The Profile Sampler

Yilun Li

Nov 18th, 2021

Table of Contents

1. The Profile Sampler
2. The Penalized Profile Sampler
3. Other Methods

The Profile Sampler

1. The Profile Sampler
2. The Penalized Profile Sampler
3. Other Methods

Recall

- Empirical log-likelihood:

$$\mathcal{L}_n(\boldsymbol{\theta}, \eta) = n\mathbb{P}_n\ell(\boldsymbol{\theta}, \eta) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}, \eta)$$

- Profile log-likelihood:

$$pL_n(\boldsymbol{\theta}) = \sup_{\eta} \mathcal{L}_n(\boldsymbol{\theta}, \eta)$$

- Quadratic expansion of profile log-likelihood: Under certain regularity conditions, for any sequence $\tilde{\boldsymbol{\theta}}_n \to_P \boldsymbol{\theta}_0$, we have

$$\begin{aligned} pL_n(\tilde{\boldsymbol{\theta}}_n) = pL_n(\boldsymbol{\theta}_0) + (\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T \sum_{i=1}^{n} \tilde{\ell}_{\boldsymbol{\theta}_0, \eta_0}(X_i) \\ - \frac{1}{2} n(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T \tilde{\mathcal{I}}_{\boldsymbol{\theta}_0, \eta_0}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_{P_0}\left(1 + \sqrt{n}||\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0||\right)^2 \end{aligned} \tag{1}$$

where $\tilde{\ell}_{\boldsymbol{\theta}, \eta}$ is the efficient score function; $\tilde{\mathcal{I}}_{\boldsymbol{\theta}, \eta}$ is the efficient Fisher information matrix; P_0 is the probability measure of X at the true parameter value $\boldsymbol{\theta}_0$.

The Profile Sampler

- The material for this section comes largely from *Lee, Kosorok and Fine (2005)*[1], which proposed inference based on sampling from a posterior distribution based on the profile likelihood.

- The quadratic expansion of the profile likelihood permits the construction of confidence sets for $\boldsymbol{\theta}$ by inverting the log-likelihood ratio. But translating this theory into practice can be computationally challenging.

[1] [LKF] Lee, B. L., Kosorok, M. R., and Fine, J. P. (2005). The profile sampler. *Journal of the American Statistical Association*, 100:960–969.

- Even if the log profile likelihood ratio can be inverted for a multivariate parameter, this inversion does not enable the construction of confidence intervals for each one-dimensional subcomponent separately, as is standard practice in data analysis. For such confidence intervals, it would be necessary to further profile over all remaining components in $\boldsymbol{\theta}$.

- A related problem for which inverting the log likelihood is not adequate is the construction of rectangular confidence regions for $\boldsymbol{\theta}$, such as minimum volume confidence rectangles.

- In principle, having an estimator of θ and its variance simplifies these inferences considerably.
- However, the computation of these quantities using the semi-parametric likelihood is more challenging compared to that in parametric models.
- Finding the maximizer of the profile likelihood is done implicitly and typically involves numerical approximations. When the nuisance parameter is not $\sqrt{n}$ estimable, nonparametric functional estimation of η for fixed θ may be required, which depends heavily on the proper choice of smoothing parameters.

- Even when η is estimable at the parametric rate, and without smoothing, $\tilde{\mathcal{I}}_0$ does not ordinarily have a closed form.
- When it does have a closed form, it may include linear operators which are difficult to estimate well, and inverting the estimated linear operators may not be straightforward.
- And the validity of these variance estimators must be established on a case-by-case basis.

- The bootstrap is a possible solution to some of these problems. Theoretical justification for the bootstrap is possible but challenging for semi-parametric models where the nuisance parameter is not $\sqrt{n}$ consistent.

- Even when the bootstrap is valid, the computational burden is substantial, since maximization over both $\boldsymbol{\theta}$ and η is needed for each bootstrap sample.

- A different approach to variance estimation is possible via Corollary 19.4 which verifies that the curvature of the profile likelihood near $\hat{\boldsymbol{\theta}}_n$ is asymptotically equal to $\tilde{\mathcal{I}}_0$.

- In practice, we can perform second order numerical differentiation by the following steps:

  1. Evaluating the profile likelihood on a hyperrectangular grid of 3^p equidistant points centered at $\hat{\boldsymbol{\theta}}_n$;
  2. Taking the appropriate differences;
  3. Dividing by $4h^2$.

  where p is the dimension of $\boldsymbol{\theta}$; h is the spacing between grid points.

