Journal of Multivariate Analysis 100 (2009) 345–362

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

The penalized profile sampler

Guang Cheng [a,*], Michael R. Kosorok [b]

[a] Department of Statistical Science, Duke University, 214 Old Chemistry Building, Durham, NC 27708, USA

[b] Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, 3101 McGavran-Greenberg Hall, Chapel Hill, NC 27599, USA

ARTICLE INFO

Article history:
Received 18 May 2007
Available online 8 May 2008

AMS 2000 subject classifications:
Primary 62G20
62F1
62F15

Keywords:
Penalized likelihood
Posterior distribution
Profile likelihood
Semiparametric inference
Smoothing parameter

ABSTRACT

The penalized profile sampler for semiparametric inference is an extension of the profile sampler method [B.L. Lee, M.R. Kosorok, J.P. Fine, The profile sampler, Journal of the American Statistical Association 100 (2005) 960–969] obtained by profiling a penalized log-likelihood. The idea is to base inference on the posterior distribution obtained by multiplying a profiled penalized log-likelihood by a prior for the parametric component, where the profiling and penalization are applied to the nuisance parameter. Because the prior is not applied to the full likelihood, the method is not strictly Bayesian. A benefit of this approximately Bayesian method is that it circumvents the need to put a prior on the possibly infinite-dimensional nuisance components of the model. We investigate the first and second order frequentist performance of the penalized profile sampler, and demonstrate that the accuracy of the procedure can be adjusted by the size of the assigned smoothing parameter. The theoretical validity of the procedure is illustrated for two examples: a partly linear model with normal error for current status data and a semiparametric logistic regression model. Simulation studies are used to verify the theoretical results.

1. Introduction

Semiparametric models are statistical models indexed by both a finite dimensional parameter of interest θ and an infinite dimensional nuisance parameter η. In order to make statistical inference about θ separately from η, we estimate the nuisance parameter with $\hat{\eta}_\theta$, its maximum likelihood estimate at each fixed θ, i.e.

$$\hat{\eta}_\theta = \operatorname*{argmax}_{\eta\in\mathcal{H}} lik_n(\theta, \eta),$$

where $lik_n(\theta,\eta)$ is the likelihood of the semiparametric model given n observations and $\mathcal{H}$ is the parameter space for η. Therefore, we can do frequentist inference about θ based on the profile likelihood, which is typically defined as

$$pl_n(\theta) = \sup_{\eta\in\mathcal{H}} lik_n(\theta, \eta).$$

The convergence rate of the nuisance parameter η is the order of $d(\hat{\eta}_{\tilde{\theta}_n}, \eta_0)$, where $d(\cdot,\cdot)$ is some metric on η, $\tilde{\theta}_n$ is any sequence satisfying $\tilde{\theta}_n = \theta_0 + o_P(1)$, and (η_0, θ_0) is the true value of (η, θ). Typically,

$$d(\hat{\eta}_{\tilde{\theta}_n}, \eta_0) = O_P(\|\tilde{\theta}_n - \theta_0\| + n^{-r}), \tag{1}$$

* Corresponding address: Department of Statistics, Purdue University, West Lafayette, IN 47907, USA.
E-mail addresses: chengg@stat.purdue.edu (G. Cheng), kosorok@unc.edu (M.R. Kosorok).

0047-259X/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2008.05.001

where $\|\cdot\|$ is the Euclidean norm and $r > 1/4$. Of course, a smaller value of r leads to a slower convergence rate of the nuisance parameter. For instance, the nuisance parameter in the Cox proportional hazards model with right censored data, the cumulative hazard function, has the parametric rate, i.e., $r = 1/2$. If current status data is applied to the Cox model instead, then the convergence rate will be slower, with $r = 1/3$, due to the loss of information provided by this kind of data.

The profile sampler is the procedure of sampling from the posterior of the profile likelihood in order to estimate and draw inference on the parametric component θ in a semiparametric model, where the profiling is done over the possibly infinite-dimensional nuisance parameter η. [9] show that the profile sampler gives a first order correct approximation to the maximum likelihood estimator $\hat{\theta}_n$ and consistent estimation of the efficient Fisher information for θ even when the nuisance parameter is not estimable at the $\sqrt{n}$ rate. Another Bayesian procedure employed to do semiparametric estimation is considered in [16] who study the marginal semiparametric posterior distribution for a parameter of interest. In particular, [16] show that marginal semiparametric posterior distributions are asymptotically normal and centered at the corresponding maximum likelihood estimates or posterior means, with covariance matrix equal to the inverse of the Fisher information. Unfortunately, this fully Bayesian method requires specification of a prior on η, which is quite challenging since for some models there is no direct extension of the concept of a Lebesgue dominating measure for the infinite-dimensional parameter set involved [8]. The advantages of the profile sampler for estimating θ compared to other methods is discussed extensively in [2,3,9].

