Estimation and Classification in a Two-component Mixture Model with Stochastic Dominance

Jingjing Wu[1], Tasnima Abedin[2] and Dan Nettleton[3]

[1]University of Calgary
[2]Alberta Health Services
[3]Iowa State University

IMS, National University of Singapore

February 13, 2018

Outline

Identifying differentially expressed genes

Consider the problem of identifying differentially expressed genes over two or more conditions (e.g. healthy vs. diseased). For this purpose, a hypothesis test is carried out for each gene. Under the null hypothesis, corresponding to a lack of difference in gene expression level, the test statistic follows a specified distribution F (e.g. normal, t). However under the alternative, corresponding to the presence of difference in gene expression level, the distribution G of the test statistic is unknown.

Treating each test statistic value as a response from the corresponding gene, the thousands of responses of all genes come from a mixture the known F and the unknown G, with some unknown mixing proportion λ.

Once λ and G have been estimated, by Bayes' rule one can estimate the probability that a gene is differentially expressed, i.e. belongs to G. Thus, using a classification criterion we can classify each gene as either differentially expressed or not with estimated misclassification probability.

Based on all the identified marker genes together, one could build a classification rule to classify each subject as, say, healthy or diseased.

More generally, F might be also unknown in practice but a training sample from F is immediately available. This means that, in the gene example, particular genes have been confidently identified by pathologists or experts as not related to patients' health condition. This more general setup incorporates the cases when F for small sample size is not exactly the same as the postulated one by asymptotics (large sample size).

The stochastic dominance of F over G arise naturally when one believes that the test statistic values for, say, marker genes tend to be larger or smaller than those for non-marker genes. For example, the most often used Student's t (strictly $|t|$) and ANOVA F statistic satisfy the stochastic ordering $F \geq G$.

Malaria example

Clinical malaria can be diagnosed by the presence of parasites and fever. However in endemic areas children can tolerate malaria parasite without the development of any sign of disease, and they may have fever due to some other reason. We can consider a mixture model where the mixture consists of parasite densities in children with fever due to malaria or due to other causes. Parasite levels in children from a community (without malaria) could be available, i.e. a sample that comes from the component of the mixture corresponding to children without clinical malaria but have parasites in their body and hence fever. Here the mixing proportion is the proportion of children whose fever is attributable to malaria.

Two-component mixture model with stochastic dominance

Suppose we have two independent samples

$$\begin{aligned} X_1, \ldots, X_m &\overset{\text{i.i.d.}}{\sim} f(x), \\ Y_1, \ldots, Y_n &\overset{\text{i.i.d.}}{\sim} h(x) = (1-\lambda) f(x) + \lambda g(x), \quad x \in \mathbb{R}, \end{aligned} \tag{1.1}$$

where the unknown mixing proportion $\lambda \in (0, 1)$, and f and g are two unknown p.d.f.s with stochastic dominance $F \geq G$.

Let $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\lambda)$ with $Z_i = 1$ if Y_i comes from G and $Z_i = 0$ if Y_i comes from F. By Bayes' rule, given an observation $Y = y$ the probability of it being from G is

$$p(y) := P(Z = 1 | Y = y) = \frac{\lambda g(y)}{(1-\lambda) f(y) + \lambda g(y)}. \tag{1.2}$$

In the marker gene example, λ and $p(y)$ correspond to the proportion of marker genes and the chance of being a marker gene given a test statistic value y respectively.

In biomedical assay case, they correspond to the proportion of patients with particular disease and the chance of having this disease given an assay index value y.

Literature review

To our best knowledge, there is not even a single work on model (1.1) in the literature. The closest work that has been done related to this model is given in Smith and Vounatsou (1997). However their model did not take the stochastic dominance constraint but instead assumed that the probability function in (1.2) is monotonically increasing.

The monotone assumption on p in these works is stronger than the stochastic dominance. To see this, note that function p being monotonic increasing generally implies that $F \geq G$. But the implication of the other direction is not true. A counter-example is $f(x) = 0.5 I_{[0,2]}(x)$ and $g(x) = (1 - 0.25x) I_{[1,3]}(x)$. Therefore, our model (1.1) is more general.

Model identifiability

Let D_f and D_g denote the lower limit of the support of function f and g respectively. Note that D_f and D_g could possibly be $-\infty$. Since $F \geq G$, we have $D_f \leq D_g$ and thus $g(x)/f(x)$ is well defined and $g(x)/f(x) \leq 1$ as $x \to D_f^+$.

Theorem 2.1. **Assume that $p(x) \to 0$ or equivalently $g(x)/f(x) \to 0$ as $x \to D_f^+$ and m is sufficiently large. Then the mixture model (1.1) is identifiable.**

Remark: The sufficient condition in Theorem 2.1 is quite weak and also easy to check. If $D_f \neq D_g$, then the condition holds.

A nonparametric estimation based on c.d.f.s

Note that for any $\alpha \in (0,1)$,

$$1 - H(F^{-1}(\alpha))/\alpha \;=\; \lambda\left[1 - \frac{G(F^{-1}(\alpha))}{\alpha}\right].$$

Since $F \geq G$, we have $\frac{G(F^{-1}(\alpha))}{\alpha} \leq 1$ and then

$$1 - H(F^{-1}(\alpha))/\alpha \leq \lambda.$$

Thus a lower bound estimate of λ is given by

$$\hat{\lambda}_\alpha \;=\; 1 - \frac{H_n(F_m^{-1}(\alpha))}{\alpha},$$

where F_m and H_n are some appropriate nonparametric estimators of F and H respectively based on the samples X_i's and Y_i's.

If the discrepancy between F and G is large enough, then $\frac{G(F^{-1}(\alpha))}{\alpha}$ may be near zero at some α value and as a result $1 - H(F^{-1}(\alpha))/\alpha$ will be close to λ. Especially when the sufficient condition for identifiability in Theorem 2.1, i.e. $\lim_{x \to D_f^+} g(x)/f(x) = 0$, holds, $\frac{G(F^{-1}(\alpha))}{\alpha}$ will be very close to zero for small α values. Thus, intuitively, we propose an estimator of λ given by

$$\hat{\lambda} = \sup_{\alpha \in (0,1)} \hat{\lambda}_\alpha = 1 - \inf_{\alpha \in (0,1)} \frac{H_n(F_m^{-1}(\alpha))}{\alpha}. \tag{2.1}$$

The function $p(y)$ in (1.2) now can be estimated by

$$\hat{p}(y) = 1 - (1 - \hat{\lambda}) \frac{f_m(y)}{h_n(y)}, \tag{2.2}$$

where f_m and h_n are the corresponding p.d.f.s of F_m and H_n.

In this work, we use kernel density estimators

$$f_m(x) = \frac{1}{mb_m}\sum_{i=1}^{m} K_0\left(\frac{x - X_i}{b_m}\right), \tag{2.3}$$

$$h_n(x) = \frac{1}{nb_n}\sum_{j=1}^{n} K_1\left(\frac{x - Y_j}{b_n}\right), \tag{2.4}$$

where K_0 and K_1 are kernel p.d.f.s and bandwidths b_n and b_m are positive sequences such that $b_m \to 0$ as $m \to \infty$ and $b_n \to 0$ as $n \to \infty$.

