AN EMPIRICAL LIKELIHOOD METHOD IN MIXTURE MODELS WITH INCOMPLETE CLASSIFICATIONS

Song Xi Chen and Jing Qin

Iowa State University and National Institute of Health

Abstract: In studying the relationship between a binary variable and a covariate, it is very common that the value of the binary variable is missing for some observations, and subsequently those observations are uncategorised. In this paper we show that the uncategorised data can be treated as auxiliary information, as in survey sampling literature. We establish a framework of parametric and nonparametric estimation by the empirical likelihood. The proposed empirical likelihood estimators improve the efficiency of estimators based on the categorised samples in the leading order. In a comparative study with the ratio estimator, we show the reveal robust performance of the empirical likelihood estimators. Applications to tax-auditing and to genetic studies are discussed.

Key words and phrases: Auxiliary information, empirical likelihood, genetic studies, mixture model, survey sampling, tax-auditing problem.

1. Introduction

Mixture models have been widely used in medical and psycho-social research and, more recently, in genetic studies, see for example, Ott (1999). Studies of finite mixture models date back to at least the late 1800s (Pearson (1893, 1895)). There is now a vast literature on mixture model inference, among them we refer readers to the books by Titterington, Smith and Makov (1985) and Lindsay (1995). Statistical analysis of mixture data is not trivial. In general, the maximum likelihood estimators do not admit closed forms and numerical methods, or an EM algorithm, are needed. Theoretical results are more difficult to attain as the mixture parameter may lie on the boundary of the parameter space. In addition, some nuisance parameters may be present under the null hypothesis, hence the null distribution of the likelihood ratio test statistic may be unknown even when the sample size is large.

Missing or incomplete data are a common phenomenon, see Little and Rubin (2002) for a comprehensive review. A missing data problem may be treated as a mixture problem. Consider the following situation: we are interested in studying the relationship between a binary variable (for example, disease status or gene mutation status) and an explanatory variable (for example, age or blood

pressure); for economic or ethical reasons, the binary variable is difficult to record. If none of the binary variables are known, then we have the standard mixture problem. On the other hand, if the binary variable is available only for part of the observations, then we have a type II mixture problem (Hosmer (1973)). Hosmer (1973) reported a study whose goal was to find the proportion of male halibut in each age class. The sex of halibut can be determined only by dissection of the fish. The International Halibut Commission has two sources of data: its own research cruises and commercial catches. Sex, age and length are available from fish taken on research cruises; only age and length can be obtained from commercial catches. Hosmer assumed normality of length distributions in each sex group in his estimation of the male proportion.

Suppose a population is composed of a mixture of two continuous sub-populations Π_1 and Π_2, with distribution functions F_1 and F_2, and in proportions π and $1 - \pi$, respectively. A sample of n observations is taken from the mixture and, for some $0 < \rho < 1$, $m = [n\rho]$ of these are selected at random and categorised correctly by further examination. Let $x_1, \ldots, x_{m_1}$ be data from F_1 and $y_1, \ldots, y_{m_2}$ be from F_2, respectively. Furthermore, denote the unclassified data as $z_1, \ldots, z_l$, where $l = n - m$. If we are interested in some characteristics of Π_1 and Π_2, for example underlying densities f_1 and f_2 or means μ_1 and μ_2, then the type II mixture problem can be treated as a problem with auxiliary information, where the explanatory variables with missing binary status contains the auxiliary information. Specifically, our interest in this type II mixture problem is motivated by the following two examples.

Example 1. A tax-auditing application.

Let w_i be the size of the ith transaction, with $\delta_i = 1$ if the transaction is taxable and 0 otherwise. The money amount w_i is known for n transactions of a company. The quantity $T = \sum_{i=1}^n \delta_i w_i$, the total taxable amount, is of interest. In practice, we can only monitor a portion of δ_i's by randomly selecting $m = [n\rho]$ transactions without replacement from the finite population $\{(\delta_1, w_1), \ldots, (\delta_n, w_n)\}$. If we assume the finite population comes from a super-population, then $\theta =: E(T/n) = E(\delta w) = \pi E(w|\delta = 1)$. Under a generalized linear model, Firth and Bennett (1998) considered robust model-based estimation of θ. Without postulating a parametric model on $P(\delta = 1|w)$, we construct more efficient estimators of the subpopulation mean $E(w|\delta = 1)$ and the mixture proportion π that lead to a more efficient estimator of θ, by using the auxiliary information.

Example 2. Genetic Applications.

Genetic studies explore the association between gene expression and adverse clinical outcomes. Although the use of genetic profiling holds great promise for

the future, currently this technology can only be applied to a small portion of cancer patients due to high cost and limited access to the tumor cells. Take prostate cancer as an example. It is known that there is an association between Her-2 over-expression and high Prostate Specific Antigen(PSA) values. Patients with the pathologically organ-confined disease were tested for specific genetic markers, including Her-2 status, along with their PSA values. Due to the high cost of genetic confirmation, Her-2 status is mostly unknown. However, PSA values are easily attainable. Clinically, it would be of much interest to find the mean PSA levels for Her-2 over-expressed and not over-expressed patients.

Motivated by ratio estimation in survey sampling (Cochran (1977)), Hall and Titterington (1985) gave a ratio-type adjustment to the standard kernel density estimators of f_i for the type II mixture problem. They showed that the ratio type density estimator has a smaller asymptotic variance than the density estimators based on the categorised samples. In a related study designed to improve the estimation of the mixture proportion π, Hall and Titterington (1984) constructed a sequence of multinomial approximations, and maximum likelihood estimators of the mixture proportions, by data binning. By allowing the number of bins to go to infinity, their estimator attains the Cramér-Rao lower bound asymptotically.

