Published in final edited form as:

Acta Math Sin Engl Ser. 2009 July ; 25(3): 389–398. doi:10.1007/s10255-008-8804-4.

Analysis of Two-sample Censored Data Using a Semiparametric Mixture Model

Gang Li¹ and Chien-tai Lin²

Gang Li: vli@ucla.edu; Chien-tai Lin: chien@math.tku.edu.tw

¹ Department of Biostatistics, University of California, Los Angeles, CA 90095, U.S.A

² Department of Mathematics, Tamkang University, Tamsui 251, Taiwan

Abstract

In this article we study a semiparametric mixture model for the two-sample problem with right censored data. The model implies that the densities for the continuous outcomes are related by a parametric tilt but otherwise unspecified. It provides a useful alternative to the Cox (1972) proportional hazards model for the comparison of treatments based on right censored survival data. We propose an iterative algorithm for the semiparametric maximum likelihood estimates of the parametric and nonparametric components of the model. The performance of the proposed method is studied using simulation. We illustrate our method in an application to melanoma.

Keywords

Biased sampling; EM algorithm; maximum likelihood estimation; mixture model; semiparametric model

1 Introduction

Let X be a nonnegative random variable with distribution functions F_1 and F_2 under treatments 1 and 2, respectively. Consider the following semiparametric mixture model:

$$F_2(x) = \int_0^x \exp[\alpha + \varphi(u; \theta)] dF_1(u), \quad (1)$$

where θ is an unknown parameter of dimension d and, for each θ , $\varphi(x; \theta)$ is a known function satisfying the identifiability constraint $\varphi(0; \theta) = 0$. The nuisance parameter $\alpha = -\log \int \exp[\varphi(u; \theta)] dF_1(u)$ is a normalizing constant.

Model (1) is a useful device for the comparison of treatments. It has the following logistic regression interpretation: define a dummy variable Q that takes value 1 if an individual is from treatment 1, and 0 otherwise; then model (1) is equivalent to

$$\log \left[\frac{P(Q=1|x)}{1 - P(Q=1|x)} \right] = \alpha^* + \varphi(x; \theta),$$

where $\alpha^* = \alpha + \log[P(Q=1)/P(Q=0)]$. Also, if X is a categorical variable with possible values $0, \dots, k$, then model (1) is equivalent to the following logistic regression model

$$\log \left[\frac{P(X=x|F_2)}{P(X=0|F_2)} \right] = \log \left[\frac{P(X=x|F_1)}{P(X=0|F_1)} \right] + \varphi(x; \theta), \text{ for } x=1, \dots, k.$$

The semiparametric mixture model (1) has been used in a variety of applications. Efron and Tibshirani[6] applied it to density estimation. Its robustness and use in ordinal data analysis has been discussed by Anderson[2–4]. Gilbert, Self and Ashby[10] adopted this model to analyze HIV vaccine trial data to assess differential vaccine protection against human immunodeficiency virus types. Fokianos et al.[8] used it to analyze a data set from spaceborne precipitation radar and spaceborne radiometer. The choice of $\varphi(x; \theta)$ has been discussed lucidly by Kay and Little[12]. For example, they found $\varphi(x; \theta) = \theta_1 \log x + \theta_2 \log(1 - x)$, $\theta = (\theta_1, \theta_2)'$, to be a good choice in an appreciation to age of menarche in girls from Warsaw. Storer, Wacholder and Breslow[14] observed that the choice of $\varphi(x; \theta) = \log(1 + x^T \theta)$ is useful in some epidemiological problems. Vardi[15,16], Gill, Vardi and Wellner[11] and others studied a bias sampling model that is a special case of model (1) with a known θ . Qin[13] used a profile empirical likelihood to construct confidence intervals for the mixture proportions and the distributions when $\varphi(x; \theta) = \theta x$. A nice discussion of model (1) can be found in Gilbert, Lele and Vardi[9] who studied the identifiability of this model and derived semiparametric maximum likelihood estimates of (θ, F_1) for complete data. Recently Zou, Fine, and Yandell [17] suggested a partial profile empirical likelihood to estimate θ and F_1 when $\varphi(x; \theta) = \theta x$ in the study of genetic loci influencing quantitative traits.

