Pairwise rank based likelihood for estimating the relationship between two homogeneous populations and their mixture proportion

Glenn Heller and Jing Qin

Department of Epidemiology and Biostatistics

Memorial Sloan-Kettering Cancer Center

1275 York Avenue, New York, New York 10021

Summary

We consider the problem of inferring the relationship between two homogeneous populations, and the prevalence of each population. Sample data from the two distributions as well as from a third population consisting of a mixture of the two will be used. Under the transformation model assumption on the two distribution functions, we develop a pairwise rank based likelihood. Simultaneous inference on the mixture proportion and the transformation parameter defining the relationship between the two populations is based on this likelihood. Under some regularity conditions, it is shown that the maximum pairwise rank likelihood estimator is consistent and has an asymptotic normal distribution. Simulation results indicate that the performance of this statistic is satisfactory. The methodology is demonstrated on a data set in prostate cancer.

Some key words: Lehmann alternative; Mixture proportion; Pairwise rank likelihood; Semiparametric model; Transformation parameter.

Introduction

We consider the two sample problem where the group identifier is missing on a subset of the observations. In this setting, we identify the outcome data as being generated from either one of two distinct univariate distributions (F, G) or from the mixture distribution $H = \lambda F + (1 - \lambda)G$, $0 < \lambda < 1$. It is presumed that the missing group identification is unrelated to either the subject's outcome or group affiliation. Our objective is to perform inference on the relationship between the distribution functions F and G and the mixture parameter λ.

Our interest in this problem emanated from an application in cancer research, where significant effort is underway to develop novel therapeutics directed at specific genetic targets on the cancer cell. Genetic profiling of cancer cells is now a reality, and with this technology comes an understanding of the genetic control of the cancer cell. The ultimate objective in implementing this technology is to identify and then degrade those genes that control the cancer cell and thus improve the clinical outcome of the patient. One path in this research, is to empirically explore the association between oncogenes and adverse clinical outcome. We anticipate that genes associated with poor clinical outcome and prevalent in the cancer population will become the primary targets of the cancer therapy.

While the use of genetic profiling holds great promise for the future, currently this technology is applied to only a small subset of the cancer population due to the high cost of the technology, as well as limited access to the tumor cell. However, there is available clinical outcome data on all subjects which can be used to infer genetic information on the larger population not included in the gene studies. Focusing on a single gene for the purpose of exposition, subjects who are tested and express the gene are members of population F, sub-

jects who do not express the gene are members of G, and the remaining patients who are not tested belong to the mixture population H. Thus we have three independent data sets:

$$\begin{aligned} &x_1, ..., x_{n_1}, \quad \textit{iid with distribution } F(x), \\ &y_1, ..., y_{n_2}, \quad \textit{iid with distribution } G(y), \\ &z_1, ..., z_{n_3}, \quad \textit{iid with distribution } H(z) = \lambda F(x) + (1-\lambda)G(x). \end{aligned} \tag{1.1}$$

The corresponding density functions are denoted by $f(x) = dF(x)/dx$, $g(y) = dG(y)/dy$ and $h(z) = dH(z)/dz$, respectively.

The problem of inference with this type of mixture data was studied by Hosmer (1973), using normality assumptions on F and G. Bayes and nonparametric estimation of mixing proportion have been discussed by Murray & Titterington (1978), Hall (1981), Titterington (1983) & Hall and Titterington (1984). Anderson (1979) proposed the semiparametric model

$$g(x) = exp(\beta_0 + x\beta_1)f(x), \tag{1.2}$$

where $f(x)$ is an arbitrary density function. The data in (1.1) then come from distributions

$$dF(x), \qquad exp(\beta_0 + x\beta_1)dF(x), \qquad [\lambda + (1-\lambda)exp(\beta_0 + x\beta_1)]dF(x) \tag{1.3}$$

respectively.

The attractive feature of (1.2) compared with the normal mixture model is that the distributions are modeled nonparametrically, except for a parametric "exponential tilt" that is used to relate one distribution to the other. This is similar to the Cox proportional hazards model and the Lehmann two sample alternative model, where the ratio of the two hazard functions is assumed to have a known parametric form (Lehmann 1953, Cox 1972).

