Journal of Multivariate Analysis 132 (2014) 25–38

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Two-sample location–scale estimation from semiparametric random censorship models

Rianka Bhattacharya, Sundarraman Subramanian *

Center for Applied Mathematics and Statistics, Department of Mathematical Sciences, New Jersey Institute of Technology, USA

ARTICLE INFO

Article history:
Received 22 December 2013
Available online 7 August 2014

AMS subject classifications:
62N01
62N02
62N03
62E20
62F03
62F10
62F12

Keywords:
Censoring rate
Cauchy link
Empirical coverage probability
Functional delta method
Gaussian process
Power function

ABSTRACT

When two survival functions belong to a location–scale family of distributions, and the available two-sample data are each right censored, the location and scale parameters can be estimated using a minimum distance criterion combined with Kaplan–Meier quantiles. In this paper, it is shown that using the estimated quantiles from a semiparametric random censorship framework produces improved parameter estimates. The semiparametric framework was originally proposed for the one-sample case (Dikta, 1998), and uses a model for the conditional probability that an observation is uncensored given the observed minimum. The extension to the two-sample setting assumes the availability of good fitting models for the group-specific conditional probabilities. When the models are correctly specified for each group, the new location and scale estimators are shown to be asymptotically as or more efficient than the estimators obtained using the Kaplan–Meier based quantiles. Individual and joint confidence intervals for the parameters are developed. Simulation studies show that the proposed method produces confidence intervals that have correct empirical coverage and that are more informative. The proposed method is illustrated using two real data sets.

1. Introduction

In statistical analysis of survival data, it is often of interest to determine the difference, if any, between two treatment effects. When the treatment-specific populations are known to be each normally distributed and when the two-sample data are each completely uncensored, the standard t-test can be used to discriminate between the treatments. But the normal family is only one member of the location–scale family of distributions, among several others, and the t-test would be inadequate for non-normal families. There may be the additional complication due to censoring which the t-test was not designed to handle. On the other hand, when the underlying distributions are unknown, the Wilcoxon test is used. However, it must be noted that, when the samples come from a location–scale family of distributions, an inferential method that also incorporates the available model information into the analysis should perform better. Indeed, when distributional difference in location and scale is suspected, the Generalized Wilcoxon test for detecting differences is inadequate, see p. 211 of [13].

For the general two-sample problem, the group-specific empirical distribution functions, or their Kaplan–Meier (KM) counterparts in the case of the random censorship model (RCM), provide the basic resource for inference. For two-sample location–scale, Zhang and Li [25] suggested a heuristic method using simple quantiles for inference. Note that in this setting the two continuous distribution functions F_1 and F_2 are related through the equation $F_1(t) = F_2(a + bt)$, where $a \in \mathbb{R}$ and $b > 0$.

* Corresponding author.
E-mail address: sundars@njit.edu (S. Subramanian).

http://dx.doi.org/10.1016/j.jmva.2014.07.011
0047-259X/

For RCM two-sample location–scale inference, the standard method is to employ a minimum distance criterion, more specifically, the Cramér–von Mises type discrepancy involving either the KM estimators of the survival functions, $S_i(t) = 1 - F_i(t),\ i = 1, 2$, or their quantiles. Hsieh [12], however, constructed a regression setup that was based on the KM quantile process and showed that his generalized least squares estimator is semiparametric efficient; see also [11] for the uncensored case. Some limitations, including practical utility of Hsieh's estimator, are pointed out by Potgieter and Lombard [19], however. Koul and Yang [15] applied the Cramér–von Mises type discrepancy to the two KM estimators but focused only on the two-sample scale model; the extension to the location–scale model can be complicated. The Cramér–von Mises type discrepancy combined with quantiles is seen to be a very convenient method for estimating the model parameters a and b, as evidenced by the fact that, under the location–scale model assumption, the quantile functions for the two groups at each point $t \in (0, 1)$ are linearly related [18,26,19]; that is, $Q_2(t) = a + bQ_1(t)$, so that minimizing $\hat{S}(a, b)$, an estimate of

$$S(a, b) = \int \{Q_2(s) - a - bQ_1(s)\}^2 \, dG(s) := E\left(\{Q_2 - a - bQ_1\}^2, G\right), \tag{1.1}$$

where $G(s)$ is a positive measure on $(0, 1)$, presents a viable option, unlike the approach founded on the KM estimators of $S_i(t),\ i = 1, 2$. In Eq. (1.1), $E(\cdot, \cdot)$ denotes the integral of the first argument with respect to the second argument. Zhang and Yu [26] developed estimation of $\boldsymbol{\theta} = (a, b)'$ using Eq. (1.1), where they plugged in the KM quantile function estimators to obtain $\hat{S}(a, b)$, which they minimized to yield $\boldsymbol{\theta}_n$, their estimator of $\boldsymbol{\theta}$. Under some regularity conditions, Zhang and Yu [26] derived the large sample distribution of $\boldsymbol{\theta}_n$ via the delta method combined with standard large sample theory for weighted KM statistics.