In clinical trials and epidemiological followup studies, the survival time of interest is often subject to censoring. The main objective of this work is to study semiparametric maximum likelihood inference for (θ, F_1) based on right censored survival data. Using a profile likelihood, we propose an iterative algorithm to obtain the semiparametric maximum likelihood estimates of the parametric component θ and the nonparametric distribution F_1 simultaneously. Our derivation begins with the case of complete data, and then is extended to right censored data using the EM algorithm principle. We propose a likelihood ratio test for the difference between two treatments and a graphical method for goodness-of-fit of model (1). The method of this paper provides an alternative to the Cox model when the proportional hazards assumption is not met.

The article is organized as follows. In Section 2 we discuss semiparametric maximum likelihood estimation and inference for model (1). In Section 3 we present a simulation study to evaluate the performance of the estimates and apply our method to a melanoma data set. We offer concluding remarks in Section 4.

2 Maximum Likelihood Estimation and Inference

In this section we develop an iterative algorithm to derive the maximum likelihood estimates of θ and F_1 for model (1). We begin with a derivation based on complete data samples and then extend the results to censored data samples.

Let X and Y be the lifetime random variables, and C and D be the corresponding censoring time random variables. Suppose that X , C , Y , and D are independent with unknown distribution functions F_1 , G_1 , F_2 , and G_2 , respectively. Define $U = \min(X, C)$, $V = \min(Y, D)$, and two indicator random variables $\Delta = I(X < C)$ and $\Gamma = I(Y < D)$. Assume that F_1 and F_2 satisfy model (1).

2.1 Complete Data

We first describe maximum likelihood estimation for model (1) with complete data, which serves as a stepping stone for studying the right censored data.

Let $(x_1, \dots, x_{n_1})$ and $(y_1, \dots, y_{n_2})$ be two independent random samples from F_1 and F_2 , respectively. Let $t_1, \dots, t_h$ represent the distinct values among the $n_1 + n_2$ observations. Define τ_{1j} and τ_{2j} to be the numbers of observations with value t_j in samples 1 and 2, respectively.

The Likelihood—Under model (1), the likelihood is

$$L(\theta, F_1) = \prod_{i=1}^{n_1} dF_1(x_i) \prod_{i=1}^{n_2} \frac{w(y_i; \theta) dF_1(y_i)}{\int_0^\infty w(y; \theta) dF_1(y)} = \prod_{j=1}^h [dF_1(t_j)]^{\tau_{1j}} \prod_{j=1}^h \left[\frac{w(t_j; \theta) dF_1(t_j)}{\int_0^\infty w(y; \theta) dF_1(y)} \right]^{\tau_{2j}},$$

where $w(u; \theta) = \exp(\varphi(u; \theta))$. Define $p_j = dF_1(t_j)$, $j = 1, 2, \dots, h$; then the likelihood can be rewritten as

$$\begin{aligned} L(\theta, F_1) &= \prod_{j=1}^h p_j^{\tau_{1j} + \tau_{2j}} \left[\int_0^\infty w(y; \theta) dF_1(y) \right]^{-n_2} \prod_{j=1}^h w^{\tau_{2j}}(t_j; \theta) \\ &= \prod_{j=1}^h p_j^{\tau_j} \left[\sum_{j=1}^h w(t_j; \theta) p_j \right]^{-n_2} \prod_{j=1}^h w^{\tau_{2j}}(t_j; \theta), \end{aligned} \quad (2)$$

where $\tau_j = \tau_{1j} + \tau_{2j}$. Maximizing $L(\theta, F_1)$ subject to the constraints $\sum_{j=1}^h p_j = 1$ and $p_j \geq 0$, we have

$$p_j(\theta) = \left[\frac{\tau_j}{w(t_j; \theta)} \right] / \left[\sum_{\ell=1}^h \frac{\tau_\ell}{w(t_\ell; \theta)} \right] \quad \text{for } j=1, \dots, h. \quad (3)$$