In this paper, we employ empirical likelihood (Owen (1988, 1990)) to improve estimation efficiency by utilising information contained in the uncategorised sample. Partial maximum empirical likelihood estimators of μ_i, $F_i(x)$ and $f_i(x)$, and π are proposed. Compared with estimators based on the categorised samples only, the proposed estimators improve the efficiency of estimation in the leading order. Our analysis shows that the empirical likelihood estimators of the densities $f_i(x)$ are as efficient as the ratio type estimators of Hall and Titterington (1984). The attraction of the empirical likelihood proposal is that it can be easily applied to estimation of other parameters, such as μ_i, $F_i(x)$ and π, within a single framework. For estimating μ_i, we show that the ratio estimator can encounter difficulties in capturing information contained in the uncategorised sample, since it may not be able to prevent the larger variance in the uncategorised sample from adversely affecting the quality of estimation. To make the ratio estimator work, a certain relationship between the mixture mean and variance has to exist. Indeed, due to the nature of kernel density estimation, this relationship is present for the density estimation considered in Hall and Titterington (1985). Our empirical likelihood estimator for π, although it may not attain the Cramér-Rao lower bound asymptotically, is easily obtained as a by-product of estimating other parameters. Simulations show that it has comparable finite sample performance with the Hall-Titterington estimator.

This paper is organised as follows. Section 2 proposes the empirical likelihood estimators for the means and the mixture proportion. Efficiency of these

estimators is evaluated in Section 3 in a comparative study with the ratio estima-
tor. Section 4 contains extentions to distribution and density function estimation.
Section 5 reports results from simulation studies. All the proofs are given in the
appendix.

2. Empirical Likelihood Based Estimation

Let $\mu = \pi\mu_1 + (1-\pi)\mu_2$ be the mean of the mixture. We start with the
estimation of μ_i. Extentions to estimation of $F_i(x)$ and $f_i(x)$ will be treated in
Section 4.

The empirical likelihood for (μ_1, μ_2, π) based on the categorised data is

$$L(\mu_1, \mu_2, \pi) = \max_{p_i, q_j} \pi^{m_1}(1-\pi)^{m_2} \prod_{i=1}^{m_1} p_i \prod_{j=1}^{m_2} q_j \tag{1}$$

subject to the constraints $\sum_{i=1}^{m_1} p_i = 1$, $\sum_{j=1}^{m_2} q_j = 1$, $\sum_{i=1}^{m_1} p_i(x_i - \mu_1) = 0$
and $\sum_{j=1}^{m_2} q_j(y_j - \mu_2) = 0$. Standard derivations in empirical likelihood (Owen
(1990)) show that the optimal p_i and q_j are $p_i = m_1^{-1}\{1 + \lambda_1(x_i - \mu_1)\}^{-1}$ and
$q_j = m_2^{-1}\{1 + \lambda_2(y_j - \mu_2)\}$, where λ_i's are Lagrange multipliers satisfying

$$\sum_{i=1}^{m_1} \frac{x_i - \mu_1}{1 + \lambda_1(x_i - \mu_1)} = 0 \quad \text{and} \quad \sum_{j=1}^{m_2} \frac{y_j - \mu_2}{1 + \lambda_2(y_j - \mu_2)} = 0. \tag{2}$$

The log empirical likelihood ratio is then

$$\ell(\mu_1, \mu_2, \pi) = m_1 \log \pi + m_2 \log(1-\pi) - \sum \log\{1 + \lambda_1(x_i - \mu_1)\}$$
$$- \sum \log\{1 + \lambda_2(y_j - \mu_2)\}. \tag{3}$$

To use the information contained in the uncategorised sample, we maximise
$\ell(\mu_1, \mu_2, \pi)$ subject to

$$\pi\mu_1 + (1-\pi)\mu_2 = \bar{W}, \tag{4}$$

where $\bar{W}$ is the grand mean based on all data. Substituting $\mu_2^* = \mu_2(\mu_1, \pi) = (\bar{W} - \pi\mu_1)/(1-\pi)$ in (3),

$$\ell(\mu_1, \mu_2^*\pi) = m_1 \log \pi + m_2 \log(1-\pi) - \sum \log\{1 + \lambda_1(x_i - \mu_1)\}$$
$$- \sum \log\{1 + \lambda_2\{y_j - \mu_2(\mu_1, \pi)\}\}.$$

Differentiating with respect to μ_1, π, we have

$$\frac{\partial \ell}{\partial \mu_1} = \sum_{i=1}^{m_1} \frac{\lambda_1}{1 + \lambda_1(x_i - \mu_1)} - \sum_{j=1}^{m_2} \frac{\frac{\lambda_2 \pi}{1-\pi}}{1 + \lambda_2\{y_j - \mu_2(\mu_1, \pi)\}} = 0 \quad \text{and}$$

$$\frac{\partial \ell}{\partial \pi} = \sum_{j=1}^{m_2} \frac{\frac{\lambda_2(\bar{W}-\mu_1)}{(1-\pi)^2}}{1 + \lambda_2\{y_j - \mu_2(\mu_1, \pi)\}} + \frac{m_1}{\pi} - \frac{m_2}{1-\pi} = 0.$$

These lead to

$$\lambda_1 m_1 - \frac{m_2 \lambda_2 \pi}{1-\pi} = 0 \quad \text{and} \quad \frac{m_1}{\pi} - \frac{m_2}{1-\pi} + m_2 \lambda_2 \frac{\bar{W}-\mu_1}{(1-\pi)^2} = 0. \tag{5}$$

It can be shown that

$$\hat{\pi} = \tilde{\pi}\{1 + \lambda_1(\bar{W} - \mu_1)\}, \tag{6}$$

where $\tilde{\pi} = m_1/(m_1 + m_2)$ is the maximum likelihood estimator of π based on the categorised samples only. We denote the solutions of the above equations as $(\hat{\mu}_1, \hat{\mu}_2, \hat{\pi})$, the empirical likelihood estimators of μ_1, μ_2 and π, respectively.

