Higher Order Semiparametric Frequentist Inference Based on the Profile Sampler

Guang Cheng
Institute of Statistics and Decision Sciences
Duke University

- Introduction
- Preliminaries
- Main Results
- Future Research Plan

Introduction

- Semiparametric Models
- The Profile Sampler
- Major Contributions

Semiparametric Models

Random Variable X is assumed to come from $\left\{P_{\theta,\eta} : \theta \in \Theta \subset R^k, \eta \in \mathcal{H}\right\}$

- θ is the Euclidean parameter of interest;
- η is an infinite dimensional nuisance parameter.

The **profile likelihood** for θ is defined as follows:

$$
\begin{aligned}
pl_n(\theta) &\equiv \sup_{\eta \in \mathcal{H}} lik_n(\theta, \eta), & (1) \\
\hat{\eta}_\theta &\equiv \arg\sup_{\eta} lik_n(\theta, \eta). & (2)
\end{aligned}
$$

The profile likelihood may be used to a considerable extend as full likelihood in the semiparametric models [Murphy and van der Vaart, 2000].

Example 1: The Cox model with right censored data

Observations: $X_1 = (Y_1, \delta_1, Z_1), \ldots, X_n = (Y_n, \delta_n, Z_n)$, i.i.d.

- $Y = T \wedge C$, $\delta = 1\{T \leq C\}$;
- T: failure time, C: censoring time;
- $Z \in \mathcal{Z}$: covariate.

$$\begin{aligned} lik(\theta, \eta) &= \exp(-e^{\theta' z}\eta(y))(e^{\theta' z}\eta\{y\})^{\delta}, && (3) \\ \log pl_n(\theta) &= \sum_{i=1}^{n} \delta_i(\theta' z_i - \log \sum_{j \in R_i} e^{\theta' z_j}), && (4) \\ \hat{\eta}_\theta(t) &= \sum_{\{y_i \leq t\}} \frac{\delta_i}{\sum_{j \in R_i} \exp(\theta' z_j)}. && (5) \end{aligned}$$

$\eta \in \mathcal{H}$, a set of nondecreasing cadlag functions on bounded subset. $\hat{\eta}_\theta(\cdot)$ is a nondecreasing step function with steps at the observed failure times.

Example 2: The proportional odds model with right censored data

The odds ratio of the survival function ($S_Z(u)$) for subjects with different covariates is independent of time, i.e.

$$
\begin{aligned}
& -\log\left(\frac{S_Z(u)}{1-S_Z(u)}\right) = \log\Lambda(u) + \theta' Z. \\
\Longrightarrow \qquad & lik(\theta,\Lambda) = \left[\frac{e^{-\theta' z}\Lambda\{y\}}{(\Lambda(y)+e^{-\theta' z})(\Lambda(y-)+e^{-\theta' z})}\right]^{\delta}\left[\frac{e^{-\theta' z}}{\Lambda(y)+e^{-\theta' z}}\right]^{1-\delta}
\end{aligned}
$$

In this example, $\hat{\Lambda}_\theta(t)$ and $\log pl_n(\theta)$ have no explicit forms.

The Profile Sampler

The profile sampler is the MCMC chain generated based on the posterior distribution of the profile likelihood given $\tilde{X} = (X_1, \ldots, X_n)$.

$$\rho(\theta) \longrightarrow pl_n(\theta).$$

Proved to be a **first order** frequentist valid procedure:

[Lee, Kosorok and Fine, 2005]

- Posterior mean: $\tilde{E}_{\theta|\tilde{X}}(\theta) = \hat{\theta}_n + o_P(n^{-1/2})$;
- Inverse of posterior variance: $\left(n\tilde{Var}_{\theta|\tilde{X}}(\theta)\right)^{-1} = \tilde{I}_0 + o_P(1)$, where $\tilde{I}_0$ is the efficient information matrix;
- Simulation Evidence.

Advantages of the Profile Sampler:

- Traditional semiparametric estimation methods rely on the use of an infinite dimensional operator;
- The profile sampler method is an automatic estimation procedure;
- The profile sampler is easy to generate:
  - Profiling $\Longleftarrow$ Iterative Convex Minorant Algorithm;
  - MCMC chain $\Longleftarrow$ Metropolis Algorithm;
- No prior on η.