Plugging (3) into (2), we obtain the following profile likelihood function for θ

$$\begin{aligned} L^{pr}(\theta) &= \prod_{j=1}^h \left[\frac{\frac{\tau_j}{w(t_j; \theta)}}{\sum_{\ell=1}^h \frac{\tau_\ell}{w(t_\ell; \theta)}} \right]^{\tau_j} \left[\frac{\frac{n_1 + n_2}{h}}{\sum_{\ell=1}^h \frac{\tau_\ell}{w(t_\ell; \theta)}} \right]^{-n_2} \prod_{j=1}^h w^{\tau_{2j}}(t_j; \theta) \\ &= (n_1 + n_2)^{-n_2} \left[\prod_{j=1}^h \frac{\tau_j^{\tau_j}}{w^{\tau_j}(t_j; \theta)} \right] \left[\sum_{\ell=1}^h \frac{\tau_\ell}{w(t_\ell; \theta)} \right]^{-n_1}. \end{aligned}$$

The logarithm of the profile likelihood function is

$$\ell^{pr}(\theta) = - \sum_{j=1}^h \tau_{1j} \log w(t_j; \theta) - n_1 \log \left[\sum_{\ell=1}^h \frac{\tau_\ell}{w(t_\ell; \theta)} \right] + \text{Constant}.$$

Estimation—Let $\hat{\theta}$ be a value of θ that maximizes $\ell^{pr}(\theta)$. The maximum likelihood estimate of F_1 is then

$$\widehat{F}_1(t) = \sum_{j=1}^h p_j(\widehat{\theta}) I(t_j \leq t).$$

2.2 Right Censored Data

Data and Likelihood—Suppose that one observes two independent right censored samples

$$(u_1, \delta_1), \dots, (u_{n_1}, \delta_{n_1}) \text{ and } (v_1, \gamma_1), \dots, (v_{n_2}, \gamma_{n_2}),$$

where $(u_1, \delta_1), \dots, (u_{n_1}, \delta_{n_1})$ are i.i.d. replicates of (U, Δ) and $(v_1, \gamma_1), \dots, (v_{n_2}, \gamma_{n_2})$ are i.i.d. replicates of (V, Γ) , respectively. Let $t_1, \dots, t_h$ represent the distinct uncensored lifetimes from the pooled sample. For $j = 1, \dots, h$, let

$$\xi_{1j} = \sum_{i=1}^{n_1} I(u_i = t_j, \delta_i = 1), \quad \xi_{2j} = \sum_{i=1}^{n_2} I(v_i = t_j, \gamma_i = 1),$$

and

$$\eta_{1j} = \sum_{i=1}^{n_1} I(u_i = t_j, \delta_i = 0), \quad \eta_{2j} = \sum_{i=1}^{n_2} I(v_i = t_j, \gamma_i = 0).$$

Then, the likelihood for (θ, F_1) is proportional to

$$\begin{aligned} L_c(\theta, F_1) &= \prod_{i=1}^{n_1} [dF_1(u_i)]^{\delta_i} [1 - F_1(u_i)]^{1-\delta_i} \prod_{i=1}^{n_2} [dF_2(v_i)]^{\gamma_i} [1 - F_2(v_i)]^{1-\gamma_i} \\ &= \prod_{j=1}^h [dF_1(t_j)]^{\xi_{1j}} [1 - F_1(t_j)]^{\eta_{1j}} [dF_2(t_j)]^{\xi_{2j}} [1 - F_2(t_j)]^{\eta_{2j}}. \end{aligned}$$