- Limitation: There are no clear cut rules on choosing the grid spacing in a given data set. Thus, it is difficult to automate this technique for practical usage.

The Profile Sampler

- Lee, Kosorok and Fine proposed an application of MCMC in their paper [LKF] to the semi-parametric profile likelihood.

- The method involves generating a Markov chain $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ...\}$ with stationary density proportional to $p_{\boldsymbol{\theta},n}(\boldsymbol{\theta}) = \exp[pL_n(\boldsymbol{\theta})]\, q(\boldsymbol{\theta})$, where $q = dQ/d\boldsymbol{\theta}$ for some prior measure Q.

- This can be accomplished by using the Metropolis-Hastings algorithm.

The Profile Sampler

- Main steps of the algorithm:

  1. Begin with an initial value $\boldsymbol{\theta}^{(1)}$ for the chain.
  2. For each $k = 2, 3, \ldots$, obtain a proposal $\tilde{\boldsymbol{\theta}}^{(k+1)}$ by random walk from $\boldsymbol{\theta}^{(k)}$.
  3. Compute $p_{\tilde{\boldsymbol{\theta}}^{(k+1)},n}(\tilde{\boldsymbol{\theta}}^{(k+1)})$, and decide whether to accept $\tilde{\boldsymbol{\theta}}^{(k+1)}$ by evaluating the ratio

$$\frac{p_{\tilde{\boldsymbol{\theta}}^{(k+1)},n}(\tilde{\boldsymbol{\theta}}^{(k+1)})}{p_{\boldsymbol{\theta}^{(k)},n}(\boldsymbol{\theta}^{(k)})} \tag{2}$$

     and applying an acceptance rule.

- After generating a sufficiently long chain, we can compute the mean of the chain to estimate the maximizer of $pL_n(\boldsymbol{\theta})$ and the variance of the chain to estimate $\tilde{\mathcal{I}}_0^{-1}$. And the output from the Markov chain can also be directly used to construct the confidence sets.

The Profile Sampler

Some remarks:

1. Whether or not a Markov chain is used to sample from the "posterior" proportional to $\exp[pL_n(\theta)]q(\theta)$, the procedure based on sampling from this posterior is referred to as the **profile sampler**.

2. Part of the computational simplicity of this procedure is that $pL_n(\theta)$ does not need to be maximized, and only needs to be evaluated.

The Profile Sampler

3 The profile likelihood is generally easy to compute as a consequence of algorithms such as the stationary point algorithm for maximizing over the nuisance parameter.

But sometimes it is hard to compute. In this case, numerical differentiation via Corollary 19.4 may be advantageous since it requires fewer evaluations of the profile likelihood.

However, numerical evidence in the paper [LKF] and some other theoretical work on the profile sampler indicate that, at least for moderately small samples, numerical differentiation does not perform as well as the profile sampler in general. And the profile sampler may still be beneficial even when the profile likelihood is hard to compute.

The Profile Sampler

4 The validity of the algorithm is established in Theorem 1 below, which enables the quadratic expansion of the profile log-likelihood around $\hat{\theta}_n$ to be valid in a fixed, bounded set, rather than only in a shrinking neighborhood.

5 The conclusion of these arguments is that the "posterior" distribution of the profile likelihood with respect to a prior on θ is asymptotically equivalent to the distribution of $\hat{\theta}_n$. And in order to do this, the new theorem will require an additional assumption on the profile likelihood.

The Profile Sampler

- Define $\Delta_n(\boldsymbol{\theta}) = \frac{1}{n}\left[pL_n(\boldsymbol{\theta}) - pL_n(\hat{\boldsymbol{\theta}}_n)\right]$.

Theorem 1

Suppose Θ is compact, $\tilde{\mathcal{I}}_0$ is positive definite, $Q(\Theta) < \infty$, and q is positive and continuous at $\boldsymbol{\theta}_0$. Also, assume that $\hat{\boldsymbol{\theta}}_n$ is efficient and for any random sequence $\tilde{\boldsymbol{\theta}}_n \to_P \boldsymbol{\theta}_0$, we have

$$pL_n(\tilde{\boldsymbol{\theta}}_n) = pL_n(\hat{\boldsymbol{\theta}}_n) - \frac{1}{2} n(\tilde{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n)^T \tilde{\mathcal{I}}_{\boldsymbol{\theta}_0,\eta_0}(\tilde{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n) + o_{P_0}\left(1 + \sqrt{n}||\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0||\right)^2. \tag{3}$$