Asymptotic bias

Theorem 2.2. **Suppose that both f and g are uniformly continuous and the bandwidths b_m and b_n make $\sum_{m=1}^{\infty} \exp(-rmb_m^2)$ and $\sum_{n=1}^{\infty} \exp(-rnb_n^2)$ converge for every $r > 0$. Then as $m, n \to \infty$,**

$$\hat{\lambda}_\alpha \xrightarrow{a.e.} \lambda - \lambda \frac{G(F^{-1}(\alpha))}{\alpha} \tag{2.5}$$

for any $\alpha \in (0, 1)$, and as a result

$$\hat{\lambda} \xrightarrow{a.e.} \lambda - \lambda \inf_{\alpha \in (0,1)} \frac{G(F^{-1}(\alpha))}{\alpha}. \tag{2.6}$$

Consistency

Corollary 2.1. Suppose the conditions in Theorem 2.2 are satisfied and in addition $g(y)/f(y) \to 0$ as $y \to D_f^+$. Then $\hat{\lambda} \xrightarrow{a.e.} \lambda$ as $m, n \to \infty$.

Remark: By Theorem 2.2, the estimator $\hat{\lambda}$ defined in (2.1) is generally biased. However when the sufficient condition for identifiability given in Theorem 2.1 is satisfied, then model (1.1) is identifiable and at the same time the estimator $\hat{\lambda}$ is consistent by Corollary 2.1.

Simulation study

We consider five mixture models that all satisfy the stochastic dominance condition. Even though our focus is on continuous mixture models, we also want to check the performance of the proposed methods for discrete cases such as M3 and M4.

Table 1: Mixture models considered in simulation study.

M1	$(1-\lambda)N(0,1)+\lambda N(1,1)$	mixture of normals that are close
M2	$(1-\lambda)N(0,1)+\lambda N(5,1)$	mixture of normals that are apart
M3	$(1-\lambda)Po(2)+\lambda Po(4)$	mixture of Poissons that are close
M4	$(1-\lambda)Po(2)+\lambda Po(6)$	mixture of Poissons that are apart
M5	$(1-\lambda)U(0,4)+\lambda U(2,6)$	mixture of uniforms

We consider varying $\lambda = 0.05, 0.20, 0.50, 0.80, 0.95$. We use two different sets of sample sizes $(m, n) = (30, 30)$ and $(100, 100)$, and replication number $N = 1000$. In the kernel density estimators f_m and h_n, we use Gaussian function for both kernels K_0 and K_1, and the same bandwidths as in Silverman's (1986)

$$\begin{aligned} b_m &= 0.9m^{-1/5} \min\left[SD_X, \frac{IQR_X}{1.34}\right], \\ b_n &= 0.9n^{-1/5} \min\left[SD_Y, \frac{IQR_Y}{1.34}\right]. \end{aligned} \tag{2.7}$$

We estimate the bias and MSE by

$$Bias(\hat{\lambda}) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\lambda}_i - \lambda\right), \tag{2.8}$$

$$MSE(\hat{\lambda}) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\lambda}_i - \lambda\right)^2. \tag{2.9}$$

To examine the performance of $\hat{p}(y)$, we check the classification results of a simple classification rule based on $\hat{p}$. Here the hard threshhold of 0.5 is used as the classification rule, i.e. an individual with observation y is classified as from G if $\hat{p}(y) > 0.5$ and F if otherwise. Then we use the misclassification rate (MR) as a measure of the performance of $\hat{p}$.

However, we can expect that the MR will be high for some models, such as M1, when the two components are close. Thus we use the *optimal misclassification rate* (OMR) as the baseline to compare with. The OMR is the MR calculated when p is assumed completely specified (the best scenario for this classification rule), i.e.

$$\mathrm{OMR} = (1-\lambda) \int_{\{y:\lambda g(y)>(1-\lambda)f(y)\}} f(y)dy + \lambda \int_{\{y:\lambda g(y)<(1-\lambda)f(y)\}} g(y)dy. \tag{2.10}$$

Table 2: Bias and MSE of $\hat{\lambda}$ and MR (%) of a classification rule based on $\hat{p}$.

Model	λ	$m = n = 30$ $Bias(\hat{\lambda})(MSE(\hat{\lambda}))$	MR	$m = n = 100$ $Bias(\hat{\lambda})(MSE(\hat{\lambda}))$	MR	OMR
M1	0.05	0.052 (0.031)	8.03	0.049 (0.019)	7.37	4.99
	0.20	0.093 (0.074)	26.30	0.085 (0.052)	26.09	18.61
	0.50	0.067 (0.078)	36.63	0.067 (0.052)	37.05	30.85
	0.80	-0.009 (0.047)	22.83	0.008 (0.026)	21.29	18.61
	0.95	-0.052 (0.029)	9.07	-0.027 (0.012)	6.80	4.99
M2	0.05	0.053 (0.022)	2.20	0.048 (0.015)	2.28	0.24
	0.20	0.095 (0.032)	3.97	0.065 (0.018)	3.02	0.48
	0.50	0.082 (0.015)	2.73	0.049 (0.005)	2.12	0.62
	0.80	0.061 (0.014)	6.93	0.057 (0.009)	6.21	0.48
	0.95	0.034 (0.002)	7.40	0.036 (0.002)	4.16	0.24
M3	0.05	0.203 (0.089)	28.40	0.035 (0.005)	5.04	4.76
	0.20	0.133 (0.064)	32.27	0.034 (0.013)	16.69	13.90
	0.50	0.019 (0.039)	34.13	-0.015 (0.001)	15.54	19.05
	0.80	-0.096 (0.034)	24.17	-0.084 (0.190)	25.20	13.35
	0.95	-0.152 (0.042)	14.40	-0.119 (0.023)	14.24	6.06
M4	0.05	0.211 (0.092)	26.30	0.044 (0.006)	3.03	3.14
	0.20	0.178 (0.071)	25.27	0.071 (0.012)	9.61	7.03
	0.50	0.117 (0.034)	21.60	0.048 (0.009)	15.54	10.19
	0.80	0.019 (0.010)	13.77	0.009 (0.006)	13.34	7.82
	0.95	-0.026 (0.004)	5.73	-0.006 (0.002)	5.81	3.21
M5	0.05	0.249 (0.136)	30.83	0.018 (0.002)	3.09	2.50
	0.20	0.151 (0.078)	29.90	0.017 (0.005)	10.99	10.00
	0.50	0.067 (0.048)	30.43	0.016 (0.008)	24.24	25.00
	0.80	0.031 (0.022)	16.20	0.012 (0.006)	12.76	10.00
	0.95	0.009 (0.006)	4.33	0.012 (0.002)	3.87	2.50

- When sample sizes increase, the performance of both $\hat{\lambda}$ in terms of bias and MSE and $\hat{p}$ in terms of MR improves.
- For $m = n = 100$, we are able to classify individuals equivalently well when $\hat{p}$ is used as when p is used.
- The performance under models M1 and M2 is about the same for $m = n = 30$ and $m = n = 100$.
- The relatively worse performance for models M3-M5 when $m = n = 30$ could be possibly explained by the fact that these models don't satisfy the conditions listed in Corollary 2.1 for $\hat{\lambda}$ being consistent. Nevertheless, when sample sizes become large, the proposed estimators work well for models M3-M5.