Define $p_j = dF_1(t_j)$, $j = 1, 2, \dots, h$, then the likelihood can be reexpressed as

$$\begin{aligned} L_c(\theta, F_1) &= \prod_{j=1}^h p_j^{\xi_{1j}} \left(\sum_{k=j+1}^h p_k \right)^{\eta_{1j}} \left[\frac{w(t_j; \theta) p_j}{\sum_{i=1}^h w(t_i; \theta) p_i} \right]^{\xi_{2j}} \left[\frac{\sum_{k=j+1}^h w(t_k; \theta) p_k}{\sum_{i=1}^h w(t_i; \theta) p_i} \right]^{\eta_{2j}} \\ &= \prod_{j=1}^h p_j^{\xi_{1j} + \xi_{2j}} \left(\sum_{k=j+1}^h p_k \right)^{\eta_{1j}} w^{\xi_{2j}}(t_j; \theta) \left[\sum_{k=j+1}^h w(t_k; \theta) p_k \right]^{\eta_{2j}} \left[\sum_{i=1}^h w(t_i; \theta) p_i \right]^{-(\xi_{2j} + \eta_{2j})}. \end{aligned}$$

The log-likelihood is

$$\ell_c = \log L_c = \sum_{j=1}^h \left\{ (\xi_{1j} + \xi_{2j}) \log p_j + \eta_{1j} \log \left(\sum_{k=j+1}^h p_k \right) + \xi_{2j} \log w(t_j; \theta) + \eta_{2j} \log \left[\sum_{k=j+1}^h w(t_k; \theta) p_k \right] - (\xi_{2j} + \eta_{2j}) \log \left[\sum_{i=1}^h w(t_i; \theta) p_i \right] \right\}. \quad (4)$$

Estimation—Let $\mathbf{p} = (p_1, \dots, p_h)'$. It is not easy to maximize the log-likelihood function ℓ_c in (4) with respect to $(\mathbf{p}, \theta)$ because the dimension of $(\mathbf{p}, \theta)$ depends on the sample size which may be large. Instead of maximizing ℓ_c directly, we exploit its structure and develop a simple EM algorithm to solve the high dimensional maximization problem.

E-step: Calculate the expected conditional log-likelihood based on the hypothetical complete data given the observed right censored data and the parameter estimates from the previous iteration. Because the complete data log-likelihood is linear in τ_{1j} and τ_{2j} , we simply need to replace them in equation (3) by $\tau_{ij}^{(s)} = E(\tau_{ij} | \text{observed data}, \mathbf{p}^{(s-1)}, \theta^{(s-1)})$, $i = 1, 2$ and $j = 1, \dots, h$, at the s -th iteration.

For sample 1, we have

$$E(\tau_{1j} | \text{observed data}, \mathbf{p}, \theta) = \sum_{i=1}^{n_1} E[I(X_i = t_j, \Delta_i = 1) | (U_i, \Delta_i) = (u_i, \delta_i), \mathbf{p}, \theta] + \sum_{i=1}^{n_1} E[I(X_i = t_j, \Delta_i = 0) | (U_i, \Delta_i) = (u_i, \delta_i), \mathbf{p}, \theta]. \quad (5)$$

Note that the first summation in (5) equals ξ_{1j} . Moreover, the i th term in the second summation of (5) is

$$\begin{aligned} & I(\delta_i = 0) I(u_i \leq t_j) \frac{P(X_i = t_j, C_i = u_i | \mathbf{p}, \theta)}{\sum_{\ell: t_\ell \geq u_i} P(X_i = t_\ell, C_i = u_i | \mathbf{p}, \theta)} \\ &= I(\delta_i = 0) I(u_i \leq t_j) \frac{dF_1(t_j) dG_1(u_i)}{\sum_{\ell: t_\ell \geq u_i} dF_1(t_\ell) dG_1(u_i)} \\ &= I(u_i \leq t_j, \delta_i = 0) \frac{p_j}{\sum_{\ell: t_\ell \geq u_i} p_\ell} \\ &= \sum_{k=1}^j I(u_i = t_k, \delta_i = 0) \frac{p_j}{\sum_{\ell: t_\ell \geq u_i} p_\ell}. \end{aligned}$$