The empirical likelihood ratio $\ell(\mu_1, \mu_2, \pi)$ can be written as $\ell_1(\pi) + \ell_2(\mu_1, \mu_2)$ where $\ell_1(\pi) = m_1 \log(\pi) + m_2 \log(1 - \pi)$. Clearly ℓ_1 is concave. Let $H = (x_{(1)}, x_{(m_1)}) \times (y_{(1)}, y_{(m_2)})$ where $x_{(1)}$ and $x_{(m_1)}$ are the smallest and largest x-sample values, and $y_{(1)}$ and $y_{(m_2)}$ are the smallest and largest y-sample values, respectively. As $\ell_2(\mu_1, \mu_2)$ is concave (Hall and La Scala (1990)) on H, $\ell(\mu_1, \mu_2, \pi)$ is concave on $H \times (0, 1)$. Let $D = \{w = \pi\mu_1 + (1-\pi)\mu_2 \mid (\mu_1, \mu_2, \pi) \in H \times (0, 1)\} = (\min\{x_{(1)}, y_{(1)}\}, \max\{x_{(m_1)}, y_{(m_2)}\})$. The constraint maximising $\ell(\mu_1, \mu_2, \pi)$ always admits a unique solution provided $\bar{W} \in D$. In a finite sample, it may happen that $\bar{W}$ is outside D, and then the above maximisation of $\ell(\mu_1, \mu_2, \pi)$ does not have a closed form. However, as n gets large, the probability of this happening goes to zero.

It should be noted that the above formulation is not a full empirical likelihood as the likelihood is constructed based on the categorized data only. This is based on considerations of computational and theoretical tractability. More discussion on this point is provided in Section 6.

3. Efficiency of Empirical Likelihood Estimators

Let $\sigma_1^2 = E(X - \mu_1)^2$, $\sigma_2^2 = E(Y - \mu_2)^2$, so $\sigma^2 = \pi\sigma_1^2 + (1 - \pi)\sigma_2^2 + \pi(\mu - \mu_1)^2/(1 - \pi)$ is the variance of the mixture population. Moreover, let $T_0 = m^{-1}(m_1\bar{X} + m_2\bar{Y}) - \bar{W}$, which is $O_p(n^{-1/2})$. Derivations given in the Appendix show that

$$\hat{\mu}_1 = \bar{X} - \sigma^{-2}\sigma_1^2 T_0 + O_p(n^{-1}), \tag{7}$$
$$\hat{\mu}_2 = \bar{Y} - \sigma^{-2}\sigma_2^2 T_0 + O_p(n^{-1}) \quad \text{and} \tag{8}$$
$$\hat{\pi} = \tilde{\pi}\{1 + \sigma^{-2}T_0(\mu - \mu_1)\} + O_p(n^{-1}). \tag{9}$$

Hence, the empirical likelihood utilises the auxiliary information by adding correction terms of $O_p(n^{-1/2})$ to the estimators $\bar{X}, \bar{Y}$ and $\tilde{\pi}$ based solely on the categorised samples.

Derivations given in the appendix show that

$$\text{Var}\,(\hat{\mu}_1) = \text{Var}\,(\bar{X}) - n^{-1}(1-\rho)\frac{\sigma_1^4}{\sigma^2\rho} + O(n^{-2}), \tag{10}$$

$$\text{Var}\,(\hat{\mu}_2) = \text{Var}\,(\bar{Y}) - n^{-1}(1-\rho)\frac{\sigma_2^4}{\sigma^2\rho} + O(n^{-2}). \tag{11}$$

From (10) and (11), we see a reduction of variance from utilising information contained in the uncategorised sample. In particular, the smaller the ρ is, the larger the variance reduction. The relative variance reduction also increases as ρ decreases.

In survey sampling, a widely used adjustment to a "standard" estimator is the ratio estimator. It is well known that ratio estimation performs well if the explanatory variable and the response are highly correlated. We now investigate its performance in the case of mean estimation.

Let $\tilde{\mu} = (m_1\bar{X} + m_2\bar{Y})/(m_1 + m_2)$ be the mean of the categorised samples. A ratio type adjustment to $\bar{X}$ is $\hat{\mu}_{1r} = \bar{X}\bar{W}/\tilde{\mu}$. It can be shown that

$$\hat{\mu}_{1r} = \bar{X}\Big\{1 - (1-\rho)(\frac{m}{m_2})^2\frac{T_0}{\mu}\Big\} + O_p(n^{-1}). \tag{12}$$

The same technique that yields (10) and (11) gives, by ignoring terms of $o(n^{-1})$,

$$\text{Var}\,(\hat{\mu}_{1,r}) = \text{Var}\,(\bar{X})\Big\{1 - \frac{(1-\rho)\mu_1\pi}{\mu}\Big\}^2 + n^{-1}(1-\rho)^2\mu_1^2\frac{\frac{\sigma^2}{1-\rho} + \frac{(1-\pi)\sigma_2^2}{\rho}}{\mu^2}.$$

This indicates that the ratio adjustment does not necessarily lead to a variance reduction, and a variance inflation is possible when μ is close to zero. Indeed, the larger variance in the uncategorised sample can penetrate into the estimator and make things worse. A key component of the ratio adjustment is T_0/μ, as revealed at (12). Equation (A.7) in the Appendix shows that $\text{Var}\,(T_0) = n^{-1}(1-\rho)\sigma^2/\rho + O(n^{-2})$, which means $\text{Var}\,(T_0/\mu) \sim n^{-1}\sigma^2/\mu^2$. To control the variance of this correction term, we either need a relationship between μ and σ^2 in order to cancel out the μ appeared in the denominator or μ has to be away from zero, and this may not be satisfied in general. However, the relationship does exists in the case of the nonparametric density estimation considered in Hall and Titterington (1985), due to the nature of kernel density estimation. A underlying cause for the problem with the ratio estimator, suggested by a referee, is that it is not invariant under the location shift transformation.

In contrast, the correction term in the empirical likelihood adjustment is T_0/σ^2 in (7) and (8). Now $\text{Var}\,(T_0/\sigma^2) \sim n^{-1}\sigma^{-2}$, which indicates a more comfortable situation than the ratio adjustment, as σ^2 is automatically away from zero.

The empirical likelihood estimation of the mixture proportion π is more efficient too. Indeed, based on derivations shown in the appendix,

$$\text{Var}\,(\hat{\pi}) = \text{Var}\,(\tilde{\pi})\left[1 - \frac{(\mu - \mu_1)^2\pi(1-\rho)}{(1-\pi)\sigma^2}\right] + O(n^{-2}). \qquad (13)$$

As $(\mu - \mu_1)^2\pi(1-\rho)/\{(1-\pi)\sigma^2\} \in [0,1)$, a variance reduction is registered for $\hat{\pi}$.

