Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Semiparametric Inference in a Genetic Mixture Model

Pengfei Li, Yukun Liu & Jing Qin

To cite this article: Pengfei Li, Yukun Liu & Jing Qin (2017) Semiparametric Inference in a Genetic Mixture Model, Journal of the American Statistical Association, 112:519, 1250-1260, DOI: 10.1080/01621459.2016.1208614

To link to this article: http://dx.doi.org/10.1080/01621459.2016.1208614


View supplementary material


Accepted author version posted online: 20 Jul 2016.
Published online: 20 Jul 2016.


Submit your article to this journal


Article views: 528


View related articles


View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=uasa20

Download by: [University of Waterloo] **Date:** 30 October 2017, At: 14:56

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 519, 1250–1260, Theory and Methods
https://doi.org/10.1080/01621459.2016.1208614

Taylor & Francis
Taylor & Francis Group

Check for updates

Semiparametric Inference in a Genetic Mixture Model

Pengfei Li[a], Yukun Liu[b], and Jing Qin[c]

[a]Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Ontario, Canada; [b]School of Statistics, East China Normal University, Shanghai, China; [c]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD

ABSTRACT

In genetic backcross studies, data are often collected from complex mixtures of distributions with known mixing proportions. Previous approaches to the inference of these genetic mixture models involve parameterizing the component distributions. However, model misspecification of any form is expected to have detrimental effects. We propose a semiparametric likelihood method for genetic mixture models: the empirical likelihood under the exponential tilting model assumption, in which the log ratio of the probability (density) functions from the components is linear in the observations. An application to mice cancer genetics involves random numbers of offspring within a litter. In other words, the cluster size is a random variable. We wish to test the null hypothesis that there is no difference between the two components in the mixture model, but unfortunately we find that the Fisher information is degenerate. As a consequence, the conventional two-term expansion in the likelihood ratio statistic does not work. By using a higher-order expansion, we are able to establish a nonstandard convergence rate $N^{-1/4}$ for the odds ratio parameter estimator $\hat{\beta}$. Moreover, the limiting distribution of the empirical likelihood ratio statistic is derived. The underlying distribution function of each component can also be estimated semiparametrically. Analogously to the full parametric approach, we develop an expectation and maximization algorithm for finding the semiparametric maximum likelihood estimator. Simulation results and a real cancer application indicate that the proposed semiparametric method works much better than parametric methods. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2014
Revised June 2016

KEYWORDS
Cluster mixture data; Degenerate Fisher information; Exponential tilting model; Genetic mixture model

Downloaded by [University of Waterloo] at 14:57 30 October 2017

1. Introduction

Finite mixture models have been widely used in psychological, social, and medical research, and more recently in biomedical and genetic studies; see, for example, Sham (1998), Ott (1999), and Efron (2010). A finite mixture model is a probabilistic model for the presence of finitely many subpopulations within an overall population, when the observed data do not have direct information on which subpopulations they come from. For an observed dataset, one needs to find the subpopulation origin using appropriate statistical modeling methods.

The finite mixture models of particular interest in this article come from the backcross design, which has recently become popular in animal study and plant research. In the backcross design, the hybrid and the progenies in subsequent generations are repeatedly backcrossed to one of the parents. Backcrossing may be deliberately employed in animals to transfer a desirable trait in an animal of inferior genetic background to an animal of superior genetic background. As a result, the genotype of the backcross progeny becomes increasingly similar to that of the recurrent parent. In backcrossing studies, the collected data often follow complex mixtures of distribution functions where the mixing proportions are known (Hoff 2000a; Zou, Fine, and Yandell 2002). Hoff (2000a, 2000b) and Hoff et al. (2002) discussed the application of this type of mixture model in cancer genetic studies. Suppose mice with the DD genotype have phenotypes distributed according to $f(x)$; while mice with the Dd genotype have phenotypes distributed according to $g(x)$. A kindred founder mouse with the Dd genotype is bred to one with the DD genotype to produce a new population. The mice in this new population are referred to as subkindred founders; see Figure 1. The law of Mendelian inheritance implies that the genotype of each subkindred founder is Dd or DD, each with probability 50%. The subkindred founders are then mated with a separate population with the DD genotype. The population of the resulting offspring is referred to as the NF population. The phenotypes of this population are then recorded. The data consist of the phenotypes of the offspring of N subkindred founders, that is, the phenotypes of an N-litter of mice,

$$\mathbf{x}_1 = (x_{11}, \ldots, x_{1n_1}), \ldots, \mathbf{x}_N = (x_{N1}, \ldots, x_{Nn_N}),$$

where within the ith litter, the number of offspring n_i is a random variable.

The Mendelian law tells us that conditioning on the litter size n_i the densities of the phenotypes of the n_i offspring have a joint mixture density with mixture proportion 0.5:

$$h(\mathbf{x}_i) = 0.5\prod_{j=1}^{n_i} f(x_{ij}) + 0.5\prod_{j=1}^{n_i}\{0.5f(x_{ij}) + 0.5g(x_{ij})\}. \quad (1.1)$$

CONTACT Yukun Liu ykliu@sfs.ecnu.edu.cn School of Statistics, East China Normal University, Shanghai, China, 200241.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.
© 2017 American Statistical Association

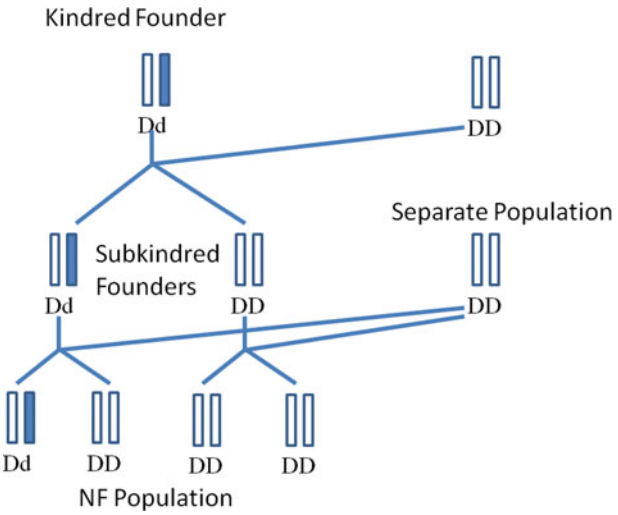

Figure 1. Genotypes of the mice in the backcross design of interest.

In this case, the main interest is to estimate f and g or some functionals of f and g. Hoff (2000a) discussed a technique for calculating the maximum likelihood estimation of probability measures when it is assumed that the measures are constrained to a compact convex set. More on the application of model (1.1) and some variations in genetic studies can be found in sec. 4 of Hoff (2000a).

A natural choice for $f(x)$ and $g(x)$ is the normal distribution. However, in practice these distributions may not be continuous, let alone normal: in the example given by Hoff (2000b) and Hoff et al. (2002), the phenotype is the tumor count. In the genetics literature, little is known about the finite mixture model when the underlying distributions are not fully parameterized. However, model misspecification is a major concern for geneticists since it may lead to biased estimation; see, for example, Sham (1998). It is therefore desirable to make inference on the underlying parameters under minimal assumptions on the underlying component distributions.

Anderson (1979) introduced the semiparametric exponential tilting model into the finite mixture model. In this model, the underlying densities of the two components are assumed to satisfy an exponential tilting model. Specifically,

$$g(x)/f(x) = \exp(\alpha + \beta x), \tag{1.2}$$

where the forms of f and g are not specified beyond this ratio. This model is analogous to the popular two-sample Lehmann's alternative and the Cox proportional hazard, where the two underlying hazard functions are not specified but the ratio of the hazards has a known parametric form. Many familiar exponential families satisfy this model, for example, two normal distributions with different means but a common variance; two exponential distributions; and two negative binomial distributions with different means but the same shape or dispersion parameter. A quadratic term is needed in the model if the two normal distributions have different variances. Moreover, the exponential tilting model has a natural connection to logistic regression if one treats $\delta = 0, 1$ as indicators for the groups with the DD and Dd genotypes, respectively. Among others, Anderson (1979) and Qin (1999) observed that the exponential tilting model is equivalent to the logistic regression model by using the fact that

$$P(\delta = 0|x) = \frac{1}{1 + \exp(\alpha^* + \beta x)},$$

where $\alpha^* = \alpha + \log\{P(\delta = 1)/P(\delta = 0)\}$. Therefore, the exponential tilting approach can be used to predict the genotype for a given phenotype.

Kay and Little (1987) found that the exponential tilting model in (1.2) can be used in various transformed versions. For example,

$$g(x)/f(x) = \exp(\alpha + \beta \log x) = x^{\beta} \exp(\alpha).$$

This is a biased sampling problem discussed by Vardi (1985). In the special cases $\beta = 1, 2, 3$, it corresponds to the case where the probability of being sampled is proportional to the associated length, area, or volume, respectively (Cox 1969; Patil and Rao 1978; Vardi 1982).