In this paper, we assume that F and G are related through a transformation function

$$\bar{G}(x) = C(\bar{F}(x), \theta) \tag{1.4}$$

where $\bar{G}(x) = 1 - G(x)$, and $C(u, \theta)$ is a continuous distribution function on $(0, 1)$ whose functional form is known and θ is a scalar parameter. Furthermore, we assume that $C(u, \theta)$ is monotonically increasing in θ and continuously differentiable in both u and θ. The transformation model includes many familiar semiparametric models. For example, $C(u, \theta) = u^{\theta}$ corresponds to the Cox proportional hazards models or the Lehmann two sample alternative model, and $C(u, \theta) = \theta u/(1 - u + \theta u)$, corresponds to the proportional odds ratio model (Bennett 1983 and Pettitt 1984). Based on a specified transformation function, likelihood based estimation and inference of the transformation and mixture parameters (θ, λ) is proposed. It is anticipated that simultaneous inference will result in additional information when compared to methodology that examines each parameter separately.

This paper is organized as follows. In section 2, we develop the pairwise rank based likelihood. Maximum pairwise rank likelihood estimation is proposed for the mixture parameter λ and the transformation parameter θ. An asymptotic normal distribution is derived for the pairwise rank likelihood estimates. In section 3, simulation studies are undertaken to examine the adequacy of the inference procedure for realistic sample sizes. An example using prostate cancer patient data is presented in section 4. We conclude in section 5 with a discussion of possible extensions of the proposed methodology.

2. Main results

Since the baseline survival function $\bar{F}$ is not specified in the transformation model (1.4), it is desirable to make inference on λ and θ without using the form of $\bar{F}(x)$. A natural approach is to consider a rank based method. Note that

$$P(Y > X) = E\{\bar{G}(X)\} = -\int \bar{G}(x) d\bar{F}(x) = -\int C(\bar{F}(x), \theta) d\bar{F}(x) = D(\theta),$$

$$P(Z > X) = -\int \{\lambda \bar{F}(x) + (1-\lambda)\bar{G}(x)\} d\bar{F}(x) = 0.5\lambda + (1-\lambda)D(\theta)$$

and

$$P(Z > Y) = -\int \{\lambda \bar{F}(x) + (1-\lambda)\bar{G}(x)\} d\bar{G}(x) = \lambda D_1(\theta) + 0.5(1-\lambda),$$

where

$$D(\theta) = -\int C(\bar{F}(x), \theta) d\bar{F}(x) = \int_0^1 C(u, \theta) du,$$

$$D_1(\theta) = -\int \bar{F}(x) dC(\bar{F}, \theta) = \int_0^1 u dC(u, \theta) = C(1, \theta) - D(\theta).$$

Denote $p_1(\theta) = P(Y > X), p_2(\lambda, \theta) = P(Z > X)$ and $p_3(\lambda, \theta) = P(Z > Y)$, then the pairwise rank likelihood is constructed as

$$\begin{aligned} L_P &= \prod_{i,j} \{p_1(\theta)\}^{I(y_j > x_i)} \{1 - p_1(\theta)\}^{I(y_j \le x_i)} \prod_{i,k} \{p_2(\lambda, \theta)\}^{I(z_k > x_i)} \\ &\quad \prod_{i,k} \{1 - p_2(\lambda, \theta)\}^{I(z_k \le x_i)} \prod_{j,k} \{p_3(\lambda, \theta)\}^{I(z_k > y_j)} \prod_{j,k} \{1 - p_3(\lambda, \theta)\}^{I(z_k \le y_j)}, \end{aligned} \tag{2.1}$$

where $I(\cdot)$ is indicator function, and $i = 1, 2, .., n_1$, $j = 1, 2, .., n_2$ and $k = 1, 2, ..., n_3$. Thus the log pairwise rank likelihood is written as

$$\begin{aligned} l_P &= \sum_{i,j} I(y_j > x_i) \log p_1(\theta) + \sum_{i,j} I(y_j \le x_i) \log\{1 - p_1(\theta)\} + \sum_{i,k} I(z_k > x_i) \log p_2(\lambda, \theta) \\ &+ \sum_{i,k} I(z_k \le x_i) \log\{1 - p_2(\lambda, \theta)\} + \sum_{j,k} I(z_k > y_j) \log p_3(\lambda, \theta) + \sum_{j,k} I(z_k \le y_j) \log\{1 - p_3(\lambda, \theta)\} \end{aligned}$$

and the maximum pairwise rank likelihood estimates $(\hat{\lambda}, \hat{\theta})$ satisfy the score equations

$$\frac{\partial l_P}{\partial \lambda} = 0, \qquad \frac{\partial l_P}{\partial \theta} = 0.$$

For convenience, denote

$$\eta = (\theta^T, \lambda)^T, \qquad \hat{\eta} = (\hat{\theta}^T, \hat{\lambda})^T.$$

Let $n = n_1 + n_2 + n_3$, $n_i/n \to \nu_i, i = 1, 2, 3$, where $0 < \nu_i < 1$.