The empirical likelihood estimator for π is "parametric" in the sense that it is a by-product of the inference for the means. This is shown in (13) where the variance reduction depends on $(\mu - \mu_1)^2/\sigma^2$. Hall and Titterington (1984) proposed an nonparametric estimator, denoted as $\hat{\pi}_{HT}$, by binning the domain of the distribution and maximizing a multinomial likelihood. The asymptotic variance of their estimator is

$$\text{Var}\,(\tilde{\pi})\{1 - \pi(1-\rho)(1-\pi)^{-1}\}\left\{\int \frac{f_1^2}{f} - 1\right\},$$

which converges to the Cramér-Rao lower bound as the number of bins goes to infinity. A simulation study reported in Section 5 reveals that our estimator $\hat{\pi}$ has similar performance.

4. Extensions

The theory developed in the previous section can be readily extended to other parameters of Π_i. In this section, we consider the tax-auditing problem and the nonparametric estimation of the distribution and density functions of Π_i respectively.

4.1. Application to the tax-auditing example

As outlined in the introduction, the parameter of interest is $\theta = \pi\mu_1$, where $\mu_1 = E(w|\delta = 1)$. An empirical likelihood estimator of θ is $\hat{\theta} = \hat{\pi}\hat{\mu}_1$ by plugging-in $\hat{\pi}$ and $\hat{\mu}_1$. An estimator that only uses the categorised sample is $\tilde{\theta} = \tilde{\pi}\bar{X}$.

Since $\hat{\theta} = \pi\mu_1 + \mu_1(\hat{\pi} - \pi) + \pi(\hat{\mu}_1 - \mu_1) + (\hat{\pi} - \pi)(\hat{\mu}_1 - \mu_1)$, it is easy to see that $E(\hat{\pi}\hat{\mu}_1) = \pi\mu_1 + O(n^{-1})$, implying that the bias is negligible. The variance of $\hat{\theta}$ is

$$\text{Var}\,(\hat{\theta}) = \mu_1^2\text{Var}\,(\hat{\pi}) + \pi^2\text{Var}\,(\hat{\mu}_1) + 2\mu_1\pi\,\text{Cov}\,(\hat{\pi}, \hat{\mu}_1) + o(n^{-1}). \qquad (14)$$

From (7) and (9),

$$\begin{aligned}
\text{Cov}\,(\hat{\pi}, \hat{\mu}_1) &= (\mu - \mu_1)\pi\sigma^{-2}\,\text{Cov}\,(T_0, \bar{X}) - (\mu - \mu_1)\pi\sigma_1^2\sigma^{-4}\text{Var}\,(T_0) + o(n^{-1}) \\
&= (\mu - \mu_1)\pi(1-\rho)\sigma_1^2\sigma^{-2} - (\mu - \mu_1)\pi(1-\rho)\sigma_1^2\sigma^{-2} + o(n^{-1}) = o(n^{-1}),
\end{aligned}$$

which mirrors the fact that $\mathrm{Cov}\,(\tilde{\pi}, \bar{X}) = 0$. This together with (10) and (13) leads to

$$\mathrm{Var}\,(\hat{\theta}) = \mu_1^2 \mathrm{Var}\,(\hat{\pi}) + \pi^2 \mathrm{Var}\,(\hat{\mu}_1) + o(n^{-1}).$$

Thus $\hat{\theta}$ has a smaller variance than $\tilde{\theta}$, as both $\hat{\pi}$ and $\hat{\mu}$ have smaller variances than those of $\tilde{\pi}$ and $\bar{X}$, respectively.

4.2. Distribution function estimation

Let F_i be the distribution functions of Π_i, $i = 1, 2$, and $F = \pi F_1 + (1 - \pi)F_2$ be the mixture distribution function. Conventional estimators of F_i and F at a fixed x are the empirical distribution functions:

$$F_{1,m_1}(x) = m_1^{-1} \sum_{i=1}^{m_1} I(X_i \leq x), \quad F_{2,m_2}(x) = m_2^{-1} \sum_{j=1}^{m_2} I(Y_j \leq x) \quad \text{and}$$

$$F_n(x) = n^{-1}\Big\{\sum_{i=1}^{m_1} I(X_i \leq x) + \sum_{j=1}^{m_2} I(Y_j \leq x) + \sum_{k=1}^{l} I(Z_k \leq x)\Big\},$$

where $I(\cdot)$ is the indicator function.

Since $F_1(x)$, $F_2(x)$ and $F(x)$ are the means of $I(X_i \leq x)$, $I(Y_j \leq x)$ and $I(Z_k \leq x)$, respectively, the scheme established in Section 2 is applicable for setting up $\ell\{F_1(x), F_2(x), \pi\}$, the empirical likelihood for $(F_1(x), F_2(x), \pi)$. Let $\hat{F}_i(x)$ be the estimators of $F_i(x)$ that minimize $\ell\{F_1(x), F_2(x), \pi\}$ subject to $\pi F_1(x) + (1 - \pi)F_2(x) = F_n(x)$. Then from the results developed in Section 3, we have, after ignoring terms of $O(n^{-2})$, for $i = 1$ and 2,

$$\mathrm{Var}\,\{\hat{F}_i(x)\} = \mathrm{Var}\,\{F_{i,m_i}(x)\} - n^{-1}(1 - \rho)F_i^2(x)\frac{\{1 - F_i(x)\}^2}{\rho\sigma^2(x)},$$

where $\sigma^2(x) = \pi F_1(x)\{1 - F_1(x)\} + (1 - \pi)F_2(x)\{1 - F_2(x)\} + \pi(1 - \pi)\{F_1(x) - F_2(x)\}^2$. This provides a variance reduction of order n^{-1} from the conventional estimators $F_{i,m_i}(x)$.