In this article, we adapt Anderson's (1979) approach to the genetic mixture model in (1.1). Throughout this article, unless otherwise stated, all developments are conditional on the litter size n_i. Under model (1.2), the joint densities of the phenotypes of the n_i offspring become

$$h(\mathbf{x}_i) = \left[0.5 + 0.5^{(1+n_i)} \prod_{j=1}^{n_i} \{1 + \exp(\alpha + \beta x_{ij})\}\right] \prod_{j=1}^{n_i} f(x_{ij}). \tag{1.3}$$

Note that the underlying parameters cannot be identified from the marginal density

$$h_m(x_{ij}) = \{0.75 + 0.25 \exp(\alpha + \beta x_{ij})\} f(x_{ij})$$

if the density f is not specified, since the finite-dimensional parameters α and β are absorbed by the nonparametric density f (Zou, Fine, and Yandell 2002). In this article, we require that on average the litter size should be at least two, so that the parameters (α, β, f) can be identified from model (1.3). In fact, $h(\mathbf{x}_i)$ and $h_m(x_{ij})$ can be consistently estimated by empirical densities. Therefore,

$$\frac{h(\mathbf{x}_i)}{\prod_{j=1}^{n_i} h_m(x_{ij})} = \frac{0.5 + 0.5^{(1+n_i)} \prod_{j=1}^{n_i} \{1 + \exp(\alpha + \beta x_{ij})\}}{\prod_{j=1}^{n_i} \{0.75 + 0.25 \exp(\alpha + \beta x_{ij})\}}$$

can be identified. As a result, (α, β) and the cumulative distribution function F of f can be consistently estimated.

The exponential tilting model has been investigated extensively in the literature due to its flexibility and efficiency. Recently, Chen and Liu (2013) had found an application of this model in the study of Canadian lumber. Liu et al. (2013) used the model to link the scalar scores of HIV patients with viral failure and those with viral suppression. Carvalho and Davison (2014) applied the model to study the dependence between extreme air temperatures under the forest canopy and in a nearby open field at 14 sites in Switzerland. It is worth pointing out that all these studies considered only situations with a standard convergence rate, that is, $N^{-1/2}$. The results in this article disclose an estimator of β with an $N^{-1/4}$ convergence rate when the true value of β is zero.

In the proposed inference procedure, we handle the nonparametric $f(x)$ by the well-known empirical likelihood method (Owen 2001). Compared with the previous approaches of

Downloaded by [University of Waterloo] at 14:57 30 October 2017

Downloaded by [University of Waterloo] at 14:57 30 October 2017

Anderson (1979), Qin (1999), Zou, Fine, and Yandell (2002), and Tan (2009), our application of empirical likelihood to the exponential tilting mixture model has three major differences:

1. We have only a single random sample in the mice application. Hence, there are no direct observations from f and g or multiple mixture samples of f and g as considered in Zou, Fine, and Yandell (2002) and Tan (2009). In other words, the model considered in this article is fundamentally different from those in Zou, Fine, and Yandell (2002) and Tan (2009), which results in different asymptotic properties of the maximum empirical likelihood estimator of β when the true value of β is zero. More details are given in Section 2.3.
2. Even in the literature on empirical likelihood and selection biased sampling problems (Vardi 1985), there is a lack of general large-sample theory for the summation of random numbers of random variables. Our mice application involves the number of offspring within a litter or the cluster size, which is a bounded random variable. The theoretical derivation becomes rather complex and tedious; see the proofs in the supplementary material.
3. Since the Fisher information is degenerate under the null $f = g$ or $\beta = 0$, a fourth-order Taylor expansion for the likelihood ratio statistic is required to derive its limiting distribution. As a result, the proof is extremely complex compared with existing proofs for related problems, where a second-order Taylor expansion suffices.

The organization of this article is as follows. In Section 2, we present the empirical likelihood inference approach for the mice genetic mixture model in (1.1) under the exponential tilting model assumption in (1.2). An EM-algorithm is suggested for finding the maximum empirical likelihood estimates of the unknown parameters/functions. We show that the limiting distribution of the empirical likelihood ratio test for testing $H_0 : \beta = \beta_0$ is $0.5\chi_0^2 + 0.5\chi_1^2$, an equal mixture of a distribution with point mass at zero and a χ_1^2 distribution, if the true value β_0 of β is 0; and it is χ_1^2 if $\beta_0 \neq 0$. Further, the convergence rate of the maximum empirical likelihood estimator of β is $N^{-1/4}$ when $\beta_0 = 0$; and it becomes $N^{-1/2}$ when $\beta_0 \neq 0$. When $\beta_0 \neq 0$, the maximum empirical likelihood estimator of (α, β) has an asymptotic joint normal distribution, and the underlying distribution function of each component can also be estimated with the rate $N^{-1/2}$. We present a simulation study in Section 3 and discuss a real example in Section 4. Section 5 provides concluding remarks. For convenience of presentation, all the proofs are given in a supplementary document.

2. Semiparametric Likelihood for Genetic Mixture Model

2.1. Parameter Estimation

Let $p_{ij} = dF(x_{ij})$. Then it follows from model (1.2) that $dG(x_{ij}) = \exp(\alpha + \beta x_{ij}) p_{ij}$. Based on the given x_{ij}'s, the empirical likelihood of (α, β, F) is defined as

$$L_N(\alpha, \beta, F) = \prod_{i=1}^{N} \left\{ \left[0.5 + 0.5^{(1+n_i)} \prod_{j=1}^{n_i} \{1 + \exp(\alpha + \beta x_{ij})\} \right] \prod_{j=1}^{n_i} p_{ij} \right\}. \tag{2.4}$$

The feasible p_{ij}'s satisfy

$$p_{ij} \geq 0, \quad \sum_{i=1}^{N} \sum_{j=1}^{n_i} p_{ij} = 1,$$
$$\sum_{i=1}^{N} \sum_{j=1}^{n_i} p_{ij} \{\exp(\alpha + \beta x_{ij}) - 1\} = 0, \tag{2.5}$$

which implies that the empirical likelihood models F by $F(x) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} p_{ij} I(x_{ij} \leq x)$.

Inferences about (α, β) are usually made through the profile empirical likelihood or log-likelihood function, $l_N(\alpha, \beta) = \sup_F \log\{L_N(\alpha, \beta, F)\}$, where the maximum is taken under constraint (2.5) given (α, β). By the Lagrange multiplier method, we find that

$$p_{ij} = p_{ij}(\alpha, \beta) = \frac{1}{\sum_{i=1}^{N} n_i} \frac{1}{1 + \gamma(e^{\alpha+\beta x_{ij}} - 1)},$$

where γ is the solution to

$$\sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{e^{\alpha+\beta x_{ij}} - 1}{1 + \gamma(e^{\alpha+\beta x_{ij}} - 1)} = 0.$$

The resulting profile empirical log-likelihood is

$$l_N(\alpha, \beta) = \sum_{i=1}^{N} \log \left\{ 0.5 + 0.5^{(1+n_i)} \prod_{j=1}^{n_i} (1 + e^{\alpha+\beta x_{ij}}) \right\} - \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log\{1 + \gamma(e^{\alpha+\beta x_{ij}} - 1)\}. \tag{2.6}$$

Let the maximum empirical likelihood estimator of (α, β) be $(\hat{\alpha}, \hat{\beta}) = \arg\sup_{\alpha,\beta} l_N(\alpha, \beta)$. We then estimate the cumulative distribution functions $F(x)$ and $G(x)$ by

$$\hat{F}(x) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \hat{p}_{ij} I(x_{ij} \leq x) \quad \text{and}$$
$$\hat{G}(x) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \hat{p}_{ij} e^{\hat{\alpha}+\hat{\beta} x_{ij}} I(x_{ij} \leq x),$$

respectively, where $\hat{p}_{ij} = p_{ij}(\hat{\alpha}, \hat{\beta})$. We further estimate the population means μ_F and μ_G of $F(x)$ and $G(x)$, respectively, by

$$\hat{\mu}_F = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \hat{p}_{ij} x_{ij} \quad \text{and} \quad \hat{\mu}_G = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \hat{p}_{ij} e^{\hat{\alpha}+\hat{\beta} x_{ij}} x_{ij},$$

Downloaded by [University of Waterloo] at 14:57 30 October 2017

and the population variances σ_F^2 and σ_G^2 of $F(x)$ and $G(x)$, respectively, by

$$\hat{\sigma}_F^2 = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \hat{p}_{ij}(x_{ij} - \hat{\mu}_F)^2 \quad \text{and}$$

$$\hat{\sigma}_G^2 = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \hat{p}_{ij} e^{\hat{\alpha}+\hat{\beta}x_{ij}}(x_{ij} - \hat{\mu}_G)^2.$$

The explicit forms of $(\hat{\alpha}, \hat{\beta})$ and the $\hat{p}_{ij}$'s are unknown in general. In the next subsection, we present an EM-algorithm to search for these estimates.

2.2. EM-Algorithm

Since the genotype of each observation is missing, we need to deal with the complicated mixture structure in the likelihood (2.4), which makes the maximization of (2.6) difficult. The EM-algorithm naturally fits into our problem. We first define the missing data. Let $z_{ij} = 1$ if the jth mouse in the ith litter has the Dd genotype, and 0 if the jth mouse in the ith litter has the DD genotype. That is, the z_{ij}'s are the missing labels for all the observations. Further, let $\mathcal{X} = (x_{11}, \ldots, x_{Nn_N})$ be the observed phenotypes, $\mathcal{Z} = (z_{11}, \ldots, z_{Nn_N})$ be the missing labels, and $\boldsymbol{\Theta} = (\alpha, \beta, p_{11}, \ldots, p_{Nn_N})$.