Thus, (5) can be reexpressed as

$$\begin{aligned}
E(\tau_{1j}|\text{observed data}, \mathbf{p}, \theta) &= \xi_{1j} + p_j \sum_{k=1}^j \sum_{i=1}^{n_1} \frac{I(u_i = t_k, \delta_i = 0)}{\sum_{\ell: t_\ell \geq u_i} p_\ell} \\
&= \xi_{1j} + p_j \sum_{k=1}^j \frac{\sum_{i=1}^{n_1} I(u_i = t_k, \delta_i = 0)}{\sum_{\ell: t_\ell \geq t_k} p_\ell} \\
&= \xi_{1j} + p_j \sum_{k=1}^j \frac{\eta_{1k}}{\sum_{\ell=k}^h p_\ell}.
\end{aligned} \tag{6}$$

Similarly, we obtain

$$E(\tau_{2j}|\text{observed data}, \mathbf{p}, \theta) = \xi_{2j} + p_j \sum_{k=1}^j \frac{\eta_{2k}}{\sum_{\ell=k}^h p_\ell} \tag{7}$$

for sample 2.

M-step: Find the value $\hat{\theta}^{(s)}$ that maximizes

$$\ell^{pr}(\theta) = - \sum_{j=1}^h \tau_{1j}^{(s)} \log w(t_j; \theta) - n_1 \log \left[\sum_{l=1}^h \frac{\tau_l^{(s)}}{w(t_l; \theta)} \right], \tag{8}$$

and update p_j by

$$p_j^{(s)} = \left[\frac{\tau_j^{(s)}}{w(t_j; \hat{\theta}^{(s)})} \right] / \left[\sum_{l=1}^h \frac{\tau_l^{(s)}}{w(t_l; \hat{\theta}^{(s)})} \right], \quad j=1, \dots, h.$$

The algorithm iterates between the E-step and the M-step until convergence.

Remark: The above algorithm can be carried out by updating θ and $\mathbf{p}$ simultaneously until convergence. An alternative approach is to update $\mathbf{p}$ only through the EM-algorithm until convergence for each fixed value of θ , and then maximize the resulting profile likelihood based on right censored data with respect to θ . The latter approach is computationally less demanding and works well in our simulations.

To test $H_0: \theta = \theta_0$, we define the following profile likelihood ratio statistic

$$\chi_{LR}^2 = -2[\ell_c^{pr}(\hat{\theta}) - \ell_c^{pr}(\theta_0)].$$

We reject H_0 at level α if $\chi_{LR}^2 > \chi_d^2(\alpha)$, where $\chi_d^2(\alpha)$ is the upper α -quantile of the chi-squared distribution with d degrees of freedom. The chi-squared approximation to the likelihood ratio statistic is examined empirically in Section 3. Consequently, a $100(1 - \alpha)\%$ confidence region for θ is defined as

$$C_\alpha = \{\theta_0: \chi_{LR}^2 \leq \chi_d^2(\alpha)\}.$$

3 Numerical Studies

In this section we investigate the performance of the estimate of θ and the likelihood ratio test through simulation. We illustrate our method in an application to melanoma.