4.3. Density estimation

Let f_i be the probability density functions of Π_i $i = 1, 2$, and $f = \pi_1 f_1 + (1 - \pi_1)f_2$ be the density of the mixture. Let $K(\cdot)$ be a kernel function and take $K_h(u) = h^{-1}K(u/h)$ where h is a smoothing bandwidth such that $h \to 0$ as $n \to \infty$. Standard kernel density estimator of f_i and f based on the categorised samples and the entire sample are, respectively,

$$\tilde{f}_1(x) = m_1^{-1} \sum_{i=1}^{m_1} K_h(x - X_i) \quad \text{and} \quad \tilde{f}_2(x) = m_2^{-1} \sum_{j=1}^{m_2} K_h(x - Y_j),$$

$$\hat{f}^\star(x) = n^{-1}\{\sum_{i=1}^{m_1} K_h(x - X_i) + \sum_{j=1}^{m_2} K_h(x - Y_j) + \sum_{k=1}^{l} K_h(Z_k - x)\}.$$

The same bandwidth h is used in the above formulation of density estimators, which is designed to reduce the variation of $\tilde{f}_1(x)$ or $\tilde{f}_2(x)$ one at a time. For instance, if the aim is to reduce the variance of $\tilde{f}_1(x)$, h should be chosen based on the first categorised sample with the aim of minimising the error of estimation of $f_1(x)$. Methods for choosing h are discussed in Silverman (1986) and Wand and Jones (1995).

Let $\mu_1(x) = E\{K_h(x - X_i)\}$ and $\mu_2(x) = E\{K_h(x - Y_j)\}$. Standard results in kernel density estimation show that $\mu_i(x) = f_i(x) + O(h^2)$ for $i = 1$ and 2. So, up to an error of $O(h^2)$, $f_1(x)$ and $f_2(x)$ can be regarded as the mean of $K_h(x - X_i)$ and $K_h(x - Y_j)$, respectively. Hence $\ell\{f_1(x), f_2(x), \pi\}$, the empirical likelihood for $(f_1(x), f_2(x), \pi)$ based on the categorised samples, can be constructed as for μ_i in (3). Let $(\hat{f}_1(x), \hat{f}_2(x), \hat{\pi})$ be the maxima of $\ell\{f_1(x), f_2(x), \pi\}$ subject to

$$\pi f_1(x) + (1 - \pi)f_2(x) = \hat{f}^\star(x). \tag{15}$$

Then, the results given in Section 3 are valid. In particular, let $R(K) = \int K^2(u)\,du$. From results on kernel density estimation we have $\sigma_i^2 = h^{-1}R(K)f_i(x)\{1 + O(h)\}$ for $i = 1$ and 2, and

$$\sigma^2 = \pi\sigma_1^2 + (1-\pi)\sigma_2^2 + \pi(1-\pi)^{-1}\{f(x) - f_1(x)\}^2 = h^{-1}R(K)f(x)\{1 + o(1)\}, \tag{16}$$

which indicates that σ^2 is proportional to the mixture mean $\mu = f(x)$. Moreover, we have the following analogies of (10) and (11):

$$\text{Var}\,\{\hat{f}_i(x)\} = \text{Var}\,\{\tilde{f}_i(x)\} - (nh)^{-1}\frac{(1 - \rho)f_i^2(x)}{\rho f(x)} + o\{(nh)^{-1}\}, \quad i = 1, 2.$$

The amount of variance reduction is $(nh)^{-1}(\rho^{-1} - 1)R(K)f_i^2(x)/f(x)$, which is the same as that achieved by the Hall and Titterington (1985) ratio estimator. The reason for the success of the ratio adjustment here is the relationship between σ^2 and μ as revealed in (16). Although the proposed empirical likelihood estimators for $f_i(x)$ does not improve the Hall-Titterington estimator, it is a bona fide density, which is not the case for the ratio estimator.

5. Simulation Results

In this section, we present simulation results designed to evaluate the performance of the empirical likelihood estimators. First we consider estimation of the mixture proportion π. The empirical likelihood estimator $\hat{\pi}$ was compared with the naive estimator $\tilde{\pi} = m_1/(m_1 + m_2)$ and $\hat{\pi}_{HT}$. We used both eight and

four bins when calculating $\hat{\pi}_{HT}$. Since four and eight bins gave similar results, only the results for eight bins are reported.

For estimation of the means μ_i, the empirical likelihood estimators $\hat{\mu}_i$ were compared with $\bar{X}$ or $\bar{Y}$, the sample means based on the categorised data, and the ratio estimators $\hat{\mu}_{1r}$ or $\hat{\mu}_{2r}$.

For $n = 400$ and $m = 50$ or 100, two simulation models were examined: a normal mixture $\pi N(0, 1) + (1 - \pi)N(\mu, 1)$, and an exponential mixture $\pi \exp(1) + (1 - \pi) \exp(1/\mu)$ for different combination of π and μ. As the simulation results for $m = 50$ and $m = 100$ share the same pattern, we report only those for $m = 50$. Tables 1 to 4 summarise the means and variances of the estimators considered in this paper based on 1,000 replications.

Table 1. Mean (variance) of different estimators discussed in Section 5 based on 1,000 simulations, $n = 400$, $m = 50$.

Normal mixture model $\pi N(0, 1) + (1 - \pi)N(\mu, 1)$.

Estimators	$\pi = 0.2$, $\mu = 1.0$		$\pi = 0.5$, $mu = 1.0$		$\pi = 0.8$, $\mu = 1.0$	
$\tilde{\pi}$	0.20070	(0.00338)	0.50434	(0.00477)	0.80148	(0.00334)
$\hat{\pi}_{HT}$	0.20090	(0.00345)	0.50370	(0.00474)	0.80012	(0.00337)
$\hat{\pi}$	0.20057	(0.00301)	0.50373	(0.00399)	0.79983	(0.00304)
$\bar{X}$	0.00466	(0.10225)	-0.00265	(0.03856)	-0.01141	(0.02333)
$\bar{Y}$	0.99822	(0.025013)	1.00155	(0.03747)	0.97646	(0.11510)
$\hat{\mu}_{1r}$	-0.01864	(0.11414)	-0.05398	(0.08377)	-0.16230	(23.68682)
$\hat{\mu}_{2r}$	1.01105	(0.01506)	1.08614	(0.13990)	1.93026	(882.06761)
$\hat{\mu}_1$	-0.00139	(0.09234)	-0.00554	(0.02622)	0.00096	(0.01024)
$\hat{\mu}_2$	1.00190	(0.01020)	1.01017	(0.02526)	0.99409	(0.10116)

Table 2. Mean (variance) of different estimators discussed in Section 5 based on 1,000 simulations, $n = 400, m = 50$.