Major Contributions

- Motivation: The Cox model with right censored data; The Cox model with current status data (Event time is not observed).
- The profile sampler procedure essentially produces second order frequentist valid inference of θ in terms of distributions, moments and quantiles; [Cheng and Kosorok, 2006a]
- The relation between the convergence rate of the nuisance parameter and estimation accuracy of the profile sampler; [Cheng and Kosorok, 2006b]
- Control the estimation accuracy through the penalized profile sampler method. [Cheng and Kosorok, 2006c]

Preliminaries

- The least favorable submodel
- Notations
- Assumptions

The Least Favorable Submodel

- $t \mapsto p_{t,\eta_t(\theta,\eta)}$ is called the submodel of $\{p_{\theta,\eta} : \theta \in \Theta, \eta \in \mathcal{H}\}$;
- **The least favorable submodel** (LFS) is the closest parametric submodel to the semiparametric models in the sense of information, i.e. $\tilde{\ell}_{\theta,\eta} = (\partial/\partial t) \log p_{t,\eta_t}$, given $t = \theta$;
- LFS can be viewed as an estimator of the profile likelihood in semiparametric models for the estimation of θ;
- Log-likelihood of the LFS is written as:

$$\ell(t, \theta, \eta) = \log lik(t, \eta_t(\theta, \eta)). \tag{6}$$

Example 1 (cont'): The Cox model with right censored data

- $\ell(t, \theta, \eta) = \log lik(t, \eta_t(\theta, \eta))$;
- $d\eta_t(\theta, \eta) = (1 + (\theta - t)' h_{\theta_0, \eta_0}) d\eta$,

where h_{θ_0, η_0} is the least favorable direction at (θ_0, η_0). By some analysis, we can establish:

$$h_{\theta,\eta}(y) = \frac{E_{\theta,\eta}(e^{\theta' Z} Z 1\{Y \geq y\})}{E_{\theta,\eta}(e^{\theta' Z} 1\{Y \geq y\})}.$$

Example 2 (cont'): The proportional odds model with right censored data

- $\ell(t, \theta, \Lambda) = \log lik(t, \Lambda_t(\theta, \Lambda))$;
- $d\Lambda_t(\theta, \Lambda) = (1 + (\theta - t)' h_{\theta_0, \Lambda_0}) d\Lambda$,

where h_{θ_0,Λ_0} is the least favorable direction at (θ_0, Λ_0). By some analysis, we can establish:

$$h_{\theta,\Lambda}(y) = (A^*_{\theta,\Lambda} A_{\theta,\Lambda})^{-1} A^*_{\theta,\Lambda} \ell_{\theta,\Lambda}.$$

Notations

- $\dot{\ell}(t, \theta, \eta) = (\partial / \partial t)\ell(t, \theta, \eta)$;
- $\ell_{t,\theta}(t, \theta, \eta) = (\partial^2 / \partial t \partial \theta)\ell(t, \theta, \eta)$;
- $\ell^{(3)}(t, \theta, \eta) = (\partial^3 / \partial t^3)\ell(t, \theta, \eta)$;
- $Pf = \int f dP$;
- $P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$;
- $G_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - Pf)$.

Assumptions

Regular Conditions:

- $d(\hat{\eta}_{\tilde{\theta}_n}, \eta_0) = O_P(n^{-1/2} \vee \|\tilde{\theta}_n - \theta_0\|)$; $\Longleftarrow$**rate assumption**
- $\tilde{I}_0$ is strictly positive definite.

Smoothness Conditions:

- The map $(t, \theta, \eta) \mapsto \ell(t, \theta, \eta)$ is smooth in each argument, e.g.
  - $(t, \theta) \mapsto (\partial^{l+m}/\partial t^l \partial \theta^m)\ell(t, \theta, \eta)$ have integrable envelope functions;
  - $G_n(\dot{\ell}(\theta_0, \theta_0, \hat{\eta}_{\tilde{\theta}_n}) - \dot{\ell}(\theta_0, \theta_0, \eta_0)) = O_P(n^{-1/2} \vee \|\tilde{\theta}_n - \theta_0\|)$. $\quad (*)$

Empirical Processes Conditions:

- P-Donsker Class: $\ddot{\ell}(t, \theta, \eta)$ and $\ell_{t,\theta}(t, \theta, \eta)$;
- P-Glivenko-Cantelli Class: $\ell^{(3)}(t, \theta, \eta)$.