MLE of multinomial approximation

Partition the support of h into L regions $R_1, \ldots, R_L$. Define

$$\begin{aligned} p_{1l} &= \int_{R_l} f(x)dx, \\ p_{2l} &= \int_{R_l} g(x)dx, \\ p_{3l} &= \int_{R_l} h(x)dx = (1-\lambda)p_{1l} + \lambda p_{2l}. \end{aligned} \tag{3.1}$$

Let m_l and n_l denote the number of observations out of m and n respectively that fall into region R_l. Then when L is large, model (1.1) could be approximated closely by the multinomial populations given in (3.1), based on which the likelihood is given by

$$\prod_{l=1}^{L} (p_{1l})^{m_l} \left[(1-\lambda)p_{1l} + \lambda p_{2l}\right]^{n_l}. \tag{3.2}$$

Let $\hat{\theta} = (\hat{\lambda}_L, \hat{p}_{11}, \ldots, \hat{p}_{1L}, \hat{p}_{21}, \ldots, \hat{p}_{2L})^\top$ denote the MLE. Then

$$\hat{p}_{1l} = \frac{m_l}{m}, \quad l = 1, \ldots, L, \tag{3.3}$$

$$\hat{p}_{2l} = \frac{1}{\lambda} \cdot \frac{n_l}{n} - \frac{1-\lambda}{\lambda} \cdot \frac{m_l}{m}, \quad l = 1, \ldots, L, \tag{3.4}$$

while the MLE of λ generally does not exist. However, we have from (3.4) that $\lambda \geq 1 - \frac{n_l}{n} \cdot \frac{m}{m_l}$ for each l such that $m_l \neq 0$. Then the MLE of the lower bound of λ is given by

$$\hat{\lambda}_L = 1 - \min_{\substack{l=1,\ldots,L \\ m_l \neq 0}} \left\{ \frac{n_l/n}{m_l/m} \right\}. \tag{3.5}$$

The MLE of $p(y)$ is given by

$$\hat{p}_L(y) = \frac{\hat{\lambda}_L \hat{p}_{2i}}{(1-\hat{\lambda}_L)\hat{p}_{1i} + \hat{\lambda}_L \hat{p}_{2i}}, \text{ if } y \in R_i.$$

Consistency

Theorem 3.1. $\hat{\lambda}_L \xrightarrow{\mathcal{P}} \lambda - \lambda \min_{\substack{l=1,\ldots,L \\ p_{1l}\neq 0}} \left\{ \frac{p_{2l}}{p_{1l}} \right\}$ as $m, n \to \infty$.

Corollary 3.1. If $p_{2l} = 0$ for some l such that $p_{1l} \neq 0$, then $\hat{\lambda}_L \xrightarrow{\mathcal{P}} \lambda$ as $m, n \to \infty$.

Corollary 3.2. Let the maximum length of the intervals $R_1, \ldots, R_L$ go to zero when $L \to \infty$. If $g(x)/f(x) \to 0$ as $x \to D_f^+$, then $\hat{\lambda}_L \xrightarrow{\mathcal{P}} \lambda$ as $m, n, L \to \infty$.

Simulation study

- We consider the same mixture models listed in Table 2.1. We chose $L = 3$ for $m = n = 30$ and $L = 10$ for $m = n = 100$.
- We examine the MLE of λ under the stochastic dominance $\sum_{i=1}^{k} p_{1i} \geq \sum_{i=1}^{k} p_{2i}$, $k = 1, \ldots, L$. Even though the MLE of λ without this constraint does not exist theoretically, our simulation results below show that the MLE with this constraint does exist.
- For simplicity, we use λ_+ in Smith and Vounatsou (1997) as the initial estimate of λ, and m_l/m as the initial of p_{1l}. To give an initial of p_{2l}, we use the relationship $g(x) = \frac{h(x)}{\lambda} - \frac{1-\lambda}{\lambda} f(x)$ with λ replaced with its initial and f and h replaced with m_l/m and n_l/n.

Table 3: Bias and MSE of $\hat{\lambda}_L$ and MR (%) of a classification rule based on $\hat{p}_L$.

Model	λ	$m = n = 30$ $Bias(\hat{\lambda}_L)(MSE(\hat{\lambda}_L))$	MR	$m = n = 100$ $Bias(\hat{\lambda}_L)(MSE(\hat{\lambda}_L))$	MR	OMR
M1	0.05	0.020 (0.031)	5.60	0.022 (0.009)	5.00	4.99
	0.20	0.032 (0.068)	22.13	0.011 (0.016)	20.16	18.61
	0.50	-0.039 (0.078)	46.77	-0.048 (0.026)	44.37	30.85
	0.80	-0.144 (0.089)	41.57	-0.121 (0.033)	27.39	18.61
	0.95	-0.181 (0.090)	25.17	-0.149 (0.035)	11.73	4.99
M2	0.05	0.022 (0.031)	4.97	0.026 (0.009)	5.00	0.24
	0.20	0.055 (0.071)	20.57	0.033 (0.016)	19.90	0.48
	0.50	0.029 (0.066)	39.13	0.017 (0.019)	25.61	0.62
	0.80	-0.009 (0.036)	19.97	0.000 (0.010)	4.56	0.48
	0.95	-0.011 (0.009)	4.53	-0.002 (0.003)	2.21	0.24
M3	0.05	0.002 (0.015)	5.40	-0.000 (0.004)	5.01	4.76
	0.20	-0.017 (0.046)	22.03	-0.022 (0.013)	20.09	13.90
	0.50	-0.076 (0.010)	43.67	-0.073 (0.029)	44.03	19.05
	0.80	-0.128 (0.101)	35.53	-0.121 (0.038)	29.74	13.35
	0.95	-0.163 (0.102)	22.57	-0.143 (0.034)	14.27	6.06
M4	0.05	0.003 (0.016)	5.43	0.000 (0.004)	5.01	3.14
	0.20	-0.007 (0.047)	21.60	-0.000 (0.012)	20.08	7.03
	0.50	-0.028 (0.084)	37.97	-0.009 (0.021)	35.49	10.19
	0.80	-0.033 (0.059)	24.20	-0.018 (0.013)	14.68	7.82
	0.95	-0.048 (0.037)	10.03	-0.023 (0.005)	5.41	3.21
M5	0.05	0.026 (0.032)	22.07	0.033 (0.009)	5.01	2.50
	0.20	0.049 (0.069)	20.80	0.035 (0.014)	20.13	10.00
	0.50	0.025 (0.065)	45.10	0.011 (0.019)	37.34	25.00
	0.80	-0.001 (0.029)	23.07	-0.001 (0.008)	13.38	10.00
	0.95	-0.012 (0.010)	5.70	-0.005 (0.002)	3.62	2.50

- $\hat{\lambda}_L$ performs very well in terms of bias and MSE in most cases.
- The estimation accuracy is higher when the two components are well separated (M2 & M4) than when they are not (M1 & M3).
- Even though we group the data and thus lose some information, we can still estimate λ quite well.
- However, the MRs are much higher than the OMRs in most cases. This is expected since with use of discretization, all the observations falling into the same interval will be classified as from the same component. When the interval is relatively wide, for example $L = 3$ or even $L = 10$, the discretization will generate a higher misclassification rate.

- When $\hat{\lambda}_L$ is compared with $\hat{\lambda}$, we observe that both perform competitively in terms of bias and MSE while $\hat{\lambda}$ has better performance in terms of MR than $\hat{\lambda}_L$.
- Therefore, if our interest is in λ only, then either method should work well. But if we are interested in classification, then the $\hat{p}$ based on $\hat{\lambda}$ works much better.

