - \prod_{\theta_0,\eta_0} l_{\theta_0,\eta_0}$$

where l_{θ_0,η_0} is the score functions for θ, and $\prod_{\theta_0,\eta_0} l_{\theta_0,\eta_0}$ is the projection of the score function for θ onto the closed linear span of the nuisance scores.

While the solution of an efficient score equation need not be a maximum likelihood estimator, it is also possible that the maximum likelihood estimator in a semiparametric model may not be expressible as the zero of an efficient score equation

Approximately Least Favorable submodels

This possibility occurs because the efficient score is a projection, thus there is no assurance that this projection is the derivative of the log-likelihood along a submodel.

Furthermore, since such projection is not necessarily a nuisance score itself, the existence of the efficient score does not imply existence of a least favorable submodel.

We assume the existence of an approximately least favorable submodel which approximates the true least favorable submodel to a useful level of accuracy that facilitates analysis of semiparametric estimators.

Appriximately Least Favorable Submodels

We will now describe the process in generality, while the specifics will depend on the situation.

Assume that for each parameter (θ, η), there exists a map $t \mapsto \eta_t(\theta, \eta)$ that maps from a fixed neighborhood of θ into the parameter set for η.

We require that

$$\eta_t(\theta, \eta) \in \hat{H}, \text{ for all } \|t - \theta\| \text{ small enough, and}$$
$$\eta_\theta(\theta, \eta) = \eta \text{ for any } (\theta, \eta) \in \Theta \times \hat{H}$$

where $\hat{H}$ is a suitable enlargement of H that includes all estimators that satisfy the constraints of the estimation process.

Define the map $t \mapsto \ell(t, \theta, \eta)(x)$ which is defined by $\ell(t, \theta, \eta)(x) = \log l(t, \eta_t(\theta, \eta))(x)$, which is twice continuously differentiable for all x.

Approximately Least Favorable Submodels

We will make additional requirement of $\ell(\cdot,\cdot,\cdot)$ at various points, leading to further restriction on $\eta_t(\theta,\eta)$

Denote $\dot{\ell}(t,\theta,\eta)(x) \equiv \frac{\partial}{\partial t}\ell(t,\theta,\eta)$

Another important structural requirement for such submodel is that it is least favorable at (θ_0,η_0) for estimating θ:

$$\dot{\ell}(\theta_0,\theta_0,\eta_0) = \tilde{l}_{\theta_0,\eta_0}$$

Meaning that for the model with likelihood $l(t,\eta_t(\theta_0,\eta_0))$, the score function for the parameter t at $t=\theta_0$ is the efficient score function for θ.

Note that we assume this only at the (θ_0,η_0) in particular, but not at every realization of the profile estimators.

Check MLE efficiency

$\left(\hat{\theta}_n, \hat{\eta}_n\right)$ are the maximum likelihood estimate, i.e maximizer of $\mathcal{P}_n \log l(\theta, \eta)$.

Clearly, we have $\mathcal{P}_n \dot{\ell}(\hat{\theta}_n, \hat{\theta}_n, \hat{\eta}_n) = 0$

Therefore, provided that $\dot{\ell}\left(\theta, \theta, \hat{\eta}_n\right)$ satisfies the conditions of Theorem 3.1 from last week, $\hat{\theta}_n$ is efficient.

It is necessary to check the conditions even for maximum likelihood estimators because $\hat{\eta}_n$ is often on the boundary (or even a little bit outside) of the parameter space.

Check MLE efficiency

To see this, recall again the Cox model setting for right-censored data. Nuisance parameter η is the baseline integrated hazard function which is usually assumed to be continuous.

However, $\hat{\eta}_n$ is the Breslow estimator, which is right-continuous step function that jumps at observed failure time. Thus not in the parameter space.

Therefore, direct differentiation of the log-likelihood at maximum likelihood estimator will not yield an efficient score equation.

The approximately least-favorable submodel structure is very useful for developing methods of inference for θ. Next we illustrate this concept by an example on right censored Cox model.