THEOREM 1 *Under some regularity conditions,*

$$\sqrt{n}(\hat{\eta} - \eta) \to N(0, \Sigma) \tag{2.2}$$

in distribution, where

$$\Sigma = A^{-1}(\eta)B(\eta)A^{-1}(\eta)$$

and A, B are given in the Appendix.

The proof of Theorem 1 will be given in the Appendix.

3. Simulation results

To examine the adequacy of the proposed method a simulation experiment was performed. Three independent samples from three distinct populations were generated. Two samples of size 50 were generated from exponential distributions. The first sample was drawn from a unit exponential and the second from an exponential distribution with scale parameter taking values 2,4,6 and 8. The third sample, derived as a mixture of the two exponentials, consisted of 200 observations. The mixture parameter ranged from .20 to .80. There were 5000 replications for each simulation run.

The simulation structure represents the Lehmann two-sample alternative model, $1 - G(x) = \{1 - F(x)\}^{\theta}$, or using the transformation function representation $C(u, \theta) = u^{\theta}$. Under this model, the pairwise probabilities in the likelihood are evaluated as

$$p_1(\theta) = \frac{1}{\theta + 1}, \quad p_2(\lambda, \theta) = 0.5\lambda + (1 - \lambda)\frac{1}{\theta + 1}, \quad p_3(\lambda, \theta) = \frac{\lambda\theta}{\theta + 1} + 0.5(1 - \lambda).$$

The results of the simulations presented in Table 1 demonstrate that the likelihood based estimation and asymptotic inference procedure for the transformation parameter θ and mixture parameter λ is accurate. The bias of the mixture parameter estimate is small over the

range of (θ, λ), with the bias improving as θ moves away from 1. In contrast, the bias of $\hat{\theta}$ becomes considerable as θ increases. This pattern holds throughout the range of λ. The empirical coverage probabilities of the asymptotic 95% confidence intervals are all close to the nominal .95 level.

4. Prostate Cancer Example

A fundamental tenet in clinical research is that the continued discovery of genetic information will result in an increased understanding of patient risk. One well-studied genetic marker is the amplification of the Her-2 gene. Its amplification has been observed in breast, ovary, lung and prostate cancer. Research is currently underway to ivestigate the role Her-2 plays in controlling the signalling pathway of a tumor cell. One path in this research development is to empirically examine the relationship between Her2 gene amplification and the level of disease burden. If it is determined that Her2 gene amplification is an important component in tumor growth and its prevalence is high in the population, then their is a strong rationale for the development of therapy that targets this gene abnormality.

At Memorial Sloan-Kettering Cancer Center, only a small percentage of prostate cancer patients with localized disease are tested for Her2 gene amplification. Data on 83 patients with pathologically organ confined disease were tested for specific genetic markers, including Her2 status along with their PSA (Prostate Specific Antigen) values. Of the 83 localized prostate cancer patients who were tested for the Her2 gene amplification, 32 tested positive. Evidence of an association between Her2 amplification and high PSA values was determined through the Mann-Whitney U-statistic (p=). An additional 200 local disease patients with only their PSA values was then included into the analysis. Our objective was to use the

association between PSA and Her2 status to enhance our estimate of the prevalence of Her2 amplification in this prostate cancer population.

The proposed pairwise likelihood is derived conditional on (n_1, n_2, n_3), with information regarding λ stemming from $z_1, ..., z_{n_3} \sim H$. However in the current example, further information on the mixture parameter λ can be obtained through consideration of the sampling distribution of (n_1, n_2), the total number of subjects observed in the Her 2 amplified and nonamplified groups. Conditioning on $n^\dagger = n_1 + n_2$, n_1 has a binomial distribution with parameters $(n^\dagger, \lambda)$. Thus a simple estimate of λ based on the marginal totals (n_1, n_2) is $\hat{\lambda}_M = n_1/n^\dagger$. A straightforward strategy to provide a more informative estimate of the mixture parameter is to combine $\hat{\lambda}_M$ with the pairwise likelihood estimate $\hat{\lambda}_P$. The convex combination that minimizes the variance of the estimate is

$$\hat{\lambda} = \left(\frac{\hat{\sigma}_M^2}{\hat{\sigma}_M^2 + \hat{\sigma}_P^2}\right)\hat{\lambda}_P + \left(\frac{\hat{\sigma}_P^2}{\hat{\sigma}_M^2 + \hat{\sigma}_P^2}\right)\hat{\lambda}_M, \tag{4.1}$$

where

$$\hat{\sigma}_M^2 \equiv \hat{var}[\hat{\lambda}_M] = \frac{\hat{\lambda}_M(1-\hat{\lambda}_M)}{n^\dagger} \qquad \text{and} \qquad \hat{\sigma}_P^2 \equiv \hat{var}[\hat{\lambda}_P],$$

the estimate of the latter term is provided in (2.2).