3.1 A Simulation Study

We assume that $F_1(x) = 1 - \exp(-x)$, $F_2(x) = 1 - \exp[-(1 - \theta)x]$, $G_1(x) = 1 - \exp[-(1/\alpha_1 - 1)x]$, and $G_2(x) = 1 - \exp[-(1/\alpha_2 - 1)x]$ under the censoring rates $1 - \alpha_1 = P(X > C)$ and $1 - \alpha_2 = P(Y > D)$. Samples were generated from these distributions for values of θ ranging from -0.6 to 0.6 , $(n_1, n_2) = (50, 50)$ and $(100, 100)$, $(\alpha_1, \alpha_2) = (0.20, 0.20)$, $(0.20, 0.50)$, $(0.50, 0.20)$ and $(0.50, 0.50)$. There were 1000 replications for each combination. Table 1 gives the bias and mean squared error of the estimate $\hat{\theta}$; it performs quite well. The variance appears to be the main contributor to the mean squared error. As one would expect, the mean squared error grows larger as the censoring rate increases.

We performed a second simulation to evaluate the likelihood ratio test described in Section 2.2. Table 2 lists the simulated rejection probabilities for testing $H_0: \theta = \theta_0$ for different values of θ_0 at three nominal levels: 10%, 5% and 1% (denoted as a , b , and c , respectively). The likelihood ratio test performed well even for heavy censoring. As expected, the rejection percentages are closer to the nominal level when the censoring rates are small and the sample sizes are large.

Finally, we examined the chi-squared approximation of the likelihood ratio statistic using the Quantile-Quantile (Q-Q) plot. We generated samples from an exponential mixture model and a normal mixture model with sample sizes of 100 in each arm ($n_1 = n_2 = 100$) and the censoring rates of 80% and 50%. Figure 1 depicts the Q-Q plots for different models, each based on 1,000 Monte Carlo samples. The Q-Q plots are all close to a straight line, suggesting that the chi-squared approximation of the likelihood ratio statistic is apt.

3.2 An Example

We demonstrate our method in an analysis of a real data consisting of 205 patients with malignant melanoma (cancer of the skin) at Odense University Hospital at Denmark in the period 1962–77. A detailed description of this data set can be found in Andersen et al.[1]. Previous studies have suggested that sex and tumor thickness are important prognostic factors for survival. Exponential distributions were used to model the survival times by Andersen et al.[1]. Here we consider the semiparametric mixture model (1) with $\varphi(x; \theta) = \theta x$ for each of the two prognostic factors, which is more flexible than the exponential model. For sex, males are in group 1 and females are in group 2. For tumor thickness, a patient is in group 1 if the tumor thickness is less than 2 mm, and group 2 otherwise.

We first check the goodness-of-fit of the semiparametric mixture model for each of the two factors. Figure 2 depicts the estimated distribution F_1 under the semiparametric mixture model

together with the Kaplan-Meier estimate. The concordance of the two curves suggests that our models are appropriate.

The maximum likelihood estimate of θ is 0.077 for the sex effect and -0.184 for tumor thickness. This suggests that female patients tend to survive longer than male patients and that patients with thicker tumor tend to have shorter survival. The likelihood ratio statistic for testing $H_0: \theta = 0$ is 7.249358 for sex and 35.741389 for tumor thickness, respectively. They both exceed $\chi_1^2(0.01)=6.63$, concluding that the effects of both sex and tumor thickness on survival are statistically significant.

Finally, we obtain a 95% confidence interval for θ for each of the two models. For sex, a 95% confidence interval for θ is (0.02, 0.12). For tumor thickness, a 95% confidence interval for θ is $(-0.26, -0.12)$.

4 Concluding Remarks

We investigate maximum likelihood estimation and inference for the semiparametric mixture model (1) with right censored survival data. An iterative EM algorithm is derived to compute the maximum likelihood estimates. Likelihood ratio inference is suggested. Empirical evidence suggests that the method performs well. Although we focus on the two-sample case in the paper, the semiparametric mixture model can be naturally extended to the general k -sample problem as described in Gilbert, Lele and Vardi[9]. Large sample theory for model (1) remains an open problem and warrants future research.