Example1: Right Censored Cox Model

H consists of all monotone increasing functions $\Lambda \in C[0, \tau]$ with $\Lambda(0) = 0$.
$\hat{H}$ is the set of all monotone, increasing functions $\Lambda \in D[0, \tau]$
From previous week, we showed that the efficient score for β is

$$\tilde{\ell}_{\beta,\Lambda} = \int_0^\tau (Z - h_0(s))\, dM(s)$$

where

$$M(t) \equiv N(t) - \int_0^t Y(s) e^{\beta' Z} d\Lambda(s)$$

N and Y are the usual counting and at-risk processes respectively, and

$$h_0(t) \equiv \frac{P\left[Z \mathbf{1}\{W \geq t\} e^{\beta_0' Z}\right]}{P\left[\mathbf{1}\{W \geq t\} e^{\beta_0' Z}\right]}$$

where P is the true probabilty measure (at the parameter values (β_0, Λ_0))

Example1: Right Censored Cox Model Cont'

The Cox model log-likelihood for a single observation is

$$\log l(\beta, \Lambda) = \left(\beta' Z + \log \Delta\Lambda(W)\right) \delta - e^{\beta' Z} \Lambda(W)$$

where $\Delta\Lambda(w)$ is the jump size in Λ at w
A convenient approximately least favorable submodel is defined by

$$d\Lambda_t(\beta, \Lambda) = \left(1 + (\beta - t)' h_0(s)\right) d\Lambda(s)$$

We can verify that

$$(1) \quad d\Lambda_\beta(\beta, \Lambda) = d\Lambda$$
$$(2) \quad \dot{\ell}\left(\beta_0, \beta_0, \Lambda_0\right) = \tilde{l}_{\beta_0, \Lambda_0}$$

Inference

A few methods exists for efficient estimation and inference for θ using the profile likelihood. We will introduce one method and leave the rest til later.

The first method is based on the very important results that under reasonable regularity conditions, a profile likelihood for θ behaves asymptotically like a parametric likelihood of a normal random variable with variance being the inverse of the efficient Fisher information $\tilde{I}_{\theta,\eta}$ in a shinking neighborhood of the maximum likelihood estimator $\hat{\theta}$.

This leads to a valid likelihood ratio based inference for θ.

The idea is based on quadratic expansion of the profile likelihood.

uadratic expansion of the profile likelihood

Still the context is we have maximum likelihood estimators $(\hat{\theta}_n, \hat{\eta}_n)$ from i.i.d samples. θ is finite-dimensional parameter of primary interest, and η is an infinite-dimensional nuisance parameter η.

The main results give the following asymptotic expansion of the profile likelihood.

Under certain regularity conditions, we have that for any random sequence $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$,

$$\begin{aligned}\log pL_n\left(\tilde{\theta}_n\right) = &\log pL_n\left(\theta_0\right) + \left(\tilde{\theta}_n - \theta_0\right)' \sum_{i=1}^{n} \tilde{\ell}_{\theta_0,\eta_0}\left(X_i\right) \\ &- \frac{1}{2} n \left(\tilde{\theta}_n - \theta_0\right)' \tilde{I}_{\theta_0,\eta_0} \left(\tilde{\theta}_n - \theta_0\right) \\ &+ o_{P_0}\left(1 + \sqrt{n}\left\|\tilde{\theta}_n - \theta_0\right\|\right)^2\end{aligned}$$

where $\tilde{\ell}_{\theta_0,\eta_0}$ is the efficient score function for θ, $\tilde{I}_{\theta_0,\eta_0}$ is the efficient Fisher information matrix, and P_0 is the probability measure of X at the true parameter values.

Corollary 1

If the aymptotic expansion of the profile likelihood holds, $\tilde{I}_{\theta_0,\eta_0}$ is positive definite, and $\hat{\theta}_n$ is consistent, i.e $\hat{\theta}_n = \theta_0 + o_{P_0}(1)$
Then MLE is aymptotically normal and has asymptotic expansion

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \sqrt{n}\mathbb{P}_n \tilde{I}^{-1}_{\theta_0,\eta_0} \tilde{\ell}_{\theta_0,\eta_0}(X) + o_{P_0}(1)$$

Log profile likelihood function can be expanded around $\hat{\theta}_n$ in the form

$$\begin{aligned}\log pL_n\left(\tilde{\theta}_n\right) =& \log pL_n\left(\hat{\theta}_n\right) - \frac{1}{2} n \left(\tilde{\theta}_n - \hat{\theta}_n\right)' \tilde{I}_{\theta_0,\eta_0} \left(\tilde{\theta}_n - \hat{\theta}_n\right) \\ &+ o_{P_0}\left(1 + \sqrt{n}\left\|\tilde{\theta}_n - \theta_0\right\|\right)^2\end{aligned}$$

Corollary 1 justifies using a semiparametric profile likelihood as an ordinary likelihood, at least asmptotically.