Conditional on $z_{ij} = 1$ or 0, x_{ij} has the cumulative distribution function $G(x)$ or $F(x)$, respectively. Further the x_{ij}'s are conditionally independent given the z_{ij}'s. It can be verified that

$$P(z_{i1} = \cdots = z_{in_i} = 0) = 0.5 + 0.5^{n_i+1} \text{ and}$$

$$P(z_{i1} = a_{i1}, \ldots, z_{in_i} = a_{in_i}) = 0.5^{n_i+1}$$

for $a_{ij} = 0$ or 1, and $(a_{i1}, \ldots, a_{in_i}) \neq (0, \ldots, 0)$. Hence, based on the complete data $\{\mathcal{X}, \mathcal{Z}\}$, the log-likelihood of $\boldsymbol{\Theta}$ has the following form (up to a constant not depending on $\boldsymbol{\Theta}$):

$$\begin{aligned} l_c(\boldsymbol{\Theta}) &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\{(1 - z_{ij}) \log dF(x_{ij}) + z_{ij} \log dG(x_{ij})\} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\{z_{ij}(\alpha + \beta x_{ij}) + \log(p_{ij})\}, \end{aligned}$$

where we have used the exponential tilting model assumption in (1.2). The concise form of the complete log-likelihood makes it convenient to develop the EM algorithm.

The core of the EM-algorithm is the EM-iteration, which contains an E-step and an M-step. We use $\boldsymbol{\Theta}^{(r-1)} = (\alpha^{(r-1)}, \beta^{(r-1)}, p_{11}^{(r-1)}, \ldots, p_{Nn_N}^{(r-1)})$ to denote the value of $\boldsymbol{\Theta}$ after $r-1$ EM-iterations, $r = 1, 2, \ldots$. When $r = 1$, $\boldsymbol{\Theta}^{(0)}$ denotes the initial value of $\boldsymbol{\Theta}$.

In the E-step of the rth iteration, we need to calculate

$$\begin{aligned} Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(r-1)}) &= \mathbb{E}\{l_c(\boldsymbol{\Theta})|\mathcal{X}, \boldsymbol{\Theta}^{(r-1)}\} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left\{\mathbb{E}\left(z_{ij}|\mathbf{x}_i, \boldsymbol{\Theta}^{(r-1)}\right)(\alpha + \beta x_{ij}) + \log p_{ij}\right\}, \end{aligned}$$

where the expectation is with respect to the conditional distribution of $\mathcal{Z}$ given $\mathcal{X}$ and using $\boldsymbol{\Theta}^{(r-1)}$ for $\boldsymbol{\Theta}$. Let $w_{ij}^{(r)} = \mathbb{E}(z_{ij}|\mathbf{x}_i, \boldsymbol{\Theta}^{(r-1)})$. In the supplementary material, we show that

$$w_{ij}^{(r)} = \frac{\prod_{k=1}^{n_i}(0.5 + 0.5e^{\alpha^{(r-1)}+\beta^{(r-1)}x_{ik}})}{1 + \prod_{k=1}^{n_i}(0.5 + 0.5e^{\alpha^{(r-1)}+\beta^{(r-1)}x_{ik}})} \cdot \frac{e^{\alpha^{(r-1)}+\beta^{(r-1)}x_{ij}}}{1 + e^{\alpha^{(r-1)}+\beta^{(r-1)}x_{ij}}}. \tag{2.7}$$

In the M-step of the rth iteration, we update $\boldsymbol{\Theta}$ by $\boldsymbol{\Theta}^{(r)}$, which maximizes

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(r-1)}) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left\{w_{ij}^{(r)}(\alpha + \beta x_{ij}) + \log(p_{ij})\right\}$$

with respect to $\boldsymbol{\Theta}$ under the constraints in (2.5). Using the approach of Zhang (2002), we can perform the above maximization in the following steps (see the detailed explanation in the supplementary material):

Step 1. Update

$$p_{ij}^{(r)}(\alpha, \beta) = \frac{1}{\sum_{i=1}^{N} n_i} \frac{1}{1 - \gamma^{(r)} + \gamma^{(r)} \exp(\alpha + \beta x_{ij})},$$

where $\gamma^{(r)} = \sum_{i=1}^{N}\sum_{j=1}^{n_i} w_{ij}^{(r)} / \sum_{i=1}^{N} n_i$.

Step 2. Substitute $p_{ij}^{(r)}(\alpha, \beta)$ into the Q-function to get

$$\begin{aligned} Q^{(r)}(\alpha, \beta) &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left[w_{ij}^{(r)}(\alpha + \beta x_{ij}) + \log\left\{p_{ij}^{(r)}(\alpha, \beta)\right\}\right] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left[w_{ij}^{(r)}(\alpha + \beta x_{ij}) - \log\left\{1 - \gamma^{(r)} \right.\right. \\ &\quad \left.\left. + \gamma^{(r)} \exp(\alpha + \beta x_{ij})\right\}\right] + \text{constant}, \end{aligned}$$

where the constant does not depend on (α, β). Maximize $Q^{(r)}(\alpha, \beta)$ to get $(\alpha^{(r)}, \beta^{(r)})$.

Step 3. Update p_{ij} via

$$p_{ij}^{(r)} = \frac{1}{\sum_{i=1}^{N} n_i} \frac{1}{1 - \gamma^{(r)} + \gamma^{(r)} \exp(\alpha^{(r)} + \beta^{(r)} x_{ij})}.$$

The E-step and M-step are iterated until convergence.

We make two remarks about the above EM-algorithm. First, following the proof in Dempster, Laird, and Rubin (1977) and that in Zhang (2002), we can show that the empirical likelihood $L_N(\alpha, \beta, F)$ does not decrease after each iteration. That is, for $r \geq 2$

$$L_N(\alpha^{(r)}, \beta^{(r)}, F^{(r)}) \geq L_N(\alpha^{(r-1)}, \beta^{(r-1)}, F^{(r-1)}).$$

Further, note that $L_N(\alpha, \beta, F) = \prod_{i=1}^{N}\{0.5\prod_{j=1}^{n_i} p_{ij} + 0.5\prod_{j=1}^{n_i}(0.5p_{ij} + 0.5q_{ij})\} \leq 1$ and $L_N(\alpha, \beta, F)$ is a continuous function of all the unknown parameters. Then the sequence $\{L_N(\alpha^{(r)}, \beta^{(r)}, F^{(r)})\}$ eventually converges to a stationary value of $L_N(\alpha, \beta, F)$ for a given initial value $\boldsymbol{\Theta}^{(0)}$ (Wu 1983). However, this stationary value may not be a global maximum. Even in the full parametric mixture model, there is no guarantee that the EM algorithm leads to the global maximum. The semiparametric approach has the same problem. To increase the possibility of finding the global maximum, we recommend using multiple initial values. Our simulation results demonstrate that this method works well. Second, in practice, we may stop the algorithm when the increment in the log empirical likelihood after an iteration is no greater than,

Downloaded by [University of Waterloo] at 14:57 30 October 2017

say, 10^{-6}. The EM-algorithm converges very fast according to our simulations. For instance, for the real data analyzed in Section 4, the algorithm stops after 21 iterations starting from $(\alpha^{(0)}, \beta^{(0)}) = (1, -1)$. Further, it takes less than 1 sec for the above calculation in an IMAC with a 3.4-GHz Intel Core i7 processor.

2.3. Asymptotic Properties

One problem of practical and scientific interest is whether or not the phenotype distributions of the mice with the DD genotype and those with the Dd genotype are the same, or equivalently $\beta = 0$ or $\beta \neq 0$. In this subsection, we first investigate the asymptotic properties of the empirical log-likelihood ratio test (ELRT) under model (1.3) for the null hypothesis $H_0 : \beta = 0$.

Define the empirical log-likelihood ratio function of β as

$$R_N(\beta) = 2\left\{l_N(\hat{\alpha}, \hat{\beta}) - \sup_{\alpha} l_N(\alpha, \beta)\right\}.$$

The following theorem presents the limiting distribution of $R_N(0)$ under H_0.

Theorem 1. Suppose $2 \leq n_i \leq C$ for some given positive integer C, the x_{ij}'s take at least two values and are independent of n_i, and $\int e^{\beta x} f(x)dx < \infty$ in a neighborhood of $\beta = 0$. Under the null hypothesis $H_0 : \beta = 0$, as $N \to \infty$ we have

(a) $\hat{\beta} = O_p(N^{-1/4})$;
(b) the limiting distribution of $R_N(0)$ is $0.5\chi_0^2 + 0.5\chi_1^2$, an equal mixture of a distribution with point mass at zero and a χ_1^2 distribution.

For presentational continuity, the proofs of Theorem 1 and those for Theorems 2 and 3 are given in the supplementary document. The assumption that $\int e^{\beta x} f(x)dx < \infty$ in a neighborhood of $\beta = 0$ implies the existence of the moment generating function of x_{ij} and therefore all its finite moments. This fact will be used in our proofs of Theorems 1 and 2.

If the true value of β is not equal to 0, the large-sample properties of the empirical likelihood estimator $\hat{\beta}$ and the empirical log-likelihood ratio for testing $H_0 : \beta = \beta_0$ for $\beta_0 \neq 0$ are different from those in Theorem 1.

Theorem 2. Let (α_0, β_0) denote the true value of (α, β) and assume $(\alpha_0, \beta_0) \neq (0, 0)$. Suppose $2 \leq n_i \leq C$ for some given positive integer C, the x_{ij}'s take at least two values and are independent of n_i, and $\int e^{\beta x} f(x)dx < \infty$ in a neighborhood of $\beta = \beta_0$ and $\beta = 0$. Let $\mathbf{\Sigma}$ be the matrix defined in Equation (A.1), and assume $\mathbf{\Sigma}$ is positive definite. As $N \to \infty$,

(a) the limiting distribution of $\sqrt{N}(\hat{\alpha} - \alpha_0, \hat{\beta} - \beta_0)^\tau$ is $N(0, \mathbf{\Sigma})$;
(b) the limiting distribution of $R_N(\beta_0)$ is χ_1^2.