Acknowledgments

Gang Li's research was supported in part by the U.S. National Institute of Health (No. CA016042, No. P01AT003960). Chien-Tai Lin's research was supported in part by the National Science Council of Taiwan (No. 89-2118-M-032-021, No. 96-2628-M-032-002-MY3).

We thank the editor and an anonymous referee for making valuable suggestions which led to considerable improvement in the presentation of this paper.

References

1. Andersen, PK.; Borgan, O.; Gill, RD.; Keiding, N. Statistical models based on counting processes. Springer-Verlag; New York: 1993.
2. Anderson JA. Robust inference using logistic models. *Bull Internat Statist Inst* 1979;48:35–53.
3. Anderson JA. Multivariate logistic compounds. *Biometrika* 1979;66:17–26.
4. Anderson JA. Regression and ordered categorical variables (with discussion). *Journal of the Royal Statistical Society (Series B)* 1984;46:1–30.
5. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society (Series B)* 1972;34:187–220.
6. Efron B, Tibshirani R. Using specially designed exponential families for density estimation. *Annals of Statistics* 1996;24:2431–2461.
7. Fang HB, Li G, Sun J. Maximum likelihood estimation in a semiparametric logistic/proportional hazards mixture model. *Scandinavian Journal of Statistics* 2005;32:69–75.
8. Fokianos K, Kedem B, Qin J, Haferman JL, Short DA. On combining instruments. *Journal of Applied Meteorology* 1998;37:220–226.
9. Gilbert PB, Lele SR, Vardi Y. Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika* 1999;86:27–43.
10. Gilbert PB, Self S, Ashby A. Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types. *Biometrics* 1998;54:799–814. [PubMed: 9750238]

11. Gill RD, Vardi Y, Wellner JA. Large sample theory of empirical distributions in biased sampling models. *Annals of Statistics* 1988;16:1069–1112.
12. Kay R, Little S. Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika* 1987;74:495–501.
13. Qin J. Empirical likelihood ratio based confidence intervals for mixture proportions. *Annals of Statistics* 1999;27:1368–1384.
14. Storer BE, Wacholder S, Breslow N. Maximum likelihood fitting of general risk models to stratified data. *Applied Statistics* 1983;32:172–181.
15. Vardi Y. Nonparametric estimation in presence of length bias. *Annals of Statistics* 1982;10:616–620.
16. Vardi Y. Empirical distribution in selection bias models. *Annals of Statistics* 1985;13:178–203.
17. Zou F, Fine JP, Yandell BS. On empirical likelihood for a semiparametric mixture model. *Biometrika* 2002;89:61–75.

Figure 1. Quantile-Quantile plots of the likelihood ratio statistic under an exponential mixture model and a normal mixture model.

Figure 2. The estimated distribution function under the semiparametric mixture model (circled lines), and the Kaplan-Meier estimate (step function) for the malignant melanoma data.

Simulated bias and mean squared error of $\hat{\theta}$ under a semiparametric exponential mixture model: the distributions of the two survival times are $F_1(x) = 1 - \exp(-x)$ and $F_2(x) = 1 - \exp[-(1 - \theta)x]$; and the corresponding censoring distributions are $G_1(x) = 1 - \exp[-(1/a_1 - 1)x]$ and $G_2(x) = 1 - \exp[-(1/a_2 - 1)x]$ with censoring rates $1 - a_1 = P(X > C)$ and $1 - a_2 = P(Y > D)$. Sample sizes are n_1 and n_2 .

