SciMedCentral

Annals of Biometrics & Biostatistics

Review Article

B-Spline-Based Sieve Estimation in Survival Analysis

Yuan Wu*

Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, USA

Corresponding author
Yuan Wu, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, USA, E-mail: yuan.wu@duke.edu

Submitted: 25 June 2013

Accepted: 02 July 2013

Published: 05 July 2013

Keywords
- The B-Spline functions
- Sieve estimation
- Survival analysis

Abstract

This manuscript reviews important work about the B-spline-based sieve maximum likelihood estimation in survival analysis including studies for nonparametric and semiparametric problems. It also discusses some advantages of spline-based sieve estimation over purely non- parametric or semiparametric estimation including its computing cost and estimation accuracy.

INTRODUCTION

Spline functions are piecewise polynomial functions and used to approximate smooth functions in numerical analysis. The B-spline functions are one type of the most popular spline functions in statistical computing and available in multiple statistical packages. Here a B-spline function is a linear combination of a set of B-spline basis functions. A set of B-spline basis functions are defined by Schumaker [1]. Sieve estimation [2] means the maximum likelihood estimation in a subspace of whole objective function space where the subspace converges to the whole space as sample size increases. Hence, intuitively sieve estimation converges to the true target function as sample size increases as well. In this manuscript we review the sieve estimation in the B-spline function space in survival analysis models. The B-spline-based sieve estimation requires that the true target function is a smooth function, which is a reasonable assumption for many real applications. The B-spline-based sieve estimation is usually more efficient than its purely nonparametric or semiparametric counterpart when the smooth assumption is valid. First we discuss the B-spline-based sieve estimation approach in nonparametric problems. Then we discuss the B-spline-based semiparametric sieve estimation approach. Finally, we discuss some of the advantages of the B-spline-based sieve estimation approach over the conventional approach; we also discuss the applicability of this approach for future studies.

B-spline-based nonparametric sieve estimation

Lu et al. [3] proposed likelihood-based sieve estimation for the mean function of the counting processes with panel count data. They actually adopted the monotone I-spline function to approximate the smooth monotone mean function of the counting process. Wu and Zhang [4] found that each I-spline basis function is a partial sum of a set of B-spline basis functions. The adoption of the I-spline function instead of the B-spline function could simplify the constraints to guarantee monotonicity [4]. The idea of estimating monotone mean function directly rather than its derivative as proposed by Cheng and Wang [5] would require monotone constraints but remove the extra integration step. They proved that the sieve estimators has faster convergence rate than their purely nonparametric alternatives proposed by Wellner and Zhang [6]. Simulation studies showed that their estimators have smaller variances and mean squared errors as well. Wu and Zhang [4] proposed tensor-spline-based sieve estimation for the joint cumulative distribution function (CDF) with bivariate current status data. In this project, we adopted the tensor spline function to approximate the bivariate CDF. The tensor spline function is a linear combination of the tensor spline basis function [7]. As in Lu et al. [3], we used the tensor I-spline function for computing to simplify the partially monotone constraints. However, due to the well developed theoretical results for the B-spline function in numerical analysis, we used the tensor B-spline function for theoretical development. We proved that the tensor spline estimator for joint CDF convergent and our convergence rate is better than n^{1/3}, the conjecture of the convergence rate for the alternative purely nonparametric approach proposed by [8]. Our simulation studies showed that our estimator has smaller bias and mean square error than its counterpart.

B-spline-based semiparametric sieve estimation

As an extension project of Lu et al. [3], Lu et al. [9] proposed the B-spline-based semiparametric sieve likelihood-based methods for panel count data with proportional mean model. They estimated the baseline mean function for counting process and regression parameter simultaneously with the B-spline functions approximating the baseline mean functions. They proved that the sieve estimator for the baseline mean has faster convergence rate than their purely semiparametric alternatives proposed by Wellner and Zhang [10]. They also established the asymptotic normality for the regression parameter. And simulation studies showed that the semiparametric sieve estimators have advantage over their counterparts in terms of computing cost however

at least behave equivalently in terms of accuracy with their counterparts. To avoid assuming any stochastic model for the underlying counting process as in Lu et al. [9], Hua and Zhang [11] proposed the B-spline-based semiparametric sieve generalized estimating equation method with the proportional mean model. They also chose working covariance matrix accounting for overdispersion to improve the estimation commutating efficiency and produce less biased variance estimations compared to Lu et al. [9] when overdispersion occurs. Zhang et al. [12] proposed B-spline-based semiparametric maximum likelihood estimation method for proportional hazard model with interval-censored data. They estimated the baseline hazard function and regression parameter simultaneously with the B-spline functions approximating the baseline cumulative hazard functions. They proved that the sieve estimator for baseline cumulative hazard function has faster convergence rate than the alternative purely semiparametric approach proposed by Huang and Wellner [13]. For regression parameter they established its asymptotic normality more importantly proposed an easy-to-compute observed information matrix for its variance estimation. They also showed computing efficiency for the sieve method over its purely semiparametric alternative by simulation studies. Related to Zhang et al. [12], Cheng and Wang [5] studied semiparametric additive transformation model under current status data with partly linear additive proportional hazards model as a special case. They adopted the B-spline functions to approximate the nonparametric parts and estimate the nonparametric terms and parametric regression coefficients simultaneously. The difference to Zhang et al. [12] is that Cheng and Wang [5] tried to find the sieve estimation for log hazard function to remove the monotone constraints in computing with tolerating an extra numerical integration step. Cheng and Wang [5] proved that the sieve estimator has faster convergence rate than its purely semiparametric alternative proposed by Ma and Kosorok [14]. They showed that parametric regression coefficients have asymptotic normality and adopting the same idea as in Zhang et al. [12] they also found the observed information matrix.