We now give some insight into the difference between the asymptotic results in Theorems 1 and 2. In the proof of Theorem 1, we encounter two types of irregularities. First, $\beta = 0$ implies that $\alpha = 0$, which means that α and β are not completely free at the null hypothesis $H_0 : \beta = 0$. This irregularity was first pointed out by Zou, Fine, and Yandell (2002) when they applied the exponential tilting model to mixtures of two univariate distributions with known mixing proportions. To overcome this irregularity, Zou, Fine, and Yandell (2002) proposed a partial empirical likelihood method. They further showed that the maximum partial likelihood estimator of β is $\sqrt{N}$ consistent and asymptotically normal whether or not $\beta = 0$, and the profile log-likelihood ratio for testing $\beta = 0$ has a χ_1^2 limiting distribution. Their results are in sharp contrast to our Theorem 1. This is because of the second type of irregularity, degenerate Fisher information at $\beta = 0$, in our setup. Let $pl_N(\beta) = \sup_\alpha l_N(\alpha, \beta)$. It can be verified that

$$\mathbb{E}\left[\frac{d^2 pl_N(\beta)}{d\beta^2}\Big|_{\beta=0}\right] = 0.$$

This implies that after profiling out α, the Fisher information of β is degenerate at $\beta = 0$. Tan (2009) showed that after α is profiled out, the Fisher information of β is not degenerate under the setup of Zou, Fine, and Yandell (2002) whether or not $\beta = 0$. Hence, $\sqrt{N}$ consistency and an asymptotic χ_1^2 limiting distribution are expected for the maximum partial likelihood estimator of β and the profile log-likelihood ratio for testing $\beta = 0$, respectively. However, in our setup, degenerate Fisher information at $\beta = 0$ results in the second-order Taylor expansion being insufficient to approximate $l_N(\alpha, \beta)$ in the neighborhood of $(0, 0)$. In the supplementary material, we show that it is necessary to use a fourth-order Taylor expansion to approximate $l_N(\alpha, \beta)$, and hence the convergence rate of $\hat{\beta}$ becomes $N^{-1/4}$ instead of $N^{-1/2}$ when the true value of β is 0. When the true value β_0 of β is not equal to 0, the above two types of irregularity do not exist. Therefore, a second-order Taylor expansion is sufficient to find the leading term of $l_N(\alpha, \beta)$ in the neighborhood of (α_0, β_0). The quadratic approximation of $l_N(\alpha, \beta)$ enables us to derive the χ_1^2 limiting distribution of $R_N(\beta_0)$ and the joint asymptotic normality of $(\hat{\alpha}, \hat{\beta})$.

Now we consider the asymptotic properties of the proposed estimators $\hat{F}(x)$ and $\hat{G}(x)$ of $F(x)$ and $G(x)$ when $(\alpha_0, \beta_0) \neq (0, 0)$. Because of the $\sqrt{N}$-consistency and joint asymptotic normality of $(\hat{\alpha}, \hat{\beta})$, we have the following results for $\hat{F}(x)$ and $\hat{G}(x)$, which imply that $\hat{F}(x)$ and $\hat{G}(x)$ are consistent and have the convergence rate $N^{-1/2}$.

Theorem 3. Assume the conditions of Theorem 2. As $N \to \infty$, we have that $\sqrt{N}\{\hat{F}(x) - F(x), \hat{G}(y) - G(y)\}$ converges weakly to a bivariate Gaussian process $\mathbf{B}(\mathbf{s})$ with zero mean, independent increment, and covariance structure $\mathbf{\Omega}(\mathbf{s}_1, \mathbf{s}_2)$ defined in (A.2). Here $\mathbf{s} = (x, y)^\tau$, $\mathbf{s}_1 = (x_1, y_1)^\tau$, and $\mathbf{s}_2 = (x_2, y_2)^\tau$.

3. Simulation Study

3.1. Setup

In this section, we conduct Monte Carlo simulation to provide insight into the following questions:

(a) When testing $H_0 : \beta = 0$, does the limiting distribution provide an accurate approximation to the finite-sample distribution of the ELRT? Is the ELRT comparable to the parametric likelihood ratio test (PLRT) when the model is correctly specified and more powerful when the model is misspecified?
(b) If the true value of β is nonzero, are the proposed maximum EL estimators for μ_F, μ_G, σ_F^2, σ_G^2, $F(x)$, and

Downloaded by [University of Waterloo] at 14:57 30 October 2017

$G(x)$ comparable to those based on the correct model, and more efficient than those based on the misspecified model?

In our simulation studies, we use a tumor-count dataset (Hoff 2000b) that contains the tumor counts of 21 litters of mice. Drinkwater and Klotz (1981) and Hoff et al. (2002) suggested using a negative binomial distribution to model this dataset. To generate data from Model (1.1), we need to specify $f(x)$, $g(x)$, and the distribution of n_i. We consider two scenarios for $f(x)$ and $g(x)$.

Scenario I: We choose $f(x)$ and $g(x)$ to be the probability mass functions of two negative binomial distributions with the common shape or dispersion parameter η and means μ_F and μ_G, respectively. That is,

$$f(x) = \frac{\Gamma(x+\eta)}{\Gamma(x+1)\Gamma(\eta)}\left(\frac{\eta}{\eta+\mu_F}\right)^{\eta}\left(\frac{\mu_F}{\eta+\mu_F}\right)^{x}, \quad (3.8)$$

$$g(x) = \frac{\Gamma(x+\eta)}{\Gamma(x+1)\Gamma(\eta)}\left(\frac{\eta}{\eta+\mu_G}\right)^{\eta}\left(\frac{\mu_G}{\eta+\mu_G}\right)^{x}. \quad (3.9)$$

Then $\log\{g(x)/f(x)\} = \alpha + \beta x$ with

$$\alpha = \eta \log\frac{\eta+\mu_F}{\eta+\mu_G} \text{ and } \beta = \log\frac{\mu_G(\eta+\mu_F)}{\mu_F(\eta+\mu_G)}. \quad (3.10)$$

Therefore, the ratio $g(x)/f(x)$ satisfies the exponential tilting model in (1.2).

Scenario II: We first fit the tumor-count dataset in Hoff (2000b) by the proposed estimation procedure and obtain the maximum empirical likelihood estimator $\hat{f}(x)$ of $f(x)$, the probability mass function for the phenotype of mice that are noncarriers of the particular allele. The cumulative distribution function of $\hat{f}(x)$ is given in Section 4. We then set $f(x)$ to $\hat{f}(x)$ and $g(x)$ such that $g(x)/f(x)$ satisfies the exponential tilting model in (1.2). The specific value of β will be given later.

In all scenarios, the n_i's are randomly generated from the set {3, 4, 4, 5, 8, 8, 8, 8, 9, 9, 10, 13, 15, 16, 16, 17, 17, 18, 19, 20, 22}; these are the litter sizes of the 21 litters of mice in the tumor-count dataset.

We fit the data generated from each scenario by the exponential tilting model in (1.2) with the empirical likelihood, and the parametric model in (3.8)–(3.9) with the parametric likelihood, respectively. It is worth mentioning that the exponential tilting model assumption is always satisfied in both scenarios. However, the parametric model in (3.8)–(3.9) is valid only in Scenario I. Hence, the parametric model is misspecified in Scenario II for any estimation and testing procedures.

3.2. Testing $H_0 : \beta = 0$

The purpose of this subsection is to address question (a). We first check the performance of the limiting distribution. We set $\eta = 5, \mu_F = \mu_G = 4$ in Scenario I, and $\beta = 0$ in Scenario II. We choose $N = 20$, which is close to the number of litters in the tumor-count dataset. We calculate the Type I error rates of the ELRT under the exponential tilting model assumption (1.2) and of the PLRT under the parametric model assumption in (3.8)–(3.9) based on 50,000 repetitions. Recall that for both scenarios, the model is correctly specified for the ELRT, while for the PLRT the model is correct under Scenario I but misspecified under Scenario II. At the 5% and 1% levels, the simulated Type I error rates of the ELRT are, respectively, 5.9% and 1.3% under Scenario I, and 6.2% and 1.4% under Scenario II. In comparison, those for the PLRT are, respectively, 6.2% and 1.3% under Scenario I, and 23.2% and 8.8% under Scenario II. Clearly, the limiting distribution of the ELRT provides a satisfactory approximation to the finite-sample distribution under both scenarios. If the parametric model is correct (i.e., Scenario I), the limiting distribution of the PLRT also works reasonably well, but if the model is misspecified (i.e., Scenario II), this distribution is stochastically much smaller than the finite-sample distribution. Hence, in this case the Type I error rates of the PLRT based on the limiting distribution are much larger than the corresponding true values.