Table 1

n_1	n_2	$1 - a_1$	$1 - a_2$	θ						
				$-.6$	$-.4$	$-.2$	0	$.2$	$.4$	$.6$
50	50	.80	bias	-.00605	-.08699	-.08865	-.03022	.06231	.13708	.07937
			mse	.30110	.36285	.42651	.42514	.36387	.22160	.12013
		.80	bias	-.04873	-.06842	-.05506	-.00745	.00997	.02129	-.00665
			mse	.17714	.23515	.24304	.23343	.17912	.14024	.10232
		.50	bias	.03179	-.03519	-.05979	-.04710	.00125	.07243	.12863
			mse	.19766	.22665	.24197	.22583	.18594	.13710	.08331
	100	.80	bias	-.04283	-.04024	-.06471	-.01128	.00925	.03313	.01478
			mse	.11431	.07792	.14441	.13165	.10240	.08107	.06038
		.80	bias	-.07560	-.09854	-.08370	-.00588	.09852	.15437	.11512
			mse	.15955	.22631	.26701	.24060	.20054	.14062	.08130
		.50	bias	-.05222	-.06761	-.02305	.00638	.02401	.03890	.02397
			mse	.11018	.13164	.14037	.11284	.09799	.08971	.06842
100	.80	bias	.02253	-.02466	-.03571	-.01143	.02591	.08630	.15297	
		mse	.12384	.13910	.13946	.11262	.09035	.07747	.06295	
	.50	bias	-.02787	-.04697	-.02673	-.00451	.02883	.02614	.02777	
		mse	.07477	.08031	.07155	.05451	.04649	.04023	.03686	

Table 2

Simulated rejection probabilities for testing $H_0: \theta = \theta_0$ for different values of θ_0 and three nominal levels 10%, 5% and 1% under a semiparametric exponential mixture model: Distributions of the two samples are $F_1(x) = 1 - \exp(-x)$ and $F_2(x) = 1 - \exp[-(1 - \theta_0)x]$; and the corresponding censoring distributions are $G_1(x) = 1 - \exp[-(1/a_1 - 1)x]$ and $G_2(x) = 1 - \exp[-(1/a_2 - 1)x]$ with censoring rates $1 - a_1 = P(X > C)$ and $1 - a_2 = P(Y > D)$. Sample sizes are n_1 and n_2 .

n_1	n_2	$1 - a_1$	$1 - a_2$	θ_0							
				$-.6$	$-.4$	$-.2$	0	$.2$	$.4$	$.6$	
50	50	.80		a	.027	.041	.066	.071	.078	.090	.067
				b	.011	.009	.020	.025	.026	.020	.020
				c	.003	.002	.001	.003	.001	.001	.006
		.50		a	.037	.095	.105	.107	.097	.092	.070
				b	.012	.032	.045	.049	.045	.041	.023
				c	.004	.006	.005	.009	.006	.005	.005
	100	.80		a	.045	.074	.098	.110	.113	.122	.130
				b	.022	.029	.045	.055	.060	.060	.048
				c	.004	.004	.006	.008	.006	.005	.002
		.50		a	.069	.111	.127	.118	.115	.107	.105
				b	.022	.060	.058	.064	.061	.061	.045
				c	.001	.012	.009	.010	.013	.009	.007
100	100	.80		a	.047	.085	.107	.097	.105	.131	.119
				b	.010	.026	.040	.053	.051	.059	.042
				c	.003	.003	.007	.010	.013	.014	.008
		.50		a	.103	.111	.129	.113	.109	.118	.080
				b	.045	.064	.071	.062	.070	.063	.043
				c	.000	.008	.018	.014	.021	.017	.005
	100	.80		a	.066	.109	.120	.105	.103	.135	.220
				b	.023	.050	.063	.058	.062	.073	.125
				c	.005	.006	.011	.013	.016	.015	.026
		.50		a	.095	.107	.120	.102	.097	.110	.103
				b	.039	.066	.069	.048	.053	.061	.058

n_1	n_2	$1 - \alpha_1$	$1 - \alpha_2$	θ_0						c
				$-.6$	$-.4$	$-.2$	0	$.2$	$.4$	$.6$
				.004	.011	.010	.011	.016	.015	.009

$^a \chi^2_1(.10)=2.71,$
 $^b \chi^2_1(.05)=3.84,$
 $^c \chi^2_1(.01)=6.63$