Moreover, assume that for every random sequence $\{\tilde{\boldsymbol{\theta}}_n\} \subseteq \Theta$,

$$\Delta_n(\tilde{\boldsymbol{\theta}}_n) = \frac{1}{n}\left[pL_n(\tilde{\boldsymbol{\theta}}_n) - pL_n(\hat{\boldsymbol{\theta}}_n)\right] = o_{P_0}(1) \Rightarrow \tilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 + o_{P_0}(1). \tag{4}$$

Then for any measurable function $g : \mathbb{R}^k \to \mathbb{R}$ satisfying

$$\limsup_{k\to\infty} \frac{1}{k^2} \log\left[\sup_{\mathbf{u}\in\mathbb{R}^k : ||\mathbf{u}|| \le k} |g(\mathbf{u})|\right] \le 0, \tag{5}$$

we have

$$\frac{\int_\Theta g(\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)) p_{\boldsymbol{\theta},n}(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_\Theta p_{\boldsymbol{\theta},n}(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \int_{\mathbb{R}^k} g(\mathbf{u}) \frac{1}{(\sqrt{2\pi})^k} |\tilde{\mathcal{I}}_0|^{1/2} \exp\left(-\frac{1}{2}\mathbf{u}^T \tilde{\mathcal{I}}_0 \mathbf{u}\right) d\mathbf{u} + o_{P_0}(1). \tag{6}$$

[a]

[a] The proof can be found in Section 19.4 of the textbook by Dr. Kosorok.

The Profile Sampler

Remarks:

1. When $g(\mathbf{u}) = O(1 + ||\mathbf{u}||)^d$ for any $d < \infty$, Condition (5) is readily satisfied. This means that the first two moments of $\sqrt{n}(T - \hat{\boldsymbol{\theta}}_n)$, where T has the posterior density proportional to $\mathbf{t} \mapsto p_{\mathbf{t},n}(\mathbf{t})$, are consistent for the corresponding probabilities and moments of the limiting Gaussian distribution. Specifically, $E(T) = \hat{\boldsymbol{\theta}}_n + o_{P_0}(n^{-1/2})$ and $nVar(T) = \tilde{\mathcal{I}}_0^{-1} + o_{P_0}(1)$.

   Thus we can calculate all the quantities needed for inference on $\boldsymbol{\theta}$ without maximizing the profile likelihood directly or computing derivatives.

2 The Condition (4) is not implied by the identifiability of the Kulback-Leibler information from the full likelihood.

Nevertheless, if it can be shown that $\Delta_n(\boldsymbol{\theta})$ converges uniformly over Θ to the profiled Kulback-Leibler information $\Delta_0(\boldsymbol{\theta})$, then identifiability of the Kulback-Leibler information for the empirical log-likelihood $\mathcal{L}_n(\boldsymbol{\theta}, \eta)$ is sufficient. And this approach works for the Cox model for right-censored data.

3 The Condition (4) is needed because the integration in (6) is over all of Θ, and thus it is important to guarantee that there are no other distinct modes besides $\hat{\boldsymbol{\theta}}_n$ in the limiting posterior.

The Profile Sampler

4 The profile sampler works well and is in general computationally efficient.

5 The Metropolis algorithm applied to $p_{\theta,n}(\theta)$ with a Lebesgue prior measure: By the ergodic theorem, there exists a sequence of finite chain lengths $\{M_n\} \to \infty$ such that

- the chain mean $\bar{\theta}_n = M_n^{-1} \sum_{j=1}^{M_n} \theta^{(j)}$ satisfies $\bar{\theta}_n = \hat{\theta}_n + o_{P_0}(n^{-1/2})$;
- the standardized sample variance $\mathbf{V}_n = M_n^{-1} \sum_{j=1}^{M_n} n(\theta^{(j)} - \bar{\theta}_n)(\theta^{(j)} - \bar{\theta}_n)^T$ is consistent for $\tilde{\mathcal{I}}_0^{-1}$;
- the empirical measure

$$G_n(A) = M_n^{-1} \sum_{j=1}^{M_n} I\left(\sqrt{n}(\theta^{(j)} - \bar{\theta}_n) \in A\right)$$

for a bounded convex $A \subseteq \mathbb{R}^k$, is consistent for the probability that a mean zero Gaussian deviate with variance $\tilde{\mathcal{I}}_0^{-1}$ lies in A.

Hence the output of the chain can be used for inference about θ_0, provided M_n is large enough so that the sampling error from using a finite chain is negligible.