Next we compare the powers of the ELRT and PLRT under alternative models. In Scenario I, we set $\eta = 5$, $\mu_G = 4$, and choose 10 values of β: $-0.05, \ldots, -0.5$ with μ_F being determined by (3.10). The same 10 values for β are considered in Scenario II. For a fair comparison, we take the simulated distributions of the ELRT and PLRT under the null hypothesis based on 50,000 repetitions as reference distributions, and we calculate their p-values and critical values. The powers of the ELRT and PLRT under the alternative models are calculated based on 2000 repetitions. We still consider the sample size $N = 20$. The results under the 5% and 1% significance levels are plotted in Figure 2. We can see that if the model is correctly specified for both the ELRT and PLRT (i.e., Scenario I), then the ELRT and PLRT have almost the same power for detecting a difference between $f(x)$ and $g(x)$. However, if the model is misspecified (i.e., Scenario II) for the PLRT, then the PLRT is less powerful than the ELRT.

3.3. Estimating μ_F, μ_G, σ_F^2, σ_G^2, $F(x)$, and $G(x)$

We now address question (b). We choose $\eta = 5$, $\mu_G = 4$, $\mu_F = 12$, 18, and 24 in Scenario I, and $\beta = -0.3, -0.45, -0.6$ in Scenario II. In Scenario II, $\mu_F = 19.58$; the values of μ_G corresponding to $\beta = -0.3, -0.45$, and -0.5 are 6.96, 4.84, and 3.47, respectively. We consider two choices of N: 20 and 200.

We first compare the estimation of (μ_F, σ_F^2) and (μ_G, σ_G^2). We use $(\hat{\mu}_{F,p}, \hat{\sigma}^2_{F,p})$ and $(\hat{\mu}_{G,p}, \hat{\sigma}^2_{G,p})$ to denote the estimates of (μ_F, σ_F^2) and (μ_G, σ_G^2), respectively, under the parametric model assumption in (3.8)–(3.9). Tables 1 and 2 give the bias, variance (Var), and mean square error (MSE) for each estimator based on 2000 repetitions under the two scenarios. As expected, the parametric maximum likelihood estimators of (μ_F, σ_F^2) and (μ_G, σ_G^2) based on the correct model (i.e., Scenario I) are more efficient than the maximum empirical likelihood estimators. We also observe that the maximum empirical likelihood estimator of (μ_F, σ_F^2) is comparable to the maximum likelihood estimator when the sample size is large. When the parametric model is misspecified (i.e., Scenario II), the maximum empirical likelihood estimators of (μ_F, σ_F^2) and (μ_G, σ_G^2) are more efficient (in most cases) than or at least comparable to the maximum

Downloaded by [University of Waterloo] at 14:57 30 October 2017

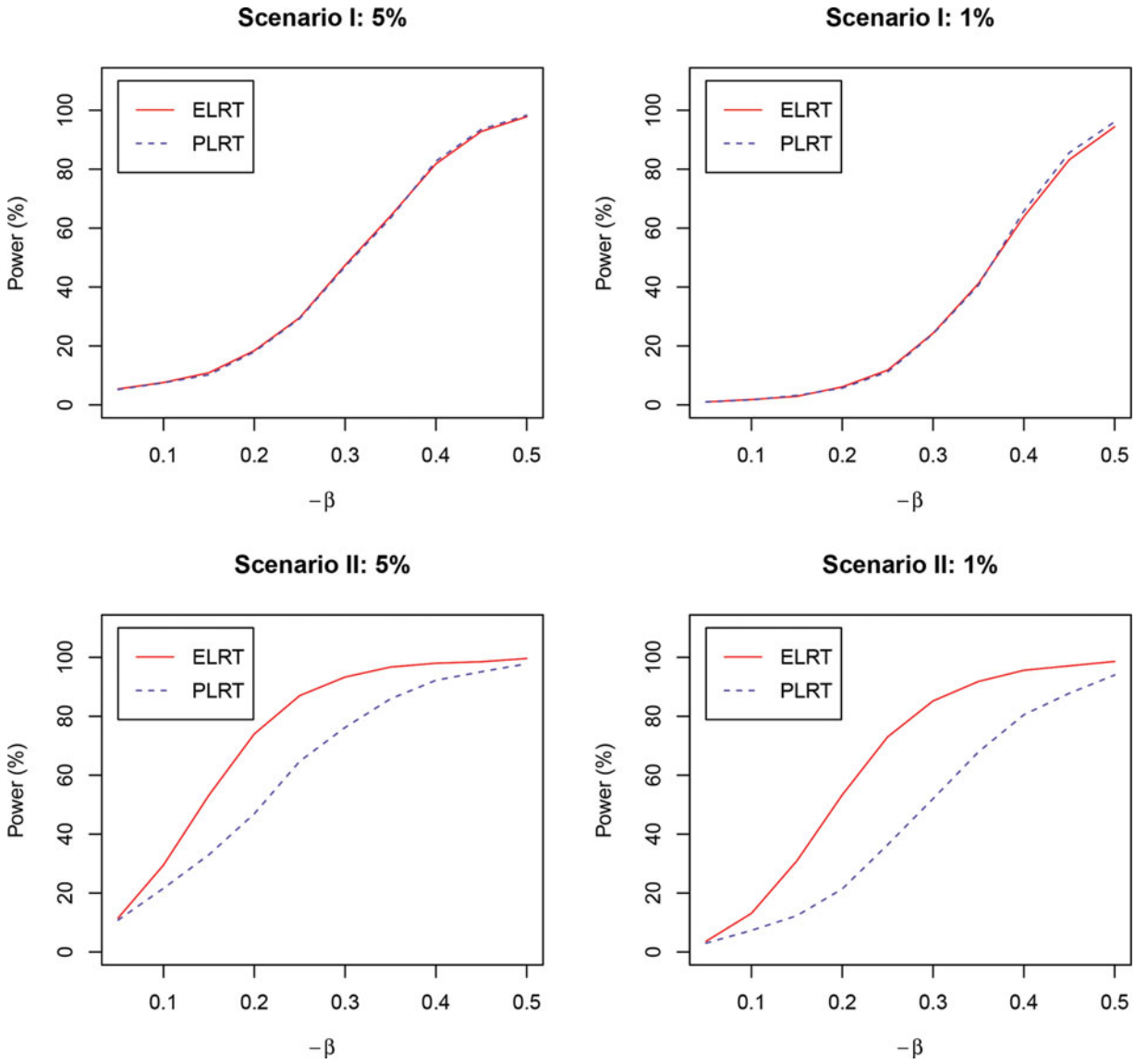

Figure 2. Power comparison between the ELRT and PLRT under Scenarios I and II at the significance levels 5% and 1%: the powers are calculated based on 2000 repetitions, and 50,000 repetitions under the null model are used to calculate the *p*-values of the ELRT and PLRT.

parametric likelihood estimators. For both methods, as N increases, the MSEs decrease, as expected.

Now we turn to the estimation of $F(x)$ and $G(x)$. We use $\hat{F}_p(x)$ and $\hat{G}_p(x)$ to denote the respective estimators of $F(x)$ and $G(x)$ under the parametric model. The Kolmogorov–Smirnov distance between the estimated and true cumulative distribution functions is used as a basis for comparison. Table 3 gives the average Kolmogorov–Smirnov distance based on 2000 repetitions. When the parametric model is correct (Scenario I), the parametric estimators of $F(x)$ and $G(x)$ are more accurate than the proposed distribution estimators. If the parametric model is misspecified (Scenario II), the proposed estimators become more accurate. This provides evidence for the robustness of the proposed estimators. Finally, the Kolmogorov–Smirnov distances of both methods decrease as N increases under each model.

In the supplementary material, we consider one more scenario in addition to Scenarios I and II, in which the parametric model is correctly specified while the exponential tilting model is misspecified. We summarize the observations as follows:

- model misspecification on the exponential tilting model seems to have no effect on the Type I error rate and the power of the ELRT for testing $H_0 : \beta = 0$;
- as expected, model misspecification deteriorates the performance of the maximum empirical likelihood estimators of (μ_F, σ_F^2), (μ_G, σ_G^2), $F(x)$, and $G(x)$.

We comment that since the exponential tilting model assumption is weaker than the full parametric model, in general we are not likely to misspecify the exponential tilting model but correctly specify the parametric model.

4. Real Example

Hoff (2000b) analyzed a tumor-count dataset collected from 74 subkindreds, with tumor counts from 968 mice. His analysis was based on the tumor counts from 21 randomly selected litters. Following Hoff (2000b), we also analyzed the observations from these 21 litters.

Panel (a) of Figure 3 presents the estimates of $F(x)$ and $G(x)$ under the exponential tilting model assumption in (1.2) and

Table 1. Comparing the biases, variances, and mean square errors (MSEs) of the estimators of (μ_F, σ_F^2) and (μ_G, σ_G^2) under the exponential tilting and parametric models under Scenario I, in which the model is correctly specified for $(\hat{\mu}_F, \hat{\sigma}_F^2, \hat{\mu}_G, \hat{\sigma}_G^2)$ and $(\hat{\mu}_{F,p}, \hat{\sigma}_{F,p}^2, \hat{\mu}_{G,p}, \hat{\sigma}_{G,p}^2)$.