Introduction of a semiparametric mixture model

Let Z denote a binary response variable and Y the associated covariate. Then the logistic regression model is given by

$$P(Z=1|Y=y) \;=\; \frac{\exp[\alpha^* + \beta^\top r(y)]}{1+\exp[\alpha^* + \beta^\top r(y)]}, \tag{4.1}$$

where $r(y) = (r_1(y), \ldots, r_p(y))^\top$ is a given vector of functions.

Let $\pi = P(Z=1)$. Let $f(y)$ and $g(y)$ denote the conditional p.d.f.s of Y given $Z=0$ and $Z=1$ respectively, then (4.1) gives

$$g(y) \;=\; \exp[\alpha + \beta^\top r(y)] f(y), \tag{4.2}$$

where $\alpha = \alpha^* + \log[(1-\pi)/\pi]$. Now model (1.1) is reduced to

$$\begin{aligned} X_1, \ldots, X_m &\overset{\text{i.i.d.}}{\sim} f(x), \\ Y_1, \ldots, Y_n &\overset{\text{i.i.d.}}{\sim} h_\theta(x) = \left\{(1-\lambda) + \lambda \exp\left[\alpha + \beta^\top r(x)\right]\right\} f(x). \end{aligned} \tag{4.3}$$

- The relationship (4.2) was first proposed by Anderson (1972). It essentially assumes that the log-likelihood ratio of the two p.d.f.s is linear in the observations.
- With $r(x) = x$ or $(x, x^2)^\top$, it has wide applications in logistic discriminant analysis (Anderson, 1972&1979) and case-control studies (Prentice and Pyke, 1979; Breslow and Day, 1980).
- For $r(x) = x$, it encompasses many common distributions, including two exponentials with different means and two normals with common variance but different means.
- With $r(x) = (x, x^2)^\top$, it coincides with the exponential family of densities considered in Efron and Tibshirani (1996).
- Moreover, (4.2) can be viewed as a biased sampling model with the 'tilt' weight function $\exp[\alpha + \beta^\top r(x)]$.
- The test of equality of f and g can be regarded as a special case of (4.3) with $\beta = 0$.

Theorem 4.1. **Model (4.3) with $r(y) = y$ is identifiable. If further $\beta > 0$ and m is sufficiently large, then $F \geq G$.**

Even though this theorem tells us that the condition (4.2) is stronger than the original stochastic dominance constraint, the thus resulted semiparametric mixture model (4.3) is identifiable and has better interpretation than the nonparametric mixture model (1.1). In addition, the estimation of (4.3) may possess better asymptotic properties, such as normality, than those of (1.1). So from now on, we will focus on model (4.3) with $r(y) = y$ and $\beta > 0$.

Empirical MLE

Let $(T_1, ..., T_{m+n}) = (X_1, ..., X_m, Y_1, ..., Y_n)$ be the pooled data and $p_i = dF(T_i)$. Then the empirical likelihood function is

$$L(\lambda, \alpha, \beta) = \prod_{i=1}^{m} dF(X_i) \prod_{j=1}^{n} dH(Y_j) = \prod_{i=1}^{m+n} p_i \prod_{j=1}^{n} \left[(1-\lambda) + \lambda e^{\alpha+\beta Y_j}\right],$$

subject to $\beta \geq 0$, $0 \leq \lambda \leq 1$, $p_i \geq 0$, $\sum_{i=1}^{m+n} p_i = 1$, $\sum_{i=1}^{m+n} p_i e^{\alpha+\beta T_i} = 1$.

With $N = m + n$ and $\rho_N = n/(m+n)$, we have

$$p_i = \frac{1}{N\left[1 + \rho_N \lambda \left(e^{\alpha+\beta T_i} - 1\right)\right]}.$$

Thus the log-likelihood function is

$$l(\lambda,\alpha,\beta) \propto \sum_{j=1}^{n} \log[(1-\lambda)+\lambda e^{\alpha+\beta Y_j}] - \sum_{i=1}^{m+n} \log[1+\rho_N \lambda (e^{\alpha+\beta T_i}-1)]. \tag{4.4}$$

Let $\hat{\theta}_{MLE} = (\hat{\lambda}_{MLE}, \hat{\alpha}_{MLE}, \hat{\beta}_{MLE})^\top$ and $\hat{p}_{MLE}(y)$ denote the MLE of θ and $p(y)$ respectively.

Asymptotic normaility

Define

$$S = \int \left(\frac{\partial w_1(y)}{\partial \theta}\right)\left(\frac{\partial w_1(y)}{\partial \theta}\right)^{\top} \frac{f}{w_1 w_2}(y)dy, \tag{4.5}$$

$$V = S - \int \frac{\partial w_1(y)}{\partial \theta}\frac{f}{w_2}(y)dy \int \left(\frac{\partial w_1(y)}{\partial \theta}\right)^{\top} \frac{f}{w_2}(y)dy, \tag{4.6}$$

where

$$w_1(y) = 1 - \lambda + \lambda e^{\alpha+\beta y}, \tag{4.7}$$

$$w_2(y) = 1 - \rho\lambda + \rho\lambda e^{\alpha+\beta y}. \tag{4.8}$$

Theorem 4.2. Assume $\rho_N \to \rho$ as $N \to \infty$. Then under some regularity conditions (for MLE in general),

$$\sqrt{N}\begin{pmatrix} \hat{\lambda}_{MLE} - \lambda \\ \hat{\alpha}_{MLE} - \alpha \\ \hat{\beta}_{MLE} - \beta \end{pmatrix} \xrightarrow{\mathcal{L}} N(0, \Sigma),$$

where $\Sigma = \frac{1}{\rho(1-\rho)} S^{-1} V S^{-1}$ with S and V defined in (4.5) and (4.6) respectively.

Simulation study

We consider the same five mixture models and for each the true values of α and β are derived under (4.2). As expected, $\beta > 0$. Note that model M5 does not satisfy (4.2).

Table 4: Mixture models considered in simulation study.

Model	Form	α	β
M1	$(1-\lambda)N(0,1)+\lambda N(1,1)$	-0.5	1
M2	$(1-\lambda)N(0,1)+\lambda N(5,1)$	-12.5	5
M3	$(1-\lambda)Po(2)+\lambda Po(4)$	-2	0.693
M4	$(1-\lambda)Po(2)+\lambda Po(6)$	-4	1.099
M5	$(1-\lambda)U(0,4)+\lambda U(2,6)$	NA	NA

- We use λ_+ as the initial of λ. Initial values of α and β are calculated by exploiting the relationship (4.3). Specifically, (4.3) indicates

$$\log \frac{h(x)/f(x) - (1-\lambda)}{\lambda} = \alpha + \beta x.$$

  Thus for each T_i in the pooled sample, we generate (T_i, R_i), where $R_i = \log \frac{h_n(T_i)/f_m(T_i) - (1-\lambda_+)}{\lambda_+}$. Finally we use (T_i, R_i), $i = 1, \ldots, N$, to fit a least-squares regression line.
- In addition to the bias, MSE and MR, we also calculate the coverage probability (CP) of 95% confidence interval constructed using $\hat{\lambda}_{MLE}$ and the asymptotic variance given in Theorem 4.2.