Normal mixture model $\pi N(0, 1) + (1 - \pi)N(\mu, 1)$.

Estimators	$\pi = 0.2$, $\mu = 2.0$		$\pi = 0.5$, $\mu = 2.0$		$\pi = 0.8$, $\mu = 2.0$	
$\tilde{\pi}$	0.19890	(0.00315)	0.50278	(0.00511)	0.79790	(0.00303)
$\hat{\pi}_{HT}$	0.20020	(0.00207)	0.50251	(0.00300)	0.79946	(0.00219)
$\hat{\pi}$	0.20018	(0.00200)	0.50286	(0.00282)	0.80005	(0.00203)
$\bar{X}$	-0.00330	(0.09825)	0.00827	(0.04334)	0.00506	(0.02719)
$\bar{Y}$	2.00354	(0.02704)	2.00489	(0.04111)	1.99994	(0.10673)
$\hat{\mu}_{1r}$	-0.01386	(0.09848)	-0.01379	(0.05090)	-0.05068	(0.69871)
$\hat{\mu}_{2r}$	2.00954	(0.02739)	2.06969	(0.16605)	2.28770	(27.04612)
$\hat{\mu}_1$	-0.01949	(0.09220)	-0.00150	(0.03350)	-0.00412	(0.01484)
$\hat{\mu}_2$	2.00538	(0.01483)	2.01388	(0.03320)	2.00598	(0.09576)

Table 3. Mean (variance) of different estimators discussed in Section 5 based on 1,000 simulations, $n = 400, m = 50$.

Exponential mixture $\pi \exp(1) + (1 - \pi) \exp(1/\mu)$.

Estimators	$\pi = 0.2$, $\mu = 1.5$	$\pi = 0.5$, $\mu = 1.5$	$\pi = 0.8$, $\mu = 1.5$
$\tilde{\pi}$	0.20072 (0.00316)	0.50018 (0.00504)	0.80112 (0.00326)
$\hat{\pi}_{HT}$	0.20097 (0.00369)	0.49948 (0.00578)	0.80079 (0.00396)
$\hat{\pi}$	0.20086 (0.00320)	0.49988 (0.00489)	0.80092 (0.00336)
$\bar{X}$	0.99747 (0.10553)	1.00160 (0.04090)	0.99704 (0.02759)
$\bar{Y}$	1.49971 (0.06159)	1.48541 (0.09091)	1.51225 (0.25036)
$\hat{\mu}_{1r}$	1.00731 (0.10081)	1.00960 (0.03207)	0.99951 (0.01152)
$\hat{\mu}_{2r}$	1.50181 (0.01344)	1.48689 (0.04110)	1.50477 (0.17712)
$\hat{\mu}_1$	1.00254 (0.10598)	1.00975 (0.03441)	0.99880 (0.01271)
$\hat{\mu}_2$	1.50316 (0.01414)	1.48722 (0.04416)	1.50258 (0.19169)

Table 4. Mean (variance) of different estimators discussed in Section 5 based on 1,000 simulations, $n = 400, m = 50$.

Exponential mixture $\pi \exp(1) + (1 - \pi) \exp(1/\mu)$.

Estimators	$\pi = 0.2$, $\mu = 4.0$	$\pi = 0.5$, $\mu = 4.0$	$\pi = 0.8$, $\mu = 4.0$
$\tilde{\pi}$	0.20050 (0.00316)	0.49846 (0.00492)	0.80014 (0.00313)
$\hat{\pi}_{HT}$	0.20133 (0.00327)	0.49836 (0.00411)	0.79868 (0.00274)
$\hat{\pi}$	0.20000 (0.00293)	0.49860 (0.00399)	0.79540 (0.00255)
$\bar{X}$	1.00452 (0.11028)	1.00917 (0.04449)	1.00243 (0.025179)
$\bar{Y}$	4.02575 (0.41346)	4.06645 (0.69684)	3.95981 (1.68587)
$\hat{\mu}_{1r}$	1.01754 (0.12342)	1.02014 (0.06266)	1.03437 (0.04107)
$\hat{\mu}_{2r}$	4.01697 (0.09403)	4.01916 (0.23856)	3.95310 (0.93674)
$\hat{\mu}_1$	1.00184 (0.10617)	1.00688 (0.04349)	1.01770 (0.02569)
$\hat{\mu}_2$	4.02037 (0.09295)	4.04098 (0.24162)	3.99416 (0.97172)

For estimation of π, $\hat{\pi}$ and $\hat{\pi}_{HT}$ had overall similar performance and both had smaller variance than the naive estimator $\tilde{\pi}$, as predicted by the theory. When the two populations were well separated from each other, $\hat{\pi}_{HT}$ was slightly better than $\hat{\pi}$. On the other hand, if the two populations were close, the proposed empirical likelihood estimator was slightly better.

For the estimation of means, the empirical likelihood estimator was the clear winner among the three estimators. Most evidently, the ratio estimators performed worse than $\bar{X}$ or $\bar{Y}$ in some cases. It is significant that the variance of $\hat{\mu}_{2r}$ is out of control in the normal mixture model when $\pi = 0.8$. The situation is a little better when m is increased to 100 and n is still 400.

6. Discussions

In this paper, empirical likelihood has been used to improve estimation efficiency in a type II mixture model. The proposed empirical likelihood formulation is effectively a partial likelihood based on the categorised sample. The information in the uncategorized sample is utilised via a constraint involving the grand mean of the entire sample. Two referees have pointed out an alternative full likelihood formulation which adds an extra component $\Pi_{l=1}^{n-m}\{\pi p_i + (1-\pi)q_i\}$ to (1) after indexing p_i from 1 to $m_1 + l$ and q_j from 1 to $m_2 + l$, respectively. This would lead to more efficient estimators. Considerations behind our choice are easier computations and tractable theoretical analyses.

There is a potential bias for the sampling methods discussed in this paper if an unequal weight sampling strategy is used. This would likely be the case for the tax-auditing problem, since large transactions are more likely to be monitored. We are exploring empirical likelihood methods to adjust for this sampling bias.