Proof

Set $\Delta_n = n^{-1/2} \sum_{i=1}^n \tilde{l}_0(X_i)$ and $\hat{h} = \sqrt{n}\left(\hat{\theta} - \theta_0\right)$

Then apply the main results with the choice $\tilde{\theta} = \hat{\theta}$ and $\tilde{\theta} = \theta_0 + n^{-1/2}\tilde{I}_0^{-1}\Delta_r$

We have

$$\log \mathrm{pl}_n(\hat{\theta}) = \log \mathrm{pl}_n(\theta_0) + \hat{h}^T \Delta_n - \frac{1}{2}\hat{h}^T \tilde{I}_0 \hat{h} + o_P(\|\hat{h}\| + 1)^2$$

and

$$\log \mathrm{pl}_n\left(\theta_0 + n^{-1/2}\tilde{I}_0^{-1}\Delta_n\right) = \log \mathrm{pl}_n(\theta_0) + \Delta_n^T \tilde{I}_0^{-1}\Delta_n - \frac{1}{2}\Delta_n^T \tilde{I}_0^{-1}\Delta_n + o_P(1)$$

Proof for Corollary 1

By the definition of $\hat{\theta}_n$, $\log \mathrm{pl}_n(\hat{\theta}) \geq \log \mathrm{pl}_n \left(\theta_0 + n^{-1/2}\tilde{I}_0^{-1}\Delta_n\right)$

$$\hat{h}^T\Delta_n - \frac{1}{2}\hat{h}^T\tilde{I}_0\hat{h} - \frac{1}{2}\Delta_n^T\tilde{I}_0^{-1}\Delta_n \geq -o_P(\|\hat{h}\| + 1)^2$$

The left side of this inequality is equal to

$$-\frac{1}{2}\left(\hat{h} - \tilde{I}_0^{-1}\Delta_n\right)^T \tilde{I}_0 \left(\hat{h} - \tilde{I}_0^{-1}\Delta_n\right) \leq -c\left\|\hat{h} - \tilde{I}_0^{-1}\Delta_n\right\|^2$$

for a positive constant c (nonsingularity of $\tilde{I}_0$)
We can conclude that

$$\left\|\hat{h} - \tilde{I}_0^{-1}\Delta_n\right\| = o_P(\|\hat{h}\| + 1)$$

This implies that $\|\hat{h}\| = O_P(1)$, and $\left\|\hat{h} - \tilde{I}_0^{-1}\Delta_n\right\| = o_P(1)$

The following two additional corollaries provide methods of using this quadratic expansion to conduct inference for θ_0

Corollary 2

If the asymptotic expansion of log profile likelihood holds, and $\tilde{I}_0$ is positive-definite, and $\hat{\theta}_n$ is consistent
Then under the null hypothesis $H_0 : \theta = \theta_0$,

$$2\left(\log pL_n\left(\hat{\theta}_n\right) - \log pL_n\left(\theta_0\right)\right) \rightsquigarrow \chi^2(k)$$

This second corollary concerns the profile likelihood ratio statistics and shows that this behaves it should.
It justifies using the set

$$\left\{ \theta : 2 \log \frac{\mathrm{pl}_n\left(\hat{\theta}_n\right)}{\mathrm{pl}_n(\theta)} \leq \chi^2_{d,1-\alpha} \right\}$$

as a confidence set of approximate coverage probability $1 - \alpha$

The third corollary concerns a discretized second derivative of the profile likelihood.