Table 5: Bias and MSE of $\hat{\theta}_{MLE}$, CP (%) of $\hat{\lambda}_{MLE}$ and MR (%) of a classification rule based on $\hat{p}_{MLE}$.

Model	λ	$m=n=30$					$m=n=100$					
		$Bias(\hat{\lambda}_{MLE})$ $(MSE(\hat{\lambda}_{MLE}))$	$Bias(\hat{\alpha}_{MLE})$ $(MSE(\hat{\alpha}_{MLE}))$	$Bias(\hat{\beta}_{MLE})$ $(MSE(\hat{\beta}_{MLE}))$	CP	MR	$Bias(\hat{\lambda}_{MLE})$ $(MSE(\hat{\lambda}_{MLE}))$	$Bias(\hat{\alpha}_{MLE})$ $(MSE(\hat{\alpha}_{MLE}))$	$Bias(\hat{\beta}_{MLE})$ $(MSE(\hat{\beta}_{MLE}))$	CP	MR	OMR
	0.05	0.139 (0.091)	-0.204 (0.198)	0.006 (0.436)	100.0	18.53	0.066 (0.036)	-0.133 (0.184)	0.025 (0.427)	96.2	11.04	4.99
M1	0.20	0.093 (0.106)	-0.156 (0.183)	0.174 (0.469)	87.4	31.20	0.053 (0.059)	-0.072 (0.159)	0.174 (0.407)	88.9	27.04	18.61
$\alpha=-0.5$	0.50	0.035 (0.102)	-0.118 (0.154)	0.225 (0.376)	100.0	38.04	0.051 (0.062)	-0.079 (0.123)	0.094 (0.198)	88.1	37.29	30.85
$\beta=1$	0.80	-0.024 (0.056)	-0.125 (0.116)	0.200 (0.245)	96.1	24.50	-0.006 (0.029)	-0.086 (0.079)	0.095 (0.113)	97.8	22.23	18.61
	0.95	-0.062 (0.032)	-0.147 (0.096)	0.219 (0.205)	94.3	10.53	-0.041 (0.014)	-0.098 (0.054)	0.116 (0.082)	98.3	8.14	4.99
	0.05	0.029 (0.005)	0.123 (0.943)	0.255 (1.125)	90.3	1.13	0.008 (0.001)	0.053 (0.865)	0.594 (0.654)	92.2	0.69	0.24
M2	0.20	0.019 (0.007)	-0.073 (0.908)	0.581 (0.648)	95.3	1.24	-0.003 (0.004)	-0.330 (0.873)	0.564 (0.560)	93.7	0.90	0.48
$\alpha=-12.5$	0.50	0.004 (0.014)	-0.197 (0.889)	0.564 (0.594)	95.9	1.27	0.001 (0.005)	-0.409 (0.889)	0.411 (0.441)	94.8	0.91	0.62
$\beta=5$	0.80	-0.008 (0.014)	-0.352 (0.900)	0.500 (0.532)	95.7	0.53	0.002 (0.002)	-0.383 (0.899)	0.316 (0.386)	90.6	0.61	0.48
	0.95	0.002 (0.002)	-0.558 (0.901)	0.428 (0.471)	95.3	0.37	0.001 (0.001)	-0.366 (0.908)	0.255 (0.331)	94.8	0.29	0.24
	0.05	0.058 (0.055)	0.544 (1.104)	0.063 (0.641)	89.6	13.63	0.019 (0.019)	0.564 (1.082)	0.078 (0.536)	97.9	7.61	4.76
M3	0.20	-0.019 (0.089)	0.089 (1.296)	0.427 (0.734)	89.9	25.47	-0.026 (0.045)	0.084 (1.292)	0.399 (0.529)	93.6	21.75	13.90
$\alpha=-2$	0.50	-0.091 (0.148)	-0.392 (1.698)	0.504 (0.599)	80.4	36.13	-0.022 (0.078)	-0.319 (1.29)	0.246 (0.296)	92.5	34.89	19.05
$\beta=0.693$	0.80	-0.109 (0.119)	-0.682 (1.686)	0.335 (0.331)	90.2	26.97	-0.023 (0.036)	-0.353 (0.851)	0.119 (0.091)	98.4	22.87	13.35
	0.95	-0.139 (0.093)	-0.774 (1.529)	0.301 (0.241)	95.3	15.73	-0.057 (0.018)	-0.392 (0.622)	0.130 (0.055)	100.0	9.55	6.06
	0.05	-0.003 (0.012)	0.651 (0.889)	0.029 (0.705)	84.5	6.70	-0.024 (0.004)	0.648 (0.765)	0.201 (0.565)	81.9	4.41	3.14
M4	0.20	-0.084 (0.037)	0.255 (0.844)	0.512 (0.492)	88.0	15.53	-0.080 (0.025)	0.189 (0.789)	0.450 (0.389)	94.2	11.14	7.03
$\alpha=-4$	0.50	-0.092 (0.087)	-0.179 (0.849)	0.337 (0.271)	82.9	19.00	-0.018 (0.025)	-0.148 (0.743)	0.115 (0.097)	96.2	16.49	10.19
$\beta=1.099$	0.80	-0.027 (0.051)	-0.263 (0.791)	0.144 (0.109)	98.4	14.23	0.004 (0.007)	-0.163 (0.618)	0.053 (0.039)	94.0	13.00	7.82
	0.95	-0.027 (0.021)	-0.375 (0.673)	0.131 (0.075)	99.4	6.07	-0.007 (0.002)	-0.223 (0.468)	0.066 (0.033)	96.1	5.18	3.21
	0.05	0.166 (0.177)	NA	NA	81.3	25.97	0.082 (0.099)	NA	NA	86.2	15.31	2.5
	0.20	-0.002 (0.114)	NA	NA	88.0	23.04	-0.093 (0.040)	NA	NA	94.9	16.59	10
M5	0.50	-0.121 (0.135)	NA	NA	84.4	29.83	-0.132 (0.057)	NA	NA	84.6	26.38	25
	0.80	-0.055 (0.076)	NA	NA	94.4	22.47	-0.048 (0.032)	NA	NA	93.0	20.16	10
	0.95	-0.022 (0.018)	NA	NA	99.7	8.23	-0.006 (0.005)	NA	NA	95.2	5.92	2.5

- The $\hat{\lambda}_{MLE}$ always gives small bias and MSE.
- The $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$ generally give relatively large bias and MSE even for larger sample sizes. Nevertheless, the MR is still reasonably close to OMR.
- Even for M5 where the assumption (4.2) doesn't hold, the MLE of λ based on (4.2) performs surprisingly well and the MR doesn't deviate from OMR too much for large sample sizes.
- The CP of the confidence intervals based on $\hat{\lambda}_{MLE}$ are close to the nominal level of 95% for most of the cases.
- When $\hat{\lambda}_{MLE}$ is compared with $\hat{\lambda}$ and $\hat{\lambda}_L$, we observe that the three perform quite competitively while $\hat{\lambda}_{MLE}$ and $\hat{\lambda}$ have better MR than $\hat{\lambda}_L$.

- For a fully parametric model $\{h_\theta : \theta \in \Theta\}$, the MHDE of θ is defined as

$$\hat{\theta} = \arg\min_{t\in\Theta} \left\| h_t^{1/2} - \hat{h}^{1/2} \right\|, \tag{5.1}$$

where $\hat{h}$ is an appropriate nonparametric estimator of h_θ.