Using the Lehmann two sample alternative in the pairwise rank based likelihood, $\hat{\theta} = x.xx(\text{se}\hat{\theta} = x.xx)$, and the estimated prevalence shows a marked decline $\hat{\lambda} = .314(\text{se}\hat{\lambda}_M = .049)$ from the binomial estimated based on the 83 tested patients. An examination of the data shows that the distribution of PSA for the mixture group was closer in shape to the nonamplified Her2 group PSA data, resulting in this downward adjustment in the probability of a Her2 amplification in this population. We are currently exploring explanations for this unusual sampling pattern. To summarize this analysis, with the addition of xx patients with PSA values only, we confirm the association between PSA and Her2 status, but the initial

finding that Her2 amplification is prevalent in this population has been called into question, requiring further investigation.

5. Discussions

By assuming a transformation model, we have used a pairwise rank likelihood approach for estimating the underlying parameters. The advantage of this approach is its simplicity. Alternatively, a triple-wise rank or all sample based rank approach can be employed. Implementation of these alternative approaches would involve additional complexity in the computation of the rank likelihood and its corresponding score and information statistics. Further research is required to determine if this increase in complexity is offset by an increase in efficiency of the resulting estimators.

We are currently exploring implementation of the pairwise likelihood to the case where outcome data is possibly right censored. This extension would enable us to examine the relationship, for example, between Her2 status and survival time in the population explored in Section 5. We are also persuing the development of a hypothesis testing framework, testing the hypothesis that the mixture parameter $\lambda = 0$.

Finally in Section 4, a convex combination of the pairwise likelihood estimate of the mixture parameter and an estimate based on the marginal group totals was established. The marginal estimate $\hat{\lambda}_M$ was developed external to the pairwise likelihood approach. We are currently exploring the development of a unified likelihood approach for estimation and inference on (θ, λ), derived by combining the conditional pairwise rank based likelihood with

the marginal binomial likelihood.

6. Appendix

In the appendix we present a proof of Theorem 1.

Denote

$$a_1 = \sum_{i,j} I(y_j > x_i), \qquad b_1 = \sum_{i,j} I(y_j \le x_i) = n_0 n_1 - a_1,$$
$$a_2 = \sum_{i,k} I(z_k > x_i), \qquad b_2 = \sum_{i,k} I(z_k \le x_i) = n_0 n_2 - a_2,$$
$$a_3 = \sum_{j,k} I(z_k > y_j), \qquad b_3 = \sum_{i,k} I(z_k \le y_j) = n_1 n_2 - a_3.$$

Then

$$l_P = \sum_{l=1}^{3} a_l \log p_l(\lambda, \theta) + b_l \log\{1 - p_l(\lambda, \theta)\}$$

and $\hat{\eta} = (\hat{\lambda}, \hat{\theta})$ satisfies

$$\frac{\partial l_P(\hat{\theta})}{\partial \eta} = \sum_{l=1}^{3} a_l \frac{\partial \log p_l(\lambda, \theta)}{\partial \eta} + b_l \frac{\partial \log\{1 - p_l(\lambda, \theta)\}}{\partial \eta} = 0.$$

By using Taylor expansion, we have

$$\sqrt{n}(\hat{\eta} - \eta) = \sqrt{n} \left(\frac{\partial^2 l_P(\eta)}{\partial \eta \eta^T} \right)^{-1} \frac{\partial l_P(\eta)}{\partial \eta} + o_p(1).$$

Note that in probability

$$\frac{1}{n_1 n_2} a_1 \to p_1, \qquad \frac{1}{n_1 n_2} b_1 \to 1 - p_1,$$
$$\frac{1}{n_1 n_3} a_2 \to p_2, \qquad \frac{1}{n_1 n_3} b_1 \to 1 - p_2,$$
$$\frac{1}{n_2 n_3} a_3 \to p_3, \qquad \frac{1}{n_2 n_3} b_3 \to 1 - p_3.$$