Corollary 3

If the asymptotic expansion of log profile likelihood holds, and $\hat{\theta}_n$ is consistent, then for all sequences $v_n \xrightarrow{P} v \in \mathbb{R}^k$ and $h_n \xrightarrow{P} 0$ such that $(\sqrt{n}h_n)^{-1} = O_P(1)$,

$$-2\frac{\log \mathrm{pl}_n\left(\hat{\theta}_n + h_n v_n\right) - \log \mathrm{pl}_n\left(\hat{\theta}_n\right)}{nh_n^2} \xrightarrow{P} v^T \tilde{I}_0 v$$

This estimator is the square of numerical derivative of the signed log-likelihood ratio statistics as discussed by Chen and Jennrich (1996). In their theorem 3.1, they showed that in the parametric setting, such derivative is the square root of the observed information about θ.

Indeed, the first derivative of the profile likelihood at $\hat{\theta}_n$ is 0, the above expression can be taken as the observed information about θ (evaluated in the direction of v_n). This corollary can thus be used to construct consistent numerical estimates of $\tilde{I}_0$

Theorem 19.5 asymptotic expansion of log profile likelihood

In addition to three conditions for the approximately least-favorable submodel $t \mapsto \eta_t(\theta, \eta)$, which is

$$\begin{aligned}
&\eta_t(\theta, \eta) \in \hat{H}, \text{ for all } \|t - \theta\| \text{ small enough, and} \\
&\eta_\theta(\theta, \eta) = \eta \text{ for any } (\theta, \eta) \in \Theta \times \hat{H} \\
&\dot{\ell}\left(\theta_0, \theta_0, \eta_0\right) = \tilde{\ell}_{\theta_0, \eta_0}
\end{aligned}$$

Several other conditions need to be satisfied in order to expand the profile likelihood.
First define $\ddot{\ell}(t, \theta, \eta) = (\partial/(\partial t))\dot{\ell}(t, \theta, \eta)$, and $\hat{\eta}_\theta \equiv \operatorname{argmax}_\eta L_n(\theta, \eta)$.
Assume that for any possibly random sequence $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$, we have

$$\begin{aligned}
&\hat{\eta}_{\tilde{\theta}_n} \xrightarrow{\mathrm{P}} \eta \quad \text{and} \\
&P_0 \dot{\ell}\left(\theta_0, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n}\right) = o_{P_0}\left(\left\|\tilde{\theta}_n - \theta_0\right\| + n^{-1/2}\right)
\end{aligned}$$

asymptotic expansion of log profile likelihood

Theorem 19.5

Assume conditions on previous page are satisfied, and assume that

$$(t, \theta, \eta) \mapsto \dot{\ell}(t, \theta, \eta)(X)$$

and

$$(t, \theta, \eta) \mapsto \ddot{\ell}(t, \theta, \eta)(X)$$

are continuous at $(\theta_0, \theta_0, \eta_0)$ for P_0-almost every X (or in measure).

Furthermore assume that for some neighborhood V of $(\theta_0, \theta_0, \eta_0)$,

the class of functions $\mathcal{F}_1 \equiv \{\dot{\ell}(t, \theta, \eta) : (t, \theta, \eta) \in V\}$ is P_0-Donsker with square-integrable envelope function,

the class of functions $\mathcal{F}_2 \equiv \{\ddot{\ell}(t, \theta, \eta) : (t, \theta, \eta) \in V\}$ is P_0-Glivenko-Cantelli and bounded in $L_1(P_0)$

$$\begin{aligned}\log pL_n\left(\tilde{\theta}_n\right) = &\log pL_n\left(\theta_0\right) + \left(\tilde{\theta}_n - \theta_0\right)' \sum_{i=1}^{n} \tilde{\ell}_{\theta_0,\eta_0}\left(X_i\right) \\ &- \frac{1}{2} n \left(\tilde{\theta}_n - \theta_0\right)' \tilde{I}_{\theta_0,\eta_0} \left(\tilde{\theta}_n - \theta_0\right) \\ &+ o_{P_0}\left(1 + \sqrt{n}\left\|\tilde{\theta}_n - \theta_0\right\|\right)^2\end{aligned}$$

We can readily verify the conditions of Theorem 19.5 for several models

- Cox model for right-censored data
- Cox model for current status data
- proportional odds model under right-censoring
- partly-linear logistic regression model
- :
- :

Several other methods for θ estimation and inference are developed by extending the idea of such quadratic expansion of the profile likelihood.

The next section will show that slightly stronger assumption can yield even more powerful methods of inference.