		Exponential tilting				Parametric model			
(μ_F, μ_G)	Summary	$\hat{\mu}_F$	$\hat{\sigma}_F^2$	$\hat{\mu}_G$	$\hat{\sigma}_G^2$	$\hat{\mu}_{F,p}$	$\hat{\sigma}_{F,p}^2$	$\hat{\mu}_{G,p}$	$\hat{\sigma}_{G,p}^2$
		Scenario I: $N = 20$							
(12,4)	Bias	−0.164	−0.542	0.457	1.972	−0.054	−0.184	0.197	1.058
	Var	0.908	40.368	4.343	74.405	0.513	34.001	1.415	40.867
	MSE	0.935	40.656	4.551	78.284	0.516	34.031	1.454	41.982
(18,4)	Bias	−0.063	0.043	0.167	1.078	0.002	−0.145	0.069	0.294
	Var	0.841	137.934	1.616	60.541	0.619	123.792	0.402	9.539
	MSE	0.845	137.919	1.644	61.695	0.619	123.798	0.407	9.624
(24,4)	Bias	−0.030	−0.154	0.088	0.757	0.007	−0.712	0.036	0.147
	Var	1.067	381.684	0.579	35.795	0.934	329.812	0.292	9.766
	MSE	1.068	381.684	0.587	36.366	0.934	330.298	0.293	9.787
		Scenario I: $N = 200$							
(12,4)	Bias	−0.004	−0.033	0.011	0.045	−0.002	−0.033	0.009	0.031
	Var	0.035	3.250	0.040	0.690	0.034	3.088	0.031	0.289
	MSE	0.035	3.251	0.040	0.692	0.034	3.088	0.031	0.290
(18,4)	Bias	0.001	−0.040	0.008	0.044	0.004	−0.051	0.005	0.015
	Var	0.059	12.900	0.033	0.786	0.057	11.750	0.023	0.186
	MSE	0.059	12.900	0.033	0.787	0.057	11.751	0.023	0.187
(24,4)	Bias	−0.003	−0.071	0.007	0.054	−0.001	−0.095	0.003	0.009
	Var	0.091	35.523	0.029	0.847	0.087	31.685	0.019	0.149
	MSE	0.091	35.525	0.029	0.849	0.087	31.692	0.019	0.149

the parametric model assumption in (3.8)–(3.9). Panel (b) of Figure 3 displays three estimates of the marginal distribution of the tumor counts x_{ij}, $0.75F(x) + 0.25G(x)$: the marginal empirical distribution, the proposed estimate $0.75\hat{F}(x) + 0.25\hat{G}(x)$ based on the exponential tilting model, and the estimate $0.75\hat{F}_p(x) + 0.25\hat{G}_p(x)$ based on the parametric model. We can see in panel (a) that the two estimates of $G(x)$ are almost the same, while the two estimates of $F(x)$ are apparently different when x is between 30 and 50. In panel (b), we observe that the proposed estimate is almost the same as the marginal empirical distribution of x_{ij}, and both are apparently different from the parametric estimate when x is between 30 and 50. Therefore, the exponential tilting model is more suitable than the parametric model for modeling the tumor counts. We conclude that the proposed estimates $\hat{F}(x)$ and $\hat{G}(x)$ are more reliable than the parametric estimates $\hat{F}_p(x)$ and $\hat{G}_p(x)$.

Further, we apply the ELRT to test $H_0 : \beta = 0$. The test statistic is found to be 57.738, which gives a p-value around 0 calibrated by the limiting distribution. For (μ_F, σ_F^2) and (μ_G, σ_G^2), under the exponential tilting model, we get $(\hat{\mu}_F, \hat{\sigma}_F^2) = (19.58, 128.11)$ and $(\hat{\mu}_G, \hat{\sigma}_G^2) = (4.32, 9.67)$. The above results provide strong evidence

Table 2. Comparing the biases, variances, and mean square errors (MSEs) of the estimators of μ_F and μ_G under the exponential tilting and parametric models under Scenario II, in which the model is correctly specified for $(\hat{\mu}_F, \hat{\sigma}_F^2, \hat{\mu}_G, \hat{\sigma}_G^2)$ but misspecified for $(\hat{\mu}_{F,p}, \hat{\sigma}_{F,p}^2, \hat{\mu}_{G,p}, \hat{\sigma}_{G,p}^2)$.

		Exponential tilting				Parametric model			
(μ_F, μ_G)	Summary	$\hat{\mu}_F$	$\hat{\sigma}_F^2$	$\hat{\mu}_G$	$\hat{\sigma}_G^2$	$\hat{\mu}_{F,p}$	$\hat{\sigma}_{F,p}^2$	$\hat{\mu}_{G,p}$	$\hat{\sigma}_{G,p}^2$
		Scenario II: $N = 20$							
(19.58, 6.96)	Bias	−0.275	−2.504	0.761	6.852	−0.701	−11.059	2.015	20.546
	Var	2.516	577.984	11.033	967.372	3.691	536.907	21.558	2482.113
	MSE	2.592	584.219	11.612	1014.265	4.182	659.167	25.617	2904.099
(19.58, 4.84)	Bias	−0.051	−0.035	0.145	1.036	−0.302	−4.890	0.351	2.673
	Var	1.092	467.747	1.127	67.078	1.168	399.889	2.269	235.196
	MSE	1.094	467.719	1.148	68.148	1.259	423.772	2.392	242.324
(19.58, 3.47)	Bias	−0.037	−0.237	0.088	0.594	−0.274	−6.389	−0.032	−0.384
	Var	0.988	467.540	0.676	33.071	0.931	374.797	0.652	42.580
	MSE	0.989	467.568	0.684	33.422	1.006	415.590	0.653	42.725
		Scenario II: $N = 200$							
(19.58, 6.96)	Bias	−0.006	−0.074	0.014	0.091	−0.222	−6.157	0.646	5.850
	Var	0.103	46.587	0.095	3.350	0.105	34.594	0.121	4.690
	MSE	0.103	46.590	0.095	3.358	0.154	72.499	0.539	38.918
(19.58, 4.84)	Bias	−0.001	−0.051	0.007	0.042	−0.244	−5.253	0.139	0.726
	Var	0.089	45.202	0.051	1.497	0.090	36.733	0.058	1.080
	MSE	0.089	45.202	0.051	1.498	0.149	64.330	0.077	1.607
(19.58, 3.47)	Bias	0.001	−0.042	0.003	0.022	−0.250	−6.506	−0.107	−0.885
	Var	0.084	45.891	0.032	0.803	0.082	35.727	0.029	0.265
	MSE	0.084	45.890	0.032	0.803	0.145	78.055	0.040	1.048

Downloaded by [University of Waterloo] at 14:57 30 October 2017

Table 3. Kolmogorov–Smirnov distance between the estimated cumulative distribution function and true cumulative distribution function under exponential tilting and parametric models.

	Exponential tilting		Parametric model	
(μ_F, μ_G)	$\|\|\hat{F} - F\|\|_\infty$	$\|\|\hat{G} - G\|\|_\infty$	$\|\|\hat{F}_p - F\|\|_\infty$	$\|\|\hat{G}_p - G\|\|_\infty$
		Scenario I: $N = 20$		
(12, 4)	0.071	0.127	0.043	0.079
(18, 4)	0.065	0.107	0.035	0.063
(24, 4)	0.069	0.102	0.034	0.056
		Scenario I: $N = 200$		
(12, 4)	0.019	0.031	0.012	0.021
(18, 4)	0.019	0.031	0.011	0.018
(24, 4)	0.019	0.030	0.010	0.017
		Scenario II: $N = 20$		
(19.58, 6.96)	0.073	0.131	0.078	0.146
(19.58, 4.84)	0.064	0.109	0.064	0.106
(19.58, 3.47)	0.062	0.105	0.060	0.097
		Scenario II: $N = 200$		
(19.58, 6.96)	0.020	0.032	0.043	0.059
(19.58, 4.84)	0.019	0.032	0.043	0.048
(19.58, 3.47)	0.019	0.031	0.042	0.060

that $F(x)$ and $G(x)$ are different, and further $\mu_G < \mu_F$, that is, the mean of the phenotype of mice with the Dd genotype is smaller than that of mice with the DD genotype. This finding is in accordance with scientific knowledge that the particular allele d suppresses the tumor-causing effects (Hoff 2000b). Under the parametric model, we get $(\hat{\mu}_{F,p}, \hat{\sigma}^2_{F,p}) = (18.09, 85.16)$ and $(\hat{\mu}_{G,p}, \hat{\sigma}^2_{G,p}) = (4.23, 7.89)$. According to our simulation experience and Figure 3, the estimates based on the exponential tilting model are more reliable.