- To give the MHDE for model (4.3), define

$$\hat{h}_t(x) = (1 - t_1 + t_1 e^{t_2 + t_3 x}) f_m(x). \tag{5.2}$$

We can use the kernel density estimator h_n for $\hat{h}$.

- Then the MHDE of $\theta = (\lambda, \alpha, \beta)^\top$ is defined as

$$\hat{\theta}_{MHDE} = T(f_m, h_n) = \arg\min_{t\in\Theta} \left\| \hat{h}_t^{1/2} - h_n^{1/2} \right\|. \tag{5.3}$$

We use $\hat{p}_{MHDE}(y)$ to denote the plug-in MHDE of $p(y)$.

Consistency

Theorem 5.1. **Under certain conditions (compact supported kernel, smooth f, etc), $\hat{\theta}_{MHDE} \xrightarrow{\mathcal{P}} \theta$ as $N \to \infty$.**

Define

$$\Delta(\theta) = \int \frac{\partial w_1}{\partial \theta}(x) \left[\frac{\partial w_1}{\partial \theta}(x)\right]^\top \frac{f}{w_1}(x)dx. \tag{5.4}$$

$$\bar{\Delta}(\theta) = \int \frac{\partial w_1}{\partial \theta}(x) \left[\frac{\partial w_1}{\partial \theta}(x)\right]^\top f(x)dx, \tag{5.5}$$

$$A_N(\theta) = \int \frac{\partial w_1}{\partial \theta}(x) \left[\frac{f_m^{1/2} h_n^{1/2}}{w_1^{1/2}}(x) - f_m(x)\right] dx. \tag{5.6}$$

Theorem 5.2. **Under certain conditions (compact supported kernel, smooth f, etc),**

$$\hat{\theta}_{MHDE} - \theta \;=\; 2\left[\Delta^{-1}(\theta) + R_N\right] A_N(\theta), \qquad (5.7)$$

where R_N is a 3×3 matrix with elements tending to zero in probability as $N \to \infty$. Furthermore, the asymptotic distribution of $\sqrt{N}(\hat{\theta}_{MHDE} - \theta)$ is $N(0, \Sigma)$ with Σ defined as

$$\begin{aligned}\Sigma &= \Delta^{-1}(\theta)\left[\frac{1}{1-\rho}\bar{\Delta}(\theta) + \frac{1}{\rho}\Delta(\theta)\right]\Delta^{-1}(\theta) \\ &= \frac{1}{\rho(1-\rho)}\Delta^{-1}(\theta)\left[\Delta(\theta) - \rho\left(\Delta(\theta) - \bar{\Delta}(\theta)\right)\right]\Delta^{-1}(\theta).\end{aligned}$$

Simulation study - efficiency

We use the truncated standard normal for the kernels and the same initials as those for $\hat{\theta}_{MLE}$. The CP is for the 95% confidence interval constructed using $\hat{\lambda}_{MHDE}$ and the asymptotic variance derived in Theorem 5.2.

Table 6: Bias and MSE of $\hat{\theta}_{MHDE}$, CP (%) for $\hat{\lambda}_{MHDE}$ and MR (%) of a classification rule based on $\hat{p}_{MHDE}$.

Model	λ	$m=n=30$ $Bias(\hat{\lambda})$ $(MSE(\hat{\lambda}))$	$Bias(\hat{\alpha})$ $(MSE(\hat{\alpha}))$	$Bias(\hat{\beta})$ $(MSE(\hat{\beta}))$	CP	MR	$m=n=100$ $Bias(\hat{\lambda})$ $(MSE(\hat{\lambda}))$	$Bias(\hat{\alpha})$ $(MSE(\hat{\alpha}))$	$Bias(\hat{\beta})$ $(MSE(\hat{\beta}))$	CP	MR	OMR
	0.05	0.126 (0.066)	-0.287 (0.197)	-0.117 (0.359)	82.4	14.90	0.059 (0.026)	-0.249 (0.184)	-0.127 (0.349)	84.3	8.81	4.99
M1	0.20	0.069 (0.041)	-0.239 (0.183)	0.023 (0.387)	82.4	28.13	0.039 (0.044)	-0.120 (0.149)	0.034 (0.342)	100.0	24.97	18.61
$\alpha=-0.5$	0.5	-0.052 (0.087)	-0.165 (0.151)	0.110 (0.328)	100.0	37.93	0.039 (0.057)	-0.065 (0.119)	-0.007 (0.193)	85.5	37.98	30.85
$\beta=1$	0.80	-0.059 (0.059)	-0.132 (0.120)	0.062 (0.223)	100.0	24.03	-0.018 (0.031)	-0.064 (0.085)	-0.011 (0.121)	100.0	23.04	18.61
	0.95	-0.099 (0.040)	-0.130 (0.102)	0.059 (0.119)	100.0	12.40	-0.058 (0.019)	-0.069 (0.062)	-0.002 (0.092)	100.0	9.04	4.99
	0.05	0.054 (0.007)	0.302 (1.072)	-1.504 (5.266)	100.0	2.37	0.013 (0.001)	0.079 (0.717)	-0.789 (3.278)	100.0	1.67	0.24
M2	0.20	0.047 (0.017)	0.221 (1.240)	0.456 (1.009)	100.0	1.70	0.001 (0.006)	-0.219 (0.477)	0.534 (0.598)	100.0	0.97	0.48
$\alpha=-12.5$	0.50	0.055 (0.017)	0.213 (1.158)	0.632 (0.641)	100.0	1.33	0.029 (0.004)	-0.242 (0.397)	0.535 (0.519)	100.0	0.89	0.62
$\beta=5$	0.80	0.009 (0.009)	-0.062 (0.697)	0.552 (0.592)	99.7	1.03	0.005 (0.002)	-0.381 (0.277)	0.463 (0.434)	100.0	0.61	0.48
	0.95	0.004 (0.001)	-0.238 (0.458)	0.468 (0.599)	100.0	0.63	0.004 (0.001)	-0.438 (0.254)	0.423 (0.385)	98.8	0.29	0.24
	0.05	0.142 (0.054)	-0.188 (0.767)	-0.299 (0.265)	85.8	12.27	0.062 (0.018)	-0.131 (0.808)	-0.229 (0.219)	98.2	6.89	4.76
M3	0.20	0.091 (0.060)	-0.349 (1.195)	-0.094 (0.168)	100.0	25.00	0.049 (0.029)	-0.215 (1.182)	-0.025 (0.104)	98.6	20.73	13.90
$\alpha=-2$	0.50	0.026 (0.062)	-0.429 (1.319)	0.044 (0.097)	99.0	34.50	0.029 (0.039)	-0.199 (0.954)	0.016 (0.068)	98.0	34.06	19.05
$\beta=0.693$	0.80	-0.059 (0.043)	-0.433 (1.076)	0.072 (0.077)	100.0	25.00	-0.043 (0.028)	-0.237 (0.705)	0.035 (0.052)	100.0	23.68	13.35
	0.95	-0.116 (0.039)	-0.455 (0.937)	0.085 (0.068)	89.8	14.73	-0.088 (0.024)	-0.241 (0.557)	0.034 (0.043)	100.0	11.27	6.06
	0.05	0.067 (0.016)	-0.140 (0.544)	-0.242 (0.610)	100.0	6.30	0.003 (0.004)	-0.542 (0.915)	-0.155 (0.579)	95.8	4.51	3.14
M4	0.20	0.027 (0.024)	-0.346 (0.809)	0.059 (0.292)	95.2	12.30	-0.005 (0.012)	-0.284 (0.901)	0.093 (0.187)	95.2	10.87	7.03
$\alpha=-4$	0.50	0.016 (0.021)	-0.471 (0.922)	0.098 (0.085)	100.0	16.97	0.002 (0.011)	-0.248 (0.813)	0.042 (0.066)	97.6	15.85	10.19
$\beta=1.099$	0.80	-0.032 (0.014)	-0.472 (0.878)	0.090 (0.073)	100.0	13.17	-0.028 (0.006)	-0.281 (0.750)	0.039 (0.049)	100.0	12.94	7.82
	0.95	-0.054 (0.009)	-0.471 (0.835)	0.089 (0.070)	98.4	7.03	-0.037 (0.004)	-0.241 (0.651)	0.033 (0.048)	99.6	5.96	3.21
	0.05	0.237 (0.190)	NA	NA	99.3	25.53	0.186 (0.170)	NA	NA	97.7	22.18	2.5
	0.20	0.178 (0.127)	NA	NA	99.8	30.87	0.093 (0.056)	NA	NA	100.0	21.61	10
M5	0.50	0.076 (0.063)	NA	NA	100.0	31.90	0.037 (0.028)	NA	NA	100.0	28.77	25
	0.80	0.046 (0.029)	NA	NA	100.0	18.70	0.063 (0.012)	NA	NA	100.0	15.77	10
	0.95	0.014 (0.005)	NA	NA	100.0	5.13	0.033 (0.002)	NA	NA	100.0	4.84	2.5