The above assumptions of $\ell(t, \theta, \eta)$ make the profile likelihood $pl_n(\theta)$ behave like a full likelihood in the parametric models asymptotically.

Main Results

- Second Order Asymptotic Inference
- Basic Theorems
- Extensions
- Examples

Second Order Asymptotic Inferences

Theorem 1:

$$\begin{aligned}\log pl_n(\tilde{\theta}_n) &= \log pl_n(\hat{\theta}_n) - \frac{n}{2}(\tilde{\theta}_n - \hat{\theta}_n)'\tilde{I}_0(\tilde{\theta}_n - \hat{\theta}_n) \\ &\quad + O_P(n^{-\frac{1}{2}} \vee n\|\tilde{\theta}_n - \hat{\theta}_n\|^3). \qquad (7)\end{aligned}$$

Remark:

The observed profile information

$$\hat{I}_n(s_n) \equiv -2\frac{\log pl_n(\hat{\theta}_n + s_n) - \log pl_n(\hat{\theta}_n)}{ns_n^2}$$

Based on (7), $\hat{I}_n(s_n) = \tilde{I}_0 + O_P(|s_n| \vee n^{-3/2}|s_n|^{-2})$.

The **optimal** step size is $s_n \asymp n^{-1/2}$.

Basic Theorems

Theorem 2: Assume that $\hat{\theta}_n$ is asymptotically unique. If the proper prior $\rho(\theta_0) > 0$ and $\rho(\cdot)$ has continuous and finite first order derivative in some neighborhood of θ_0, then

$$
\begin{aligned}
\hat{\theta}_n &= \tilde{E}_{\theta|\tilde{X}}(\theta) + O_P(n^{-1}), \\
\tilde{I}_0 &= \left(n\tilde{Var}_{\theta|\tilde{X}}(\theta)\right)^{-1} + O_P(n^{-1/2}),
\end{aligned}
$$

provided that the prior $\rho(\cdot)$ has finite second moment.

Remark:

- The posterior profile mean is the second order approximation of $\hat{\theta}_n$;
- The inverse of the posterior profile variance is the second order estimator of the efficient information matrix.

Theorem 3: Let $\tilde{P}_{\theta|\tilde{X}}(\sqrt{n}\tilde{I}_0^{1/2}(\theta - \hat{\theta}_n) \le \kappa_{n\alpha}) = \alpha$. Assume that $\tilde{\ell}_0(X)$ has finite third moment and nondegenerate distribution, then there exists a unique $\hat{\kappa}_{n\alpha}$ based on the data such that $P(\sqrt{n}\tilde{I}_0^{1/2}(\hat{\theta}_n - \theta_0) \le \hat{\kappa}_{n\alpha}) = \alpha$ and $\sqrt{n}(\hat{\kappa}_{n\alpha} - \kappa_{n\alpha}) = O_P(1)$.

Remark:

- Theorem 3 implies that the Wald-type confidence interval for θ can be approximated by the Wald-type credible set based on the profile sampler with error of the order $O_P(n^{-1/2})$;
- **Conjecture:** The above $O_P(1)$ converges to the product of two different non-trivial but uniformly integrable Gaussian processes.

  $\Longrightarrow$ sharp rate.

  $\Longrightarrow P(\sqrt{n}\tilde{I}_0^{1/2}(\hat{\theta}_n - \theta_0) \le \kappa_{n\alpha}) = \alpha + O(n^{-1/2})$

Partial Simulations Results:

The Cox model with right censored data;

True $\theta_0 = 1$;

The true standard error of $\hat{\theta}_n$ is 3.7523.

500 Datasets are analyzed;

MCMC of length 5,000 with burn-in period of 1,000.

Table 1. The Cox model with right censored data.

n	MLE	Chain Mean	Std. Err.$_{\mathrm{M}}$	Std. Err.$_{\mathrm{N}}$
20	1.1049	1.1376	4.4128	4.3004
50	1.0202	1.0262	3.9869	3.9548
100	1.0156	1.0181	3.8592	3.8561
200	1.0131	1.0147	3.8124	3.8105
500	1.0012	1.0016	3.7598	3.7691

Std. Err.$_{\mathrm{M}}$, estimated standard errors based on MCMC; Std. Err.$_{\mathrm{N}}$, estimated standard errors based on numerical derivatives.