Acknowledgements

Appendix. Technical Details

Derivations of (7), (8) and (9). Let $\eta = m_1^{-1} \sum_{i=1}^{m_1}(X_i - \mu_1)^2$ and $\xi = m_2^{-1} \sum_{i=1}^{m_2}(Y_i - \mu_2)^2$. A standard proof in empirical likelihood, for instance that given in Owen (1990), leads to $\lambda_i = O_p(n^{-1/2})$ for $i = 1$ and 2. Inverting (2),

$$\lambda_1 = (\bar{X} - \hat{\mu}_1)\eta^{-1} + O_p(n^{-1}) \quad \text{and} \quad \lambda_2 = (\bar{Y} - \hat{\mu}_2)\xi^{-1} + O_p(n^{-1}). \tag{A.1}$$

From (6), $(1-\hat{\pi})/\hat{\pi} = m_1^{-1}m_2\{1 - \lambda_1(\bar{W} - \hat{\mu}_1)(m_1 + m_2)/m_2\} + O_p(n^{-1})$. This and (5) imply

$$\lambda_2 = \lambda_1 + O_p(n^{-1}). \tag{A.2}$$

Let $T = \bar{Y} + m_1 m_2^{-1}\bar{X} - (m_1 m_2^{-1} + 1)\bar{W}$. Clearly, $T_0 = m_2 m^{-1}T$. From (A.1) and the fact that $\hat{\mu}_2 = (\bar{W} - \hat{\pi}\hat{\mu}_1)/(1 - \hat{\pi})$,

$$\begin{aligned}
\bar{Y} - \hat{\mu}_2 &= \{\bar{Y} + m_1 m_2^{-1}\hat{\mu}_1 - (m_1 m_2^{-1} + 1)\bar{W} + m_1 m_2^{-1}\lambda_1(\bar{W} - \hat{\mu}_1)(\hat{\mu}_1 - \bar{Y})\} \\
&\quad \times \{1 - m_1 m_2^{-1}\lambda_1(\bar{W} - \hat{\mu}_1)\}^{-1} \\
&= T - m_1 m_2^{-1}\{1 + (m_1 m_2^{-1} + 1)(\bar{W} - \bar{X})^2\eta^{-1}\}(\bar{X} - \hat{\mu}_1) + O_p(n^{-\frac{3}{2}}). \tag{A.3}
\end{aligned}$$

From (A.1), (A.2) and (A.3), by ignoring terms of $O_p(n^{-1})$,

$$(\bar{X} - \hat{\mu}_1)\eta^{-1} = \left[T - m_1 m_2^{-1}\{1 + (m_1 m_2^{-1} + 1)(\bar{W} - \hat{\mu}_1)^2\eta^{-1}\}(\bar{X} - \hat{\mu}_1)\right]\xi^{-1}. \tag{A.4}$$

Let $\hat{\alpha} = \xi + m_1 m_2^{-1}\eta + m_1 m_2^{-1}(m_1 m_2^{-1} + 1)(\bar{W} - \bar{X})^2$. Then (A.4) implies

$$\bar{X} - \hat{\mu}_1 = \hat{\alpha}^{-1}\eta T + O_p(n^{-1}) = \sigma^{-1}\sigma_1^2 T_0 + O_p(n^{-1}),$$

where $\alpha = \sigma_2^2 + \pi\sigma_1^2/(1-\pi) + \pi(\mu-\mu_1)^2/(1-\pi)^2$ produces is the leading order term of $E(\hat{\alpha})$. This, and the fact that $\alpha = \sigma^2/(1-\pi)$, produces (7). A similar exercise leads to (8). From (A.1), $\lambda_1 = \alpha^{-1}T + O_P(n^{-1})$. This and (6) imply (9).

Derivations of (10) and (11). We first need to derive $\text{Var}(T)$. Let $\bar{Z}$ be the average of the uncategorised sample and $T = n^{-1}\{m_1 m_2^{-1}\bar{X} + \bar{Y} - mm_2^{-1}\bar{Z}\}$. Then

$$E(T|m_1, m_2) = n^{-1}(m_1 m_2^{-1} + 1)(\tilde{\pi} - \pi)(\mu_1 - \mu_2) + O_p(n^{-1}), \tag{A.5}$$

$$\text{Var}(T|m_1, m_2) = \frac{l}{nm_2}\{\alpha - (1 - \frac{l}{n})m_1 m_2^{-1}(m_1 m_2^{-1} + 1)(\mu - \mu_1)^2\} + O_p(n^{-2}). \tag{A.6}$$

Then $E(T) = O(n^{-1})$. Since $\mu - \mu_1 = -(1-\pi)(\mu_1 - \mu_2)$,

$$\begin{aligned}
\text{Var}(T) &= \text{Var}\{E(T|m_1, m_2)\} + E\{\text{Var}(T|m_1, m_2)\} \\
&= \frac{(1-\rho)^2(\mu_1 - \mu_2)^2\pi}{(1-\pi)\rho n} + \frac{(1-\rho)\alpha}{(1-\pi)\rho n} - \frac{(1-\rho)^2\pi(\mu-\mu_1)^2}{(1-\pi)^3\rho n} + O(n^{-2}) \\
&= n^{-1}\frac{(1-\rho)\sigma^2}{(1-\pi)^2\rho} + O(n^{-2}). \tag{A.7}
\end{aligned}$$

A by-product of this is that $T = O_p(n^{-1/2})$ as claimed.