Also note $n = n_1 + n_2 + n_3$ and $n_i/n \to \nu_i, i = 1, 2, 3$. Then in probability

$$\begin{aligned}
\frac{1}{n^2}\frac{\partial^2 l_P(\eta)}{\partial\eta\partial\eta^T} &\to A(\eta) \\
&= \nu_1\nu_2 p_1 \frac{\partial^2 \log p_1(\eta)}{\partial\eta\partial\eta^T} + \nu_1\nu_2(1-p_1)\frac{\partial^2 \log(1-p_1(\eta))}{\partial\eta\partial\eta^T} \\
&+ \nu_1\nu_3 p_2 \frac{\partial^2 \log p_2(\eta)}{\partial\eta\partial\eta^T} + \nu_1\nu_3(1-p_2)\frac{\partial^2 \log(1-p_2(\eta))}{\partial\eta\partial\eta^T} \\
&+ \nu_2\nu_3 p_3 \frac{\partial^2 \log p_3(\eta)}{\partial\eta\partial\eta^T} + \nu_2\nu_3(1-p_3)\frac{\partial^2 \log(1-p_3(\eta))}{\partial\eta\partial\eta^T}. \qquad (6.1)
\end{aligned}$$

We can prove that

$$\frac{\partial l_P(\eta)}{\partial\eta} = d_1 \sum_{j,i}\{I(y_j > x_i) - p_1\} + d_2 \sum_{k,i}\{I(z_k > x_i) - p_2\} + d_3 \sum_{k,j}\{I(z_k > y_j) - p_3\},$$

where

$$d_l = \left[\frac{\partial \log p_l(\eta)}{\partial\eta} - \frac{\partial \log\{1-p_l(\eta)\}}{\partial\eta}\right],\ l = 1, 2, 3.$$

By using the results of Wilcox statistic and considering the correlation between terms, we can show that

$$\sqrt{n}\frac{1}{n^2}\frac{\partial l_P(\eta)}{\partial\eta} \to N(0, B)$$

in distribution, where the expression of $B = B(\eta)$ is extremly long involving twenty four terms. We refer this to a technical report by the authors. As a result

$$\sqrt{n}(\hat{\eta} - \eta) \to N(0, \Sigma),$$

where

$$\Sigma = A^{-1}(\eta)B(\eta)A^{-1}(\eta).$$

References

Anderson, J. A. (1979). Multivariate Logistic Compounds. *Biometrika.* **66** 17-26.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statist. Med.* **2**, 237-7.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc. B*, **34**, 187-220.

Hall, P. (1981). On the nonparametric estimation of mixture proportions. *J. R. Statist. Soc. B*, **43**, 147-56.

Hall, P. and Titterington, D. M. (1984). Efficient nonparametric estimation of mixture proportions. *J.Roy. Statist. Soc. Ser. B*, **46** 465-73.

Hosmer, D. W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three types of sample. *Biometrics* **29** 761-70.

Lehmann, E. L. (1953). The power of rank tests. *Ann. Math. Statist.* **24** 23-43

Murray, G. D. and Titterington, D. M. (1978). Estimation problems with data from a mixture. *Appl. Statist.* **27** 325-34.

Pettitt, A. N. (1984). Proportional odds model for survival data and estimates using ranks. *Appl. Statist.* **33**, 169-75.

Titterington, D. M. (1983). Minimum distance nonparametric estimation of mixture proportions. *J. R. Statist. Soc. B*, **45**, 37-46.

Table 1. Estimated mean values and observed empirical coverages, in parentheses, for nominal 95% confidence intervals based on 5000 simulations

λ	$\theta = 2$	$\theta = 4$	$\theta = 6$	$\theta = 8$
0.20	2.045 (0.929)	4.174 (0.942)	6.269 (0.943)	8.488 (0.932)
	0.181 (0.846)	0.194 (0.911)	0.196 (0.928)	0.198 (0.921)
0.35	2.039 (0.944)	4.148 (0.952)	6.294 (0.947)	8.476 (0.941)
	0.338 (0.906)	0.350 (0.934)	0.351 (0.925)	0.350 (0.937)
0.50	2.046 (0.952)	4.141 (0.954)	6.325 (0.951)	8.468 (0.947)
	0.503 (0.940)	0.499 (0.934)	0.501 (0.941)	0.503 (0.942)
0.65	2.041 (0.945)	4.142 (0.949)	6.298 (0.950)	8.477 (0.948)
	0.663 (0.944)	0.655 (0.941)	0.650 (0.933)	0.654 (0.936)
0.80	2.038 (0.940)	4.126 (0.945)	6.260 (0.943)	8.420 (0.939)
	0.824 (0.950)	0.806 (0.930)	0.805 (0.936)	0.804 (0.937)