Table 2. The Cox model with right censored data.

n	$n\|\text{MLE} - \text{Chain Mean}\|$	$\sqrt{n}\|\text{Std. Err.}_{\text{M}} - \text{Std. Err.}_{\text{N}}\|$
20	0.6541	0.5027
50	0.3062	0.2270
100	0.2587	0.0311
200	0.3218	0.0279
500	0.2017	0.2080

Conclusions:

The profile sampler is proved to generate **second order** frequentist valid inference about θ in terms of moment (theorem 2) and quantile (theorem 3) under mile conditions of prior.

Extensions

- The estimation accuracy of the profile sampler procedure depends on the convergence rate of the nuisance parameter:

  Faster convergence rate $\Longrightarrow$ Higher estimation accuracy;

  (When convergence rate is parametric rate or slower rate.)

  [Cheng and Kosorok, 2006b]

  - Therefore, the frequentist inference about the Cox model with right censored data ($\|\hat{\eta}_{\tilde{\theta}_n} - \eta_0\|_\infty = O_P(n^{-1/2} + \|\tilde{\theta}_n - \theta_0\|)$) is more accurate than the Cox model with current status data ($\|\hat{\eta}_{\tilde{\theta}_n} - \eta_0\|_2 = O_P(n^{-1/3} + \|\tilde{\theta}_n - \theta_0\|)$).
  - Simulation results support the above theoretical judgement.

- Control the estimation accuracy by proposing the penalized profile sampler in which we profile the penalized log-likelihood:

  Assign smaller size of the smoothing parameter $\Longrightarrow$ More precise estimation;

  [Cheng and Kosorok, 2006c]

Note: The above phenomena can be realized only when we consider higher order asymptotic results.

Selected Examples from Other Areas

- **Econometrics** [Cheng and Kosorok, 2006c]
  - The partly linear model with current status;
  - Semiparametric logistic regression.
- **Epidemiology** [Cheng and Kosorok, 2006a]
  - Case-control studies with partially observed data.

Future Research Plan

Extensions of thesis work

- Apply the semiparametric Bayes methods, e.g. the profile sampler or fully Bayesian procedure [Shen, 2002], to the Biostatistics or Business models.
- Select the proper smoothing parameter of the penalized profile sampler in applications.

One Applied Problem:

- Evaluating the Microbial Role in Soil Carbon Dynamics Using Markov Chain Analysis. [Liang, Cheng and Balser, 2006]

One Theoretical Problem:

- Study the efficient estimation in semiparametric isotonic regression with many parameters; [Cheng, 2007]

$$Y = X'\theta + \eta(W) + \epsilon = X'\theta + \sum_{j=1}^{J_n} \eta_j(W_j) + \epsilon,$$

where observation $(Y, X, W) \in R \times R^d \times R^{J_n}$ and each $\eta_j(\cdot)$ is a monotone function.

Reference

Cheng, G. AND KOSOROK, M.R. (2006a). Higher Order Semiparametric Frequentist Inference with the Profile Sampler. *Tentatively accepted by the Annals of Statistics*

Cheng, G. AND KOSOROK, M.R. (2006b). General Frequentist Properties of the Posterior Profile Distribution. *Submitted to the Annals of Statistics*

Cheng, G. AND KOSOROK, M.R. (2006c). The Penalized Profile Sampler. *Submitted to the Annals of Statistics*

Cheng, G. (2007). Semiparametric Isotonic Regression with Many Parameters. *Technical Report*

Reference

LEE, B. L., KOSOROK, M. R. AND FINE, J. P. (2005). The Profile Sampler. *Journal of the American Statistical Association* **100** 960–969.

LIANG, C., **Cheng, G.** AND BALSER, T.C. (2006). Evaluating the Microbial Role in Soil Carbon Dynamics Using Markov Chain Analysis. *Accepted by the 18th World Congress of Soil Sciences*

MURPHY, S. A. AND VAN DER VARRT, A. W. (2000). On Profile Likelihood. *Journal of the American Statistical Association* **93** 1461–1474.

SHEN, X. (2002). Asymptotic Normality in Semiparametric and Nonparametric Posterior Distributions. *Journal of the American Statistical Association* **97** 222–235.