The Profile Sampler

Example 1 of Verification of Condition (4): $\Delta_n(\tilde{\theta}_n) = o_{P_0}(1) \Rightarrow \tilde{\theta}_n = \theta_0 + o_{P_0}(1)$ (Cox Model for Right-censored Data)

- In this example, we use the identifiability of the profile Kulback-Leibler information to verify the condition (4).

- Notation and assumptions:
  - Let B be the compact parameter space for $\boldsymbol{\beta}$;
  - The true parameter vector $\boldsymbol{\beta}_0$ is known to be in the interior of B;
  - The norm of the covariate vector $||\mathbf{Z}||$ is bounded by a constant.

- By discussion in previous section,

$$\hat{\Lambda}_{\beta}(s) = \int_0^s \frac{\mathbb{P}_n dN(u)}{\mathbb{P}_n Y(u) \exp(\beta^T \mathbf{Z})} \tag{7}$$

  is the maximizer of the log-likelihood function over Λ given any fixed β.

- Combined the conclusion above with the log-likelihood function, we have

$$\frac{1}{n} pL_n(\beta) = H_n(\beta) + C_0 \tag{8}$$

$$H_n(\beta) = \mathbb{P}_n \left[\int_0^{\tau} \left(\beta^T \mathbf{Z} - \log \left\{ \mathbb{P}_n \left[Y(s) \exp(\beta^T \mathbf{Z}) \right] \right\} \right) dN(s) \right] \tag{9}$$

  where C_0 is a constant independent of β.

- It can also be proved that $||H_n - H_0||_B \to_P 0$, where

$$H_0(\beta) = P_0 \left[\int_0^\tau \left(\beta^T \mathbf{Z} - \log \left\{ P_0 \left[Y(s) \exp(\beta^T \mathbf{Z}) \right] \right\} \right) dN(s) \right] \tag{10}$$

- So the first derivative of H_0 w.r.t. β, denoted by $U_0(\beta)$, is:

$$\mathbf{U}_0(\beta) = P_0 \left\{ \int_0^\tau \left[\mathbf{Z} - \mathbf{E}(s, \beta) \right] dN(s) \right\} \tag{11}$$

$$\mathbf{E}(s, \beta) = \frac{P_0 \left[\mathbf{Z} Y(s) \exp(\beta^T \mathbf{Z}) \right]}{P_0 \left[Y(s) \exp(\beta^T \mathbf{Z}) \right]} \tag{12}$$

- And the second derivative of H_0 is:

$$-\mathbf{V}(\beta) = -\int_0^\tau \left(\frac{P_0 \left[\mathbf{Z}\mathbf{Z}^T Y(s) \exp(\beta^T \mathbf{Z}) \right]}{P_0 \left[Y(s) \exp(\beta^T \mathbf{Z}) \right]} - \left\{ \frac{P_0 \left[\mathbf{Z} Y(s) \exp(\beta^T \mathbf{Z}) \right]}{P_0 \left[Y(s) \exp(\beta^T \mathbf{Z}) \right]} \right\}^{\otimes 2} \right) P_0 \left[Y(s) \exp(\beta^T \mathbf{Z}) \right] d\Lambda(s) \tag{13}$$

The Profile Sampler

Example 1 of Verification of Condition (4): $\Delta_n(\tilde{\theta}_n) = o_{P_0}(1) \Rightarrow \tilde{\theta}_n = \theta_0 + o_{P_0}(1)$ (Cox Model for Right-censored Data)

- By the boundedness of $||Z||$ and B, it can be verified that $\exists c_1 > 0$ independent of β such that the matrix $[\mathbf{V}(\beta) - c_1 Var(Z)]$ is positive semi-definite.
- Thus, H_0 is strictly concave and has a unique maximum in B.
- Since we have $\mathbf{U}_0(\beta_0) = \mathbf{0}$ by equation (11), the unique maximizer of H_0 is exactly β_0.

- We define $\Delta_0(\beta) = H_0(\beta) - H_0(\beta_0)$, which is continuous and non-positive, and is strictly negative whenever $\beta \neq \beta_0$. Thus,

$$\begin{aligned} ||\Delta_n(\beta) - \Delta_0(\beta)||_B &= ||H_n(\beta) - H_n(\hat{\beta}_n) - H_0(\beta) + H_0(\beta_0)||_B \\ &= ||H_n(\beta) - H_n(\hat{\beta}_n) - H_0(\beta) + H_0(\hat{\beta}_n) - H_0(\hat{\beta}_n) + H_0(\beta_0)||_B \\ &\leq 2||H_n - H_0||_B + ||H_0(\hat{\beta}_n) - H_0(\beta_0)|| \to_P 0 \end{aligned} \tag{14}$$

- Since $\Delta_0(\beta_0) = 0$ and (14) holds, if $\Delta_n(\tilde{\beta}_n) = o_{P_0}(1)$, we can conclude that $\tilde{\beta}_n = \beta_0 + o_{P_0}(1)$ by the identifiability of the profile Kulback-Leibler information $\Delta_0(\beta)$.