In this paper, we showcase the efficacy of utilizing alternate quantiles, obtained from semiparametric survival function estimators, for two-sample location–scale inference. These survival function estimators arise from the framework of semiparametric random censorship models, SRCMs henceforth, introduced by Dikta [4]. Let δ_i denote the censoring indicator for the ith group, $i = 1, 2$. Start with the parametric model $m_i(t, \boldsymbol{\gamma}_i)$, where $m_i(t, \boldsymbol{\gamma}_{i0}) = E(\delta_i | X_i = t)$, and $\boldsymbol{\gamma}_{i0}$ is the true value of $\boldsymbol{\gamma}_i$. Use the ith group sample data to obtain $\hat{\boldsymbol{\gamma}}_i$, the maximum likelihood estimator (MLE) of $\boldsymbol{\gamma}_i \in \Gamma_i \subset \mathbb{R}^k$. The estimated m_i is used as a "surrogate" for the censoring indicator in the ith group, leading to a semiparametric estimator of the group-specific subdistribution function corresponding to uncensored failures. Plugging in this last estimator and the usual "at-risk" function into a standard sequence of mappings [9] yields the group-specific SRCM-based survival function estimator, see Section 2.1 for details.

There are compelling reasons why it would be desirable to incorporate SRCMs into the two-sample location–scale analysis. The KM estimator, although the primary choice under the RCM, ceases to provide optimal performance under the framework of SRCMs. Specifically, Dikta [4] proved that, under correct model specification, the asymptotic variance of the SRCM-based survival function estimator is no greater than that of the KM estimator, with equality attained only in rare and unrealistic cases. In fact, from a more general result derived by Dikta [5] recently, it is now evident that the SRCM-based survival function estimator is asymptotically efficient with respect to the class of all regular estimators under the SRCM. It stands to reason that, if proper parametric models can be identified for each group-specific conditional probability function, inference for the censored two-sample location–scale problem can be improved. Note that the function to be estimated from binary response data is a "success probability" function, for which models are readily available. In reality, fitting the standard Cauchy model to estimate this function appears to produce better estimates most of the times, see [23,17]; see Section 5 for discussion about various possible models for analyzing the binary response data. Furthermore, when the censoring is rather heavy, the KM estimator has fewer jumps leading to a patchy result, which is not a problem with an SRCM-incorporated analysis. Finally, when the censoring indicators are missing at random for a subset of the study subjects [21], RCM-based inference for the location–scale problem becomes difficult compared with our proposed SRCM-based inference. Indeed, with minor modifications, the SRCMs approach readily applies to the case of missing censoring indicators, see Section 5 for some discussion.

We propose to plug in the SRCM-based quantiles into Eq. (1.1) and obtain $\hat{\boldsymbol{\theta}}$ as the minimizer of the resulting criterion function. Under mostly the same regularity conditions as in [26] we derive the limiting distribution of $\hat{\boldsymbol{\theta}}$, from which we are able to obtain confidence intervals for a, b or any function of the two parameters thereof. Numerical results reported in Section 3 indicate that fitting the standard probit or Cauchy link for the binary response data produces estimators with approximately correct coverage and relative reduction over the estimators based on KM quantiles amounting to between 5% and 15%. The Cauchy fit provides the overall coverage closest to the nominal value. A power study confirms the superiority of SRCMs over RCMs. A theoretical analysis of the asymptotic variances reinforces the numerical evidence. Specifically, when the models are correctly specified, we show that the proposed estimators are asymptotically as or more efficient than Zhang and Yu's [26] estimators. Thus, there appears to be strong theoretical and numerical support for SRCMs to be incorporated into censored two-sample location–scale analysis.

The paper is organized as follows. In Section 2, we review the SRCMs and then present our proposed approach. In Section 3, we present our simulation results. In Section 4, we illustrate our method using a mouse leukemia data set [13] and an acute myelogenous data set [14]. In Section 5, we give some concluding remarks.