The variance of $\hat{\mu}_1$ is $\text{Var}(\hat{\mu}_1) = E\{\text{Var}(\hat{\mu}_1|m_1, m_2)\} + \text{Var}\{E(\hat{\mu}_1|m_1, m_2)\}$. From (7) and (A.5)

$$E(\hat{\mu}_1|m_1, m_2) = \mu_1 - n^{-1}\sigma_1^2\alpha^{-1}l(m_1 m_2^{-1} + 1)(\tilde{\pi} - \pi)(\mu_1 - \mu_2) + O(n^{-2}). \tag{A.8}$$

Thus, by ignoring the terms of $O(n^{-2})$ and noticing $\mu_1 - \mu_2 = -(\mu - \mu_1)/(1-\pi)$,

$$\begin{aligned}
\text{Var}\{E(\hat{\mu}_1|m_1, m_2)\} &= \sigma_1^4(\mu_1 - \mu_2)^2\text{Var}\{\alpha^{-1}(\frac{l}{n})(m_1 m_2^{-1} + 1)(\tilde{\pi} - \pi)\} \\
&= \sigma_1^4(\mu - \mu_1)^2(1-\pi)^{-3}(1-\rho)^2\alpha_0^{-1}\pi\rho^{-1}n^{-1}, \tag{A.9}
\end{aligned}$$

where $\alpha_0 = \sigma_2^2 + \pi(1-\pi)^{-1}\sigma_1^2 + \pi(1-\pi)^{-2}(\mu - \mu_1)^2$.

From (A.6) and the fact that $\text{Cov}(\bar{X}, T|m_1, m_2) = (l/n)\sigma_1^2/m_2$,

$$\begin{aligned}
&\text{Var}(\hat{\mu}_1|m_1, m_2) \\
&= \frac{\sigma_1^2}{m_1} + \sigma_1^4\alpha^{-2}\{\frac{l}{nm_2}\}\{\alpha - (1 - \frac{l}{n})m_1 m_2^{-1}(m_1 m_2^{-1} + 1)(\mu - \mu_1)^2\} - 2\sigma_1^4\frac{\alpha^{-1}l}{nm_2} \\
&= \frac{\sigma_1^2}{m_1} - \frac{\sigma_1^4\alpha^{-1}l}{nm_2} - \sigma_1^4\alpha^{-2}(1 - \frac{l}{n})m_1 m_2^{-1}(m_1 m_2^{-1} + 1)(\mu - \mu_1)^2.
\end{aligned}$$

Therefore,

$$E\{\text{Var}\,(\hat{\mu}_1|m_1,m_2)\} = \text{Var}\,(\bar{X}) - \frac{(1-\rho)\sigma_1^4}{\alpha_0\rho(1-\pi)n} - \frac{(1-\rho)^2\pi(\hat{\mu}_1-\hat{\mu}_2)^2\sigma_1^4}{(1-\pi)\rho\alpha_0^2 n} + O(n^{-2}).$$

This together with (A.9) gives (10). A similar analysis leads to (11).

Derivation of (13). Recall from (9) that $\hat{\pi} = \tilde{\pi} + \tilde{\pi}\hat{\alpha}^{-1}T(\bar{W} - \bar{X}) + O_p(n^{-1})$. Then

$$\text{Var}\,(\hat{\pi}) = \text{Var}\,(\tilde{\pi}) + 2\,\text{Cov}\,(\tilde{\pi}, \tilde{\pi}T)\frac{\mu - \mu_1}{\alpha_0} + \text{Var}\,(\tilde{\pi}T)\frac{(\mu - \mu_1)^2}{\alpha_0^2} + O(n^{-2}).$$

From (A.5) and (A.6),

$$\begin{aligned}
\text{Cov}\,(\tilde{\pi}, \tilde{\pi}T) &= \text{Cov}\,\{\tilde{\pi}, \tilde{\pi}E(T|m_1,m_2)\} \\
&= \text{Cov}\,\{\tilde{\pi}, \tilde{\pi}(\frac{l}{n})(m_1m_2-1+1)(\tilde{\pi} - \pi)(\mu_1 - \mu_2)\} \\
&= n^{-1}\pi(1-\rho)(\mu_1 - \mu_2)\frac{\text{Var}\,(\tilde{\pi})}{1-\pi} \\
&= -n^{-1}\frac{\pi^2(1-\rho)(\mu - \mu_1)}{(1-\pi)\rho} + O(n^{-2}),
\end{aligned} \tag{A.10}$$

$$\begin{aligned}
\text{Var}\,(\tilde{\pi}T) &= \text{Var}\,\{\tilde{\pi}E(T|m_1,m_2)\} + E\{\text{Var}\,(T|m_1,m_2)\} \\
&= \text{Var}\,\{\tilde{\pi}(\frac{l}{n})(m_1m_2^{-1}+1)(\tilde{\pi} - \pi)(\mu_1 - \mu_2)\} \\
&\quad + E\left[\tilde{\pi}^2\left\{\frac{l\alpha}{nm_2} - l^2m^2(\tilde{\pi} - \pi)^2\frac{(\mu_1 - \mu_2)^2}{n^2m_2^2}\right\}\right] \\
&= \frac{\pi^2(1-\rho)\alpha_0}{(1-\pi)\rho n} + O(n^{-2}).
\end{aligned} \tag{A.11}$$

Combining (A.11) with (A.10), gives (13).

References

Cochran, W. C. (1977). *Sampling Techniques*. John Wiley, New York.

Firth, D. and Bennett, K. E. (1998). Robust models in probability sampling. *J. Royal Statist. Soc. Ser. B* **60**, 3-21.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *Internat. Statist. Rev.* **58**, 109-127.

Hall, P. and Titterington, D. M. (1984). Efficient nonparametric estimation of mixture proportions. *J. Roy. Statist. Soc. Ser. B* **46**, 465-473.

Hall, P. and Titterington, D. M. (1985). The use of uncategorised data to improve the performance of a nonparametric estimator of a mixture density. *J. Roy. Statist. Soc. Ser. B* **47**, 155-163.

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**, 761-770.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics **5**, Institute for Mathematical Statistics: Hayward, CA.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd edition. John Wiley, New York.

Ott, J. (1999). *Analysis of Human Genetic Linkage*. 3rd edition. The Johns Hopkins University Press, Baltimore.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90-120.

Pearson, K. (1893). Contributions to the mathematical theory of evolution. *Philos. Trans.* **A 185**, 71-110.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. skew variation in homogeneous material. *Philos. Trans.* **A 186**, 342-414.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley, New York.

Wand, W. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Department of Statistics, Iowa State University, Ames, IA 50010-1210, U.S.A.

E-mail: songchen@iastate.edu

National Institute of Allergy and Infectious Disease, National Institute of Health, Bethesda, MD 20892, U.S.A.

E-mail: jingqin@niaid.nih.gov