- The $\hat{\lambda}_{MHDE}$ always gives small bias and MSE.
- The $\hat{\alpha}_{MHDE}$ and $\hat{\beta}_{MHDE}$ generally give relatively large bias and MSE even for larger sample sizes. Nevertheless, the MR is still reasonably close to OMR.
- Even for M5 where the assumption (4.2) doesn't hold, the MHDE of λ based on (4.2) performs surprisingly well and the MR doesn't deviate from OMR too much for large sample sizes.
- The CP of the 95% confidence interval based on $\hat{\lambda}_{MHDE}$ is higher than the nominal level 95% for most of the cases.
- When $\hat{\lambda}_{MHDE}$ is compared with $\hat{\lambda}$ and $\hat{\lambda}_L$, we observe that the three perform quite competitively while $\hat{\lambda}_{MHDE}$ and $\hat{\lambda}$ have better MR than $\hat{\lambda}_L$.

Simulation study - robustness

- We examine the behaviour of all the proposed estimators when data are contaminated by a single outlying observation. Presence of several outliers will be similar and thus omitted.
- Here we only consider the case when the outlier comes from h. We look at the change in estimate before and after data contamination.
- For this purpose, we use an adaptive version of α-IF as in Lu et al. (2003) which uses the change in estimate, divided by contamination rate (proportion of outlying observations).
- After drawing two independent samples from f and h, we replace the last observation generated from h with a single outlier, an integer with range $[-30, 20]$.

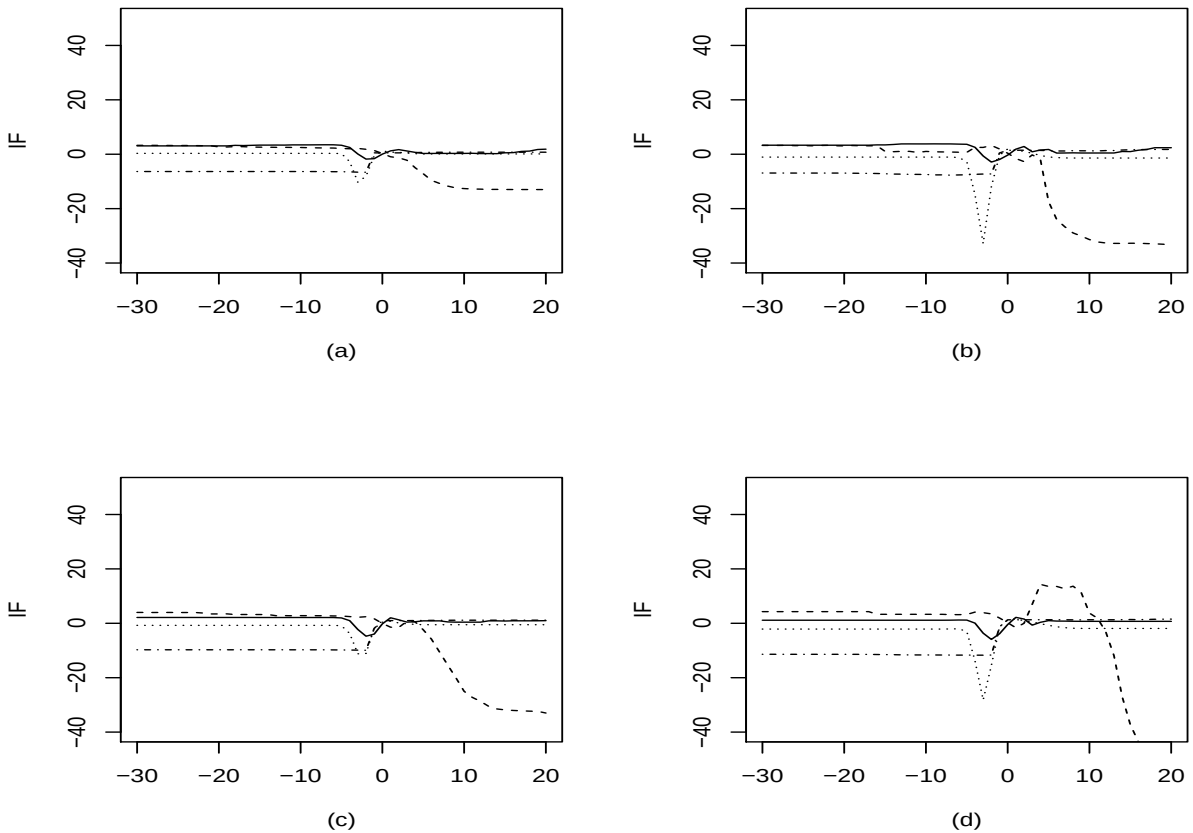

Figure 1: The α-IFs of $\hat{\lambda}$ (dotted), $\hat{\lambda}_L$ (dot-dashed), $\hat{\lambda}_{MLE}$ (dashed) and $\hat{\lambda}_{MHDE}$ (solid) for mixture model M1 $(1-\lambda)N(0,1)+\lambda N(1,1)$: (a) $\lambda = 0.15$ and $m = n = 30$; (b) $\lambda = 0.15$ and $m = n = 100$; (c) $\lambda = 0.55$ and $m = n = 30$; (d) $\lambda = 0.55$ and $m = n = 100$.