- Recall:
  - Current status data arises when each subject is observed at a single examination time Y to determine whether an event has occurred, and the event time T cannot be observed exactly.
  - Along with the covariate vector $\mathbf{Z}$, the observed data consists of n i.i.d. realizations of $X = (Y, \delta, \mathbf{Z})$, where $\delta = I(T \leq Y)$.
  - The log-likelihood function for a single observation has the form of:

$$\ell(\boldsymbol{\beta}, \Lambda) = \delta \log \left\{ 1 - \exp\left[-\Lambda(Y) \exp(\boldsymbol{\beta}^T \mathbf{Z}) \right] \right\} - (1 - \delta) \exp(\boldsymbol{\beta}^T \mathbf{Z}) \Lambda(Y). \tag{15}$$

- In this example, we verify condition (4) directly.

Example 2 of Verification of Condition (4): $\Delta_n(\tilde{\theta}_n) = o_{P_0}(1) \Rightarrow \tilde{\theta}_n = \theta_0 + o_{P_0}(1)$ (Cox Model for Current Status Data)

- Let $\{\tilde{\beta}_n\}$ be a random sequence satisfying $\Delta_n(\tilde{\beta}_n) = o_{P_0}(1)$.
- Fix any $\alpha \in (0, 1)$. Since

$$\Delta_n(\tilde{\beta}_n) = \frac{1}{n}\left[pL_n(\tilde{\beta}_n) - pL_n(\hat{\beta}_n)\right] = o_{P_0}(1),\ \Delta_n(\beta_0) = \frac{1}{n}\left[pL_n(\beta_0) - pL_n(\hat{\beta}_n)\right] \leq 0,$$

we have

$$\frac{1}{n}\left[pL_n(\tilde{\beta}_n) - pL_n(\beta_0)\right] \geq o_{P_0}(1),$$

or equivalently,

$$\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; \mathbf{X}_i)}{f(\beta_0, F_0; \mathbf{X}_i)}\right] \geq o_{P_0}(1), \tag{16}$$

where the likelihood function

$f(\beta, F; \mathbf{X}) = \delta\left\{1 - \exp\left[-\Lambda(Y)\exp(\beta^T\mathbf{Z})\right]\right\} + (1-\delta)\exp\left[-\Lambda(Y)\exp(\beta^T\mathbf{Z})\right]$;

$\Lambda = -\log(1-F)$; $\hat{F}_\beta = 1 - \exp(-\hat{\Lambda}_\beta)$ is the maximizer of the likelihood function over the nuisance parameter for fixed β.

- Since $\alpha \log(x) \leq \log[1 + \alpha(x-1)]$ for any $x > 0$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \log\left\{1 + \alpha\left[\frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; \mathbf{X}_i)}{f(\beta_0, F_0; \mathbf{X}_i)} - 1\right]\right\} \geq o_{P_0}(1), \tag{17}$$

- (Lemma 2) The class $\mathcal{F} = \{f(\beta, F; \mathbf{X}) : \beta \in B, F \in \mathcal{M}\}$ is P_0-Donsker, where $\mathcal{M}$ is the class of distribution functions on $[0, \tau]$. [1]

- So we have

$$P_0\left(\log\left\{1 + \alpha\left[\frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; \mathbf{X})}{f(\beta_0, F_0; \mathbf{X})} - 1\right]\right\}\right) \geq o_{P_0}(1), \tag{18}$$

- On the other hand, by Jensen's Inequality and the strict concavity of the function $x \mapsto \log(x)$, it can be proved that

$$P_0\left(\log\left\{1 + \alpha\left[\frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; \mathbf{X})}{f(\beta_0, F_0; \mathbf{X})} - 1\right]\right\}\right) \leq \log\left(P_0\left\{1 + \alpha\left[\frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; \mathbf{X})}{f(\beta_0, F_0; \mathbf{X})} - 1\right]\right\}\right) \leq 0, \tag{19}$$

[1] Proof of the lemma is given in Section 19.4 of the textbook by Dr. Kosorok.