Based on the estimates of $F(x)$ and $G(x)$ under the exponential tilting and parametric models, we calculate the posterior probability that a subkindred founder carries the allele d. Panel (a) of Figure 4 compares the posterior probabilities for all 21 subkindred founders under the two models. To have a clearer picture of the lower-left and upper-right corners of Panel (a), we further compare the posterior probabilities for the eight subkindred founders under the lower left corner of Panel (a) in Panel (b) and those for the seven subkindred founders under the upper

Downloaded by [University of Waterloo] at 14:57 30 October 2017

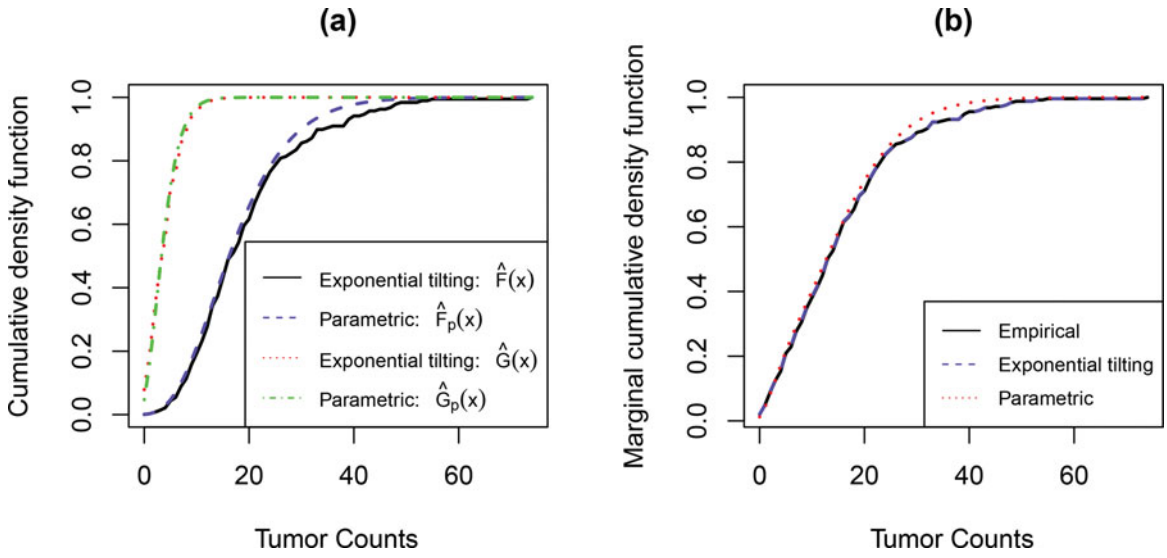

Figure 3. Panel (a) compares $\hat{F}(x)$ and $\hat{F}_p(x)$ in the lower part and $\hat{G}(x)$ and $\hat{G}_p(x)$ in the upper part; panel (b) compares the marginal empirical distribution of x_{ij}, $0.75\hat{F}(x) + 0.25\hat{G}(x)$, with $0.75\hat{F}_p(x) + 0.25\hat{G}_p(x)$.

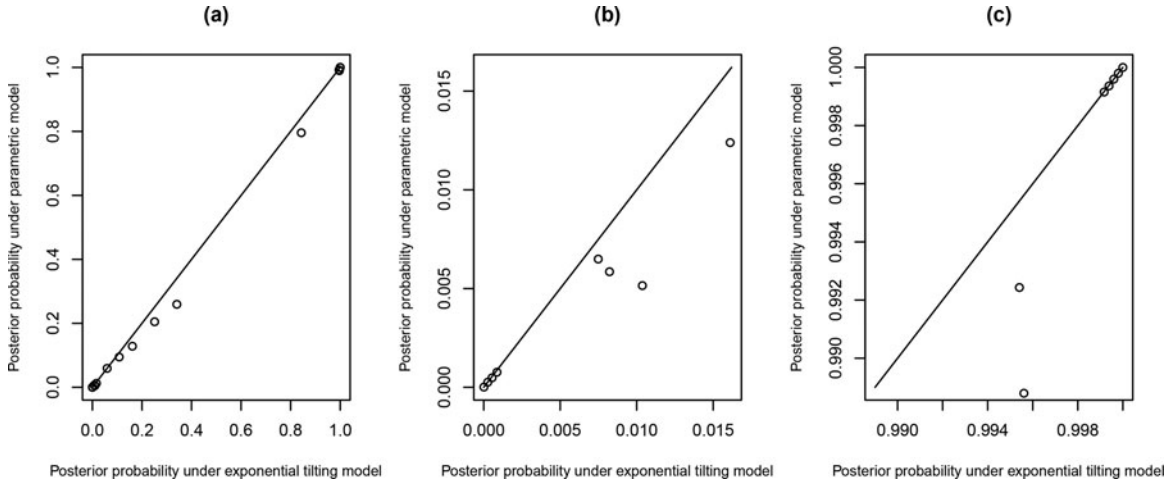

Figure 4. Panel (a) compares the posterior probabilities for all 21 subkindred founders; Panel (b) compares the posterior probabilities for the eight subkindred founders under the lower left corner of Panel (a); Panel (c) compares the posterior probabilities for the seven subkindred founders under the upper right corner of Panel (a).

Downloaded by [University of Waterloo] at 14:57 30 October 2017

right corner of Panel (a) in Panel (c). The posterior probabilities from the exponential tilting model are larger than those from the parametric model. Since the exponential tilting model provides a better fit, we believe that the posterior probabilities from this model are more reliable.

5. Concluding Remarks

In this article, we have explored a semiparametric approach to the genetic mixture model. The theoretical results show that the semiparametric approach has many nice statistical properties that are analogous to those of the full parametric likelihood. Moreover, the numerical results demonstrate that the efficiency loss from the semiparametric mixture model compared to the full parametric model is small if the underlying full parametric model is correctly specified. On the other hand, the semiparametric method is more robust than the full parametric approach. The results of the mice data analysis strongly support our semiparametric approach.

The methods discussed in this article can be generalized to other applications of mixture models. For example, Hoff (2000a) proposed an alternative model for the data presented in the Introduction. This model considers both a *zygotic* effect and a *maternal* effect of allele *d*. A mouse in the NF population experiences the zygotic effect if the animal has the Dd genotype and experiences the maternal effect if the animal's mother has the Dd genotype. This gives rise to four categories of animals with four potentially different tumor count distributions. The tumor count of a mouse in the NF population is distributed according to

- $f_1(x)$ if the animal is subject to both maternal and zygotic effects, that is, the animal and its mother both have the Dd genotype;
- $f_2(x)$ if the animal is subject to the maternal effect only, that is, the animal has the DD genotype and its mother has the Dd genotype;
- $f_3(x)$ if the animal is subject to the zygotic effect only, that is, the animal has the Dd genotype and its mother has the DD genotype;
- $f_4(x)$ if the animal is subject to neither, that is, the animal and its mother both have the DD genotype.

If the ith subkindred founder is female, then conditional on n_i, the joint density of $\mathbf{x}_i$ is

$$0.5\prod_{j=1}^{n_i} f_4(x_{ij}) + 0.5\prod_{j=1}^{n_i}\{0.5f_1(x_{ij}) + 0.5f_2(x_{ij})\}; \qquad (5.11)$$

if the ith subkindred founder is male, then conditional on n_i, the joint distribution of $\mathbf{x}_i$ is

$$0.5\prod_{j=1}^{n_i} f_4(x_{ij}) + 0.5\prod_{j=1}^{n_i}\{0.5f_3(x_{ij}) + 0.5f_4(x_{ij})\}. \qquad (5.12)$$

We can model $f_k(x)/f_4(x)$ for $k = 1, 2, 3$ by the exponential tilting model

$$f_k(x)/f_4(x) = \exp(\alpha_k + \beta_k x).$$

The empirical likelihood method can then be used to estimate (α_k, β_k) and $F_k(x)$, the cumulative distribution function of $f_k(x)$. We plan to investigate the asymptotic properties of the resultant estimators in future research.

Since we have made only the assumption that the two-component density ratio satisfies the exponential tilting model, our method can also be applied to any F2 or recombinant inbred-line experiments as long as there are repeated observations or there are direct observations from each of the component distributions. An example is the setup of Anderson (1979). In general, without training samples or repeated observations, the density ratio is not identifiable.

Appendix: Forms of Σ and $\Omega(s_1, s_2)$

A.1. Form of Σ

Let $\boldsymbol{\theta} = (\alpha, \beta)^\tau$, $\mathbf{y}_{ij} = (1, x_{ij})^\tau$, and $\psi_i(\boldsymbol{\theta}) = \prod_{j=1}^{n_i}\{0.5 + 0.5\exp(\boldsymbol{\theta}^\tau \mathbf{y}_{ij})\}$. We use $\boldsymbol{\theta}_0$ to denote the true value of $\boldsymbol{\theta}$ and $\psi_i = \psi_i(\boldsymbol{\theta}_0)$. We note that the profile empirical log-likelihood in (5) can be written as $l_N(\alpha, \beta) = \inf_\gamma l(\boldsymbol{\theta}, \gamma)$ with

$$l(\boldsymbol{\theta}, \gamma) = \sum_{i=1}^{N}\log\{0.5 + 0.5\psi_i(\boldsymbol{\theta})\} - \sum_{i=1}^{N}\sum_{j=1}^{n_i}\log\{1 + \gamma(e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ij}} - 1)\}.$$

Equivalently, $l_N(\alpha, \beta) = l(\boldsymbol{\theta}, \gamma)$ with γ being the solution to $\partial l(\boldsymbol{\theta}, \gamma)/\partial\gamma = 0$.