The motivation for studying second order asymptotic properties of the profile sampler comes from the observed simulation differences in the Cox model with different types of data, i.e. right censored data [2] and current status data [9]. The profile sampler generated based on the first model yields much more accurate estimation results comparing to the second model when the sample size is relatively small. [2,3] have successfully explored the theoretical reasons behind the above phenomena by establishing the relation between the estimation accuracy of the profile sampler, measured in terms of second order asymptotics, and the convergence rate of the nuisance parameters. Specifically, the profile sampler generated from a semiparametric model with a faster convergence rate usually yields more precise frequentist inference of θ. These second order results are verified in [2,3] for several examples, including the proportional odds model, case-control studies with missing covariates, and the partly linear model. The convergence rates for these models range from the parametric to the cubic. The work in [3] has shown clearly that the accuracy of the inference for θ based on the profile sampler method is intrinsically determined by the semiparametric model specifications through its entropy number.

In many semiparametric models involving a smooth nuisance parameter, it is often 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 semiparametric models leads to over-fitting. Seminal applications of penalized maximum likelihood estimation include estimation of a probability density function in [17] and nonparametric linear regression in [18]. Note that penalized likelihood is a special case of penalized quasi-likelihood studied in [13]. Under certain reasonable regularity conditions, penalized semiparametric log-likelihood estimation can yield fully efficient estimates for θ (see, for example, [13]). As far as we are aware, the only general procedure for inference for θ in this context known to be theoretically valid is a weighted bootstrap with bounded random weights (see [11]). It is even unclear whether the usual nonparametric bootstrap will work in this context when the nuisance parameter has a convergence rate $r < 1/2$.

The purpose of this paper is to ask the somewhat natural question: does sampling from the exponential of a profiled penalized log-likelihood (which process we refer hereafter to as the penalized profile sampler) yield first and even second order accurate frequentist inference? The conclusion of this paper is that the answer is yes and, moreover, the accuracy of the inference depends in a fairly simple way on the size of the smoothing parameter.

The unknown parameters in the semiparametric models we study in this paper include θ, which we assume belongs to some compact set $\Theta \subset \mathbb{R}^d$, and η, which we assume to be a function in the Sobolev class of functions $\mathcal{H}_k$ or its subset $\mathcal{H}_k^M \equiv \mathcal{H}_k \cap \{\eta : \|\eta\|_\infty \le M\}$ for some known $M < \infty$ supported on some compact set on the real line. The Sobolev class of functions $\mathcal{H}_k$ is defined as the set $\{\eta : J^2(\eta) \equiv \int_{\mathcal{Z}} (\eta^{(k)}(z))^2 dz < \infty\}$, where $\eta^{(j)}$ is the j-th derivative of η with respect to z. Obviously $J^2(\eta)$ is some measurement of complexity of η. We denote $\mathcal{H}_k$ as the Sobolev function class with degree k. The penalized log-likelihood in this context is:

$$\log lik_{\lambda_n}(\theta, \eta) = \log lik(\theta, \eta) - n\lambda_n^2 J^2(\eta), \quad (2)$$

where $\log lik(\theta, \eta) \equiv n\mathbb{P}_n \ell_{\theta,\eta}(X)$, $\ell_{\theta,\eta}(X)$ is the log-likelihood of the single observation X, and λ_n is a smoothing parameter, possibly dependent on data. In practice, λ_n can be obtained by cross-validation [21] or by inspecting the various curves for different values of λ_n. The penalized maximum likelihood estimators $\hat{\theta}_n$ and $\hat{\eta}_n$ depend on the choice of the smoothing parameter λ_n. Consequently, we use the notation $\hat{\theta}_{\lambda_n}$ and $\hat{\eta}_{\lambda_n}$ for the remainder of this paper to denote the estimators obtained from maximizing (2). In particular, a larger smoothing parameter usually leads to a less rough penalized estimator of η_0. It is of interest to establish the asymptotic property of the proposed penalized profile sampler procedure with a data-driven λ_n. Further studies on this issue are needed, but it is beyond the scope of this paper.

For the purpose of establishing first order accuracy of inference for θ based on the penalized profile sampler, we assume that the bounds for the smoothing parameter are as follows:

$$\lambda_n = o_P(n^{-1/4}) \quad \text{and} \quad \lambda_n^{-1} = O_P(n^{k/(2k+1)}). \quad (3)$$