- Thus by (18) and (19), we have

$$P_0\left(\log\left\{1+\alpha\left[\frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; \mathbf{X})}{f(\beta_0, F_0; \mathbf{X})} - 1\right]\right\}\right) = o_{P_0}(1), \tag{20}$$

$$\log\left(P_0\left\{1+\alpha\left[\frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; \mathbf{X})}{f(\beta_0, F_0; \mathbf{X})} - 1\right]\right\}\right) = o_{P_0}(1). \tag{21}$$

- Then it can be proved that

$$P_0\left|\left(1-\hat{F}_{\tilde{\beta}_n}\right)^{\exp(\tilde{\beta}_n^T \mathbf{Z})} - (1-F_0)^{\exp(\beta_0^T \mathbf{Z})}\right| = o_{P_0}(1), \tag{22}$$

  which can further imply that

$$P_0\left(\left\{\left(\tilde{\beta}_n - \beta_0\right)^T [Z - E(Z|Y)] - c_n(Y)\right\}^2 \middle| Y\right) = o_{P_0}(1) \tag{23}$$

  for almost surely all Y, where $c_n(Y)$ is uncorrelated with $[Z - E(Z|Y)]$, and the desired result $\tilde{\beta}_n = \beta_0 + o_{P_0}(1)$ can be derived based on (23).

The Penalized Profile Sampler

1 The Profile Sampler

2 The Penalized Profile Sampler

3 Other Methods

The Penalized Profile Sampler

- In many semi-parametric models involving a smooth nuisance parameter, it is convenient and beneficial to perform estimation using penalization.

- One motivation for this is that, in the absence of any restrictions on the form of the function η, maximum likelihood estimation for some semi-parametric models leads to over-fitting.

- And under certain regularity conditions, penalized semi-parametric log-likelihood estimation can yield fully efficient estimates for $\boldsymbol{\theta}$.

The Penalized Profile Sampler

- We will consider a modification of the profile sampler that works with profiled penalized likelihoods.

- Assume the nuisance parameter η is a function in Sobolev class of functions supported on some compact set $\mathcal{U}$ on the real line, whose d_{th} derivative exists and is absolutely continuous with $J(\eta) < \infty$, where

$$J^2(\eta) = \int_{\mathcal{U}} \left[\eta^{(d)}(u)\right]^2 du. \tag{24}$$

  Here d is a fixed, positive integer, and $\eta^{(j)}$ is the j_{th} derivative of η with respect to u. And we denote $\mathcal{H}$ to be the Sobolev function class with degree d on $\mathcal{U}$.

The Penalized Profile Sampler

- Then the penalized log-likelihood is:

$$\tilde{L}_n(\boldsymbol{\theta}, \eta) = \frac{1}{n}\mathcal{L}_n(\boldsymbol{\theta}, \eta) - \lambda_n^2 J^2(\eta), \tag{25}$$

where $\mathcal{L}_n(\boldsymbol{\theta}, \eta)$ is the empirical log-likelihood function; λ_n is a smoothing parameter.

- Assumptions about the smoothing parameter λ_n:

$$\begin{cases} \lambda_n = o_{P_0}(n^{-1/4}) \\ \lambda_n^{-1} = O_{P_0}(n^{d/(2d+1)}) \end{cases} \tag{26}$$

e.g., $\lambda_n = n^{-d/(2d+1)}$, or $\lambda_n = n^{-1/3}$ which is independent of d.

- The penalized profile log-likelihood is defined as

$$p\tilde{L}_n(\boldsymbol{\theta}) = \tilde{L}_n(\boldsymbol{\theta}, \tilde{\eta}_{\boldsymbol{\theta}}), \tag{27}$$

where $\tilde{\eta}_{\boldsymbol{\theta}} = \arg\max_{\eta \in \mathcal{H}} \tilde{L}_n(\boldsymbol{\theta}, \eta)$ for fixed $\boldsymbol{\theta}$ and λ_n.

- The penalized profile sampler is the procedure of sampling from the posterior distribution of $p\tilde{L}_n(\boldsymbol{\theta})$ by assigning a prior on $\boldsymbol{\theta}$.

Other Methods

- The exchangeable bootstrap (Cheng and Huang, *Annals of Statistics*);
- *m* within *n* subsampling (Bickel, Götze and van Zwet, 1997).
- Subsampling (Politis and Ramono, 1994).
- Block jackknife (Ma and Kosorok, 2005a).
- Bayesian methods (Shen, 2002).
- .....