Chen et al. [15] proposed a sieve maximum likelihood estimation of a general semiparametric copula model for bivariate data without censoring. One type of sieve estimation they considered is the B-spline-based sieve estimation and their sieve estimation is for marginal distributions. They showed that both sieve estimators for marginal distributions and the copula association parameter have asymptotic normality. For bivariate interval censored data, Wu and Gao [7] proposed a two-stage B-spline-based semiparametric sieve maximum likelihood estimation. The sieve estimation is also for the marginal distributions. By simulation studies we showed that our method has computing efficiency over the purely semiparametric alternative proposed by Sun et al. [16].

DISCUSSION AND CONCLUSION

In this manuscript, we reviewed important papers on B-spline-based sieve estimation in survival analysis. These studies include nonparametric models for univariate panel counting data and bivariate current status data; and semiparametric models for univariate panel counting data, univariate interval censored data and bivariate data with copula. We found that theoretically the sieve estimation methods have faster convergence rate than their purely nonparametric or semiparametric alternatives; for semiparametric models, even though the sieve estimators for nonparametric parts converge in a rate slower that $\sqrt{n}$, the regression parameters can still be showed to have asymptotic normality, more importantly, the variance for parameter estimators can be easily estimated efficiently for sieve estimation method as in Zhang et al. [12]. The computing cost for sieve estimation is less expensive, since the number of spline knots is much smaller than the sample size and the dimension of the problem is largely reduced.

Many other problems can be studied by the B-spline-based sieve estimation method. For example, the sieve method in Wu and Zhang [4] can be easily applied to bivariate interval censored. And as mentioned in Zhang et al. [12], the sieve method is applicable to any semiparametric maximum likelihood estimation problems in which the estimator for the regression parameter has asymptotic normality but the computing is not efficient.

ACKNOWLEDGEMENTS

We owe thanks to Dr. Guang Cheng, associate professor of Statistics at Purdue University for very helpful discussions.

REFERENCES

1. Schumaker L. Spline Functions: Basic Theory. 2nd ed, Wiley, New York. 1981.
2. Geman A, Hwang C. Nonparametric Maximum Likelihood Estimation by the Method of Sieves. Ann Statist. 1982; 10: 401-414.
3. Lu M, Zhang Y, Huang J. Estimation of the Mean Function with Panel Count Data Using Monotone Polynomial Splines. Biometrika. 2007; 94: 705-718.
4. Wu Y, Zhang Y. Partially Monotone Tensor Spline Estimation of the Joint Distribution Function with Bivariate Current Status Data. Ann Statist. 2012; 40: 1609-1636.
5. Cheng G, Wang X. Semiparametric Additive Transformation Model under Current Status Data. Electronic Journal of Statistics. 2011; 5: 1735-1764.
6. Wellner J, Zhang Y. Two Estimators of the Mean of a Counting Process with Panel Count Data. Ann Statist. 2000; 28: 779-814.
7. Wu Y, Gao X. Sieve Estimation with Bivariate Interval Censored Data. Journal of Statistics: Advances in Theory and Applications 2011; 16: 37-61.
8. Maathuis MH. Reduction Algorithm for the NPMLE for the Distribution Function of Bivariate Interval-Censored Data. J Comput Graph Statist. 2005; 14: 352-362.
9. Lu M, Zhang Y, Huang J. Semiparametric Estimation Methods for Panel Count Data Using Monotone B-Splines. J Amer Statist Assoc. 2009; 104: 1060-1070.
10. Wellner J, Zhang Y. Likelihood-based Semiparametric Estimation Methods for Panel Count Data with Covariates. Ann Statist. 2007; 35: 2106-2142.
11. Hua L, Zhang Y. Spline-based semiparametric projected generalized estimating equation method for panel count data. Biostatistics 2012; 13: 440-454.

12. Zhang Y, Hua L, Huang J. A Spline-Based Semiparametric Maximum Like lihood Estimation Method for the Cox Model with Interval-Censored Data. Scand J Statist. 2010; 37: 338-354.

13. Huang J, Wellner JA. Interval Censored Survival Data: A Review of Recent Progress. In Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis (D. Lin and T. Fleming, eds.) Springer, New York, 1997; 123-169.

14. Ma S, Kosorok MR. Penalized Log-likelihood Estimator for Partly Linear TransformationModels with Current Status Data. Ann Statist. 2005; 33: 2256-2290.

15. Chen X, Fan Y, Tsyrennikov V. Efficient Estimation of Semiparametric Multivariate Copula Models. J Amer Statist Assoc. 2006; 475: 1228-1240.

16. Sun L, Wang L, Sun J. Estimation of the Association for Bivariate Intervalcensored Failure Time Data. Scand J Statist. 2006; 33: 637-649.

Cite this article

Wu Y (2013) B-Spline-Based Sieve Estimation in Survival Analysis. Ann Biom Biostat 1: 1003.