The form of $\boldsymbol{\Sigma}$ depends on the second derivatives of $l(\boldsymbol{\theta}, \gamma)$ with respect to $\boldsymbol{\theta}$ and γ, which are given below:

$$\frac{\partial^2 l(\boldsymbol{\theta}, \gamma)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\tau} = \sum_{i=1}^{N}\frac{\psi_i(\boldsymbol{\theta})}{\{1 + \psi_i(\boldsymbol{\theta})\}^2}\sum_{j=1}^{n_i}\frac{e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ij}}}{1 + e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ij}}}\mathbf{y}_{ij}\sum_{k=1}^{n_i}\frac{e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ik}}}{1 + e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ik}}}\mathbf{y}_{ik}^\tau + \sum_{i=1}^{N}\frac{\psi_i(\boldsymbol{\theta})}{1 + \psi_i(\boldsymbol{\theta})}\sum_{j=1}^{n_i}\left\{\frac{e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ij}}}{(1 + e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ij}})^2}\mathbf{y}_{ij}\mathbf{y}_{ij}^\tau\right\} - \sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{\gamma(1-\gamma)e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ij}}}{\{1 + \gamma(e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ij}} - 1)\}^2}\mathbf{y}_{ij}\mathbf{y}_{ij}^\tau,$$

$$\frac{\partial^2 l(\boldsymbol{\theta}, \gamma)}{\partial\gamma^2} = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{\{e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ij}} - 1\}^2}{\{1 + \gamma(e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ij}} - 1)\}^2},$$

$$\frac{\partial^2 l(\boldsymbol{\theta}, \gamma)}{\partial\boldsymbol{\theta}\partial\gamma} = -\sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ij}}}{\{1 + \gamma(e^{\boldsymbol{\theta}^\tau \mathbf{y}_{ij}} - 1)\}^2}\mathbf{y}_{ij}.$$

Then $\boldsymbol{\Sigma}$ is defined to be

$$\boldsymbol{\Sigma} = \left(\mathbf{D}_{12}\mathbf{D}_{22}^{-1}\mathbf{D}_{21} - \mathbf{D}_{11}\right)^{-1}, \qquad (A.1)$$

where $\mathbf{D}_{11} = \mathbb{E}\{\frac{1}{N}\frac{\partial^2 l(\boldsymbol{\theta}_0, \gamma_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\tau}\}$, $\mathbf{D}_{12} = \mathbf{D}_{21}^\tau = \mathbb{E}\{\frac{1}{N}\frac{\partial^2 l(\boldsymbol{\theta}_0, \gamma_0)}{\partial\boldsymbol{\theta}\partial\gamma}\}$, $\mathbf{D}_{22} = \mathbb{E}\{\frac{1}{N}\frac{\partial^2 l(\boldsymbol{\theta}_0, \gamma_0)}{\partial\gamma^2}\}$, and $\gamma_0 = 0.25$. The meaning of γ_0 is discussed in Section 3 of the supplementary document.

A.2. Form of $\Omega(\mathbf{s}_1, \mathbf{s}_2)$

Let $w(x_{ij}) = 0.5 + 0.5e^{\boldsymbol{\theta}_0^\tau \mathbf{y}_{ij}}$ and

$$\Delta_{N,F}(x) = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{I(x_{ij} \le x)}{N\mathbb{E}(n_1)} \frac{1}{0.5 + 0.5w(x_{ij})} - F(x) \right\}$$

$$- C_{1,x}(\hat{\gamma} - \gamma_0) - \mathbf{C}_{2,x}^\tau(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

with $C_{1,x} = \mathbb{E}[I(x_{ij} \le x)(e^{\boldsymbol{\theta}_0^\tau \mathbf{y}_{ij}} - 1)\{0.5 + 0.5w(x_{ij})\}^{-2}]$ and $\mathbf{C}_{2,x} = \mathbb{E}[I(x_{ij} \le x)\gamma_0 e^{\boldsymbol{\theta}_0^\tau \mathbf{y}_{ij}}\{0.5 + 0.5w(x_{ij})\}^{-2}\mathbf{y}_{ij}]$. Further, let

$$\Delta_{N,G}(x) = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{I(x_{ij} \le x)}{N\mathbb{E}(n_1)} \frac{e^{\boldsymbol{\theta}_0^\tau \mathbf{y}_{ij}}}{0.5 + 0.5w(x_{ij})} - G(x) \right\}$$

$$- C_{3,x}(\hat{\gamma} - \gamma_0) - \mathbf{C}_{4,x}^\tau(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

with $C_{3,x} = \mathbb{E}[I(x_{ij} \le x)e^{\boldsymbol{\theta}_0^\tau \mathbf{y}_{ij}}(e^{\boldsymbol{\theta}_0^\tau \mathbf{y}_{ij}} - 1)\{0.5 + 0.5w(x_{ij})\}^{-2}]$ and $\mathbf{C}_{4,x} = \mathbb{E}[I(x_{ij} \le x)(1 - \gamma_0)e^{\boldsymbol{\theta}_0^\tau \mathbf{y}_{ij}}\{0.5 + 0.5w(x_{ij})\}^{-2}\mathbf{y}_{ij}]$.

It is shown in the supplementary material that

$$\hat{F}(x) - F(x) = \Delta_{N,F}(x) + O_p(N^{-1}),$$

$$\hat{G}(x) - G(x) = \Delta_{N,G}(x) + O_p(N^{-1}).$$

Further, let $\boldsymbol{\Delta}_N(\mathbf{s}) = (\Delta_{N,F}(x), \Delta_{N,G}(y))^\tau$ with $\mathbf{s} = (x, y)^\tau$. Then $\boldsymbol{\Omega}(\mathbf{s}_1, \mathbf{s}_2)$ is defined as

$$\boldsymbol{\Omega}(\mathbf{s}_1, \mathbf{s}_2) = \lim_{N \to \infty} N\mathbb{E}\left\{\boldsymbol{\Delta}_N(\mathbf{s}_1)\boldsymbol{\Delta}_N^\tau(\mathbf{s}_2)\right\}. \quad \text{(A.2)}$$

Supplementary Materials

The online web appendix contains more simulation studies, more details for the EM-algorithm in Section 2.2, and detailed proofs of Theorems 1–3.

Acknowledgments

The authors thank the joint editors, the associate editor, and the two referees for constructive comments and suggestions that led to a significant improvement. The first two authors contributed equally to this work. Liu is the corresponding author.

Funding

Dr. Li's research is supported in part by NSERC Grant RGPIN-2015-06592. Dr. Liu's research is supported by grants from the National Natural Science Foundation of China (Numbers 11371142, 11171112, 11101156, 11501208, and 11501354), the Program of Shanghai Subject Chief Scientist (14XD1401600), and the 111 Project (B14019).

References

Anderson, J. A. (1979), "Multivariate Logistic Compounds," *Biometrika*, 66, 17–26. [1251,1259]

Carvalho, M. D., and Davison, A. C. (2014), "Spectral Density Ratio Models for Multivariate Extremes," *Journal of the American Statistical Association*, 109, 764–776. [1251]

Chen, J., and Liu, Y. (2013), "Quantile and Quantile-Function Estimations Under Density Ratio Model," *The Annals of Statistics*, 41, 1669–1692. [1251]

Cox, D. R. (1969), "Some Sampling Problems in Technology," in *New Developments in Survey Sampling*, eds. N. L. Johnson and H. Smith, New York: Wiley-Interscience, pp. 506–527. [1251]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Series B, 39, 1–38. [1253]

Drinkwater, N. R., and Klotz, J. H. (1981), "Statistical Methods for the Analysis of Tumor Multiplicity Data," *Cancer Research*, 41, 113– 119. [1255]

Efron, B. (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (*Institute of Mathematical Statistics Monographs*, Vol. I), Cambridge, UK: Cambridge University Press. [1250]

Hoff, P. D. (2000a), "Constrained Nonparametric Maximum Likelihood via Mixtures," *Journal of Computational and Graphical Statistics*, 9, 633–641. [1250,1251,1259]

——— (2000b), "Constrained Nonparametric Estimation via Mixtures," Ph.D. dissertation, University of Wisconsin-Madison, Madison, WI. [1250,1251,1255,1256,1258]

Hoff, P. D., Halberg, R. B., Shedlovsky, A., Dove, W. F., and Newton, M. A. (2002), "Identifying Carriers of a Genetic Modifier Using Nonparametric Bayes Methods," *Case Studies in Bayesian Statistics*, 5, 327–342. [1250,1251,1255]

Kay, R., and Little, S. (1987), "Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data," *Biometrika*, 74, 495–501. [1251]

Liu, T., Hogan, J. W., Wang, L., Zhang, S., and Kantor, R. (2013), "Optimal Allocation of Gold Standard Testing Under Constrained Availability: Application to Assessment of HIV Treatment Failure," *Journal of the American Statistical Association*, 108, 1173–1188. [1251]

Ott, J. (1999), *Analysis of Human Genetic Linkage* (3rd ed.), Baltimore, MD: The Johns Hopkins University Press. [1250]

Owen, A. B. (2001), *Empirical Likelihood*, New York: Chapman & Hall/CRC. [1251]

Patil, G. P., and Rao, C. R. (1978), "Weighted Distributions and Size-Based Sampling with Applications to Wildlife Populations and Human Families," *Biometrics*, 34, 179–189. [1251]

Qin, J. (1999), "Empirical Likelihood Ratio Based Confidence Intervals for Mixture Proportions," *Annals of Statistics*, 27, 1368– 1384. [1251]

Sham, P. (1998), *Statistics in Human Genetics*, New York: Arnold. [1250,1251]

Tan, Z. (2009), "On Profile Likelihood for Exponential Tilt Mixture Models," *Biometrika*, 96, 229–236. [1252,1254]

Vardi, Y. (1982), "Nonparametric Estimation in the Presence of Length Bias," *Annals of Statistics*, 10, 616–620. [1251]

——— (1985), "Empirical Distribution in Selection Bias Models," *Annals of Statistics*, 13, 178–203. [1251]

Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103. [1253]

Zhang, B. (2002), "An EM Algorithm for a Semiparametric Finite Mixture Model," *Journal of Statistical Computation and Simulation*, 72, 791–802. [1253]

Zou, F., Fine, J. P., and Yandell, B. S. (2002), "On Empirical Likelihood for a Semiparametric Mixture Model," *Biometrika*, 89, 61–75. [1250,1251,1254]

Downloaded by [University of Waterloo] at 14:57 30 October 2017