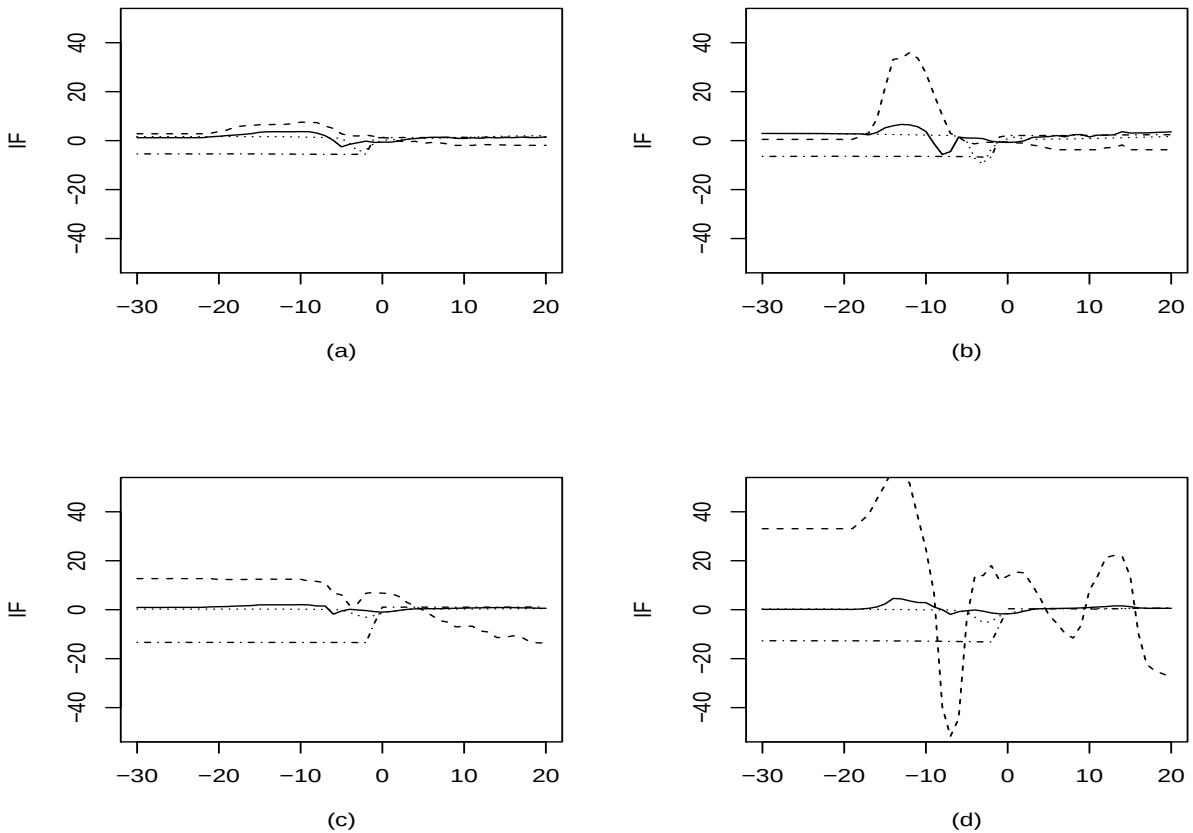

Figure 2: The α-IFs of $\hat{\lambda}$ (dotted), $\hat{\lambda}_L$ (dot-dashed), $\hat{\lambda}_{MLE}$ (dashed) and $\hat{\lambda}_{MHDE}$ (solid) for mixture model M2 $(1-\lambda)N(0,1)+\lambda N(5,1)$: (a) $\lambda = 0.25$ and $m = n = 30$; (b) $\lambda = 0.25$ and $m = n = 100$; (c) $\lambda = 0.75$ and $m = n = 30$; (d) $\lambda = 0.75$ and $m = n = 100$.

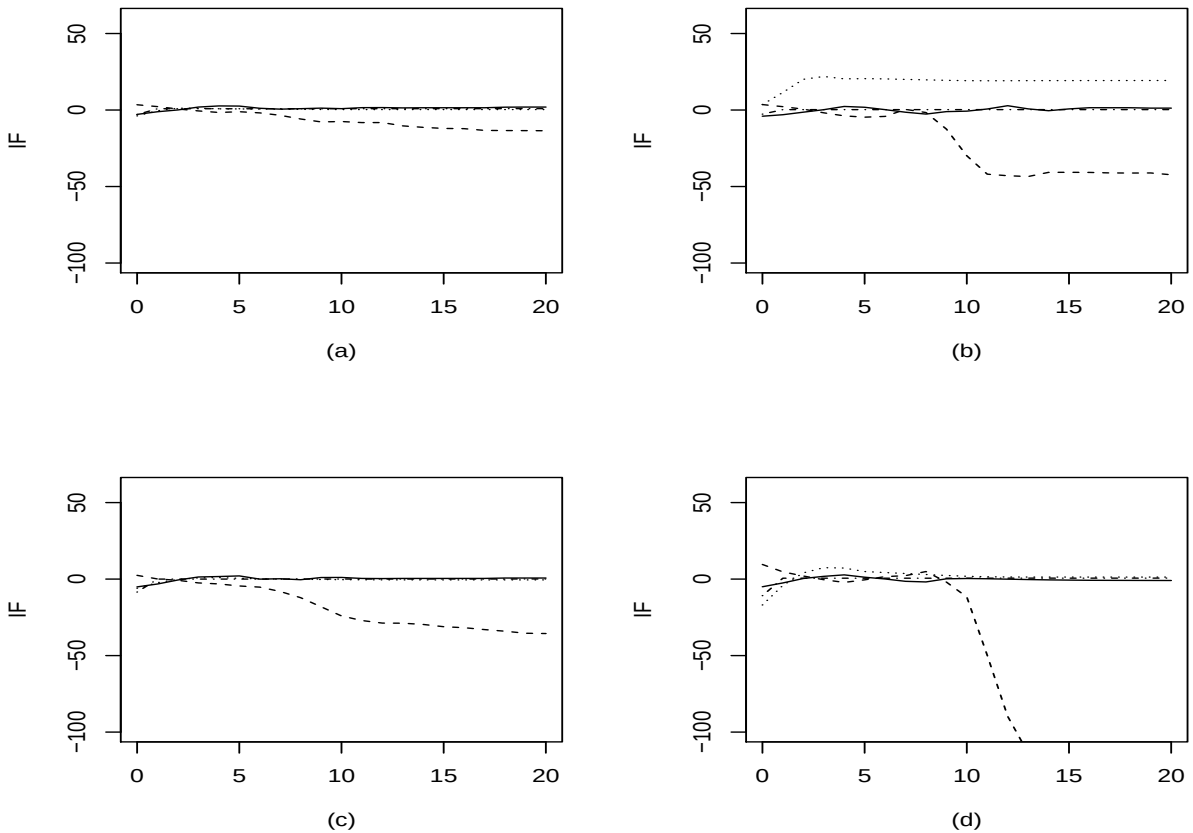

Figure 3: The α-IFs of $\hat{\lambda}$ (dotted), $\hat{\lambda}_L$ (dot-dashed), $\hat{\lambda}_{MLE}$ (dashed) and $\hat{\lambda}_{MHDE}$ (solid) for mixture model M3 $(1-\lambda)Po(2)+\lambda Po(4)$: (a) $\lambda = 0.25$ and $m = n = 30$; (b) $\lambda = 0.25$ and $m = n = 100$; (c) $\lambda = 0.75$ and $m = n = 30$; (d) $\lambda = 0.75$ and $m = n = 100$.

- $\hat{\lambda}_{MLE}$ always performs the worst, $\hat{\lambda}_{MHDE}$ performs the best and the behavior of $\hat{\lambda}$, $\hat{\lambda}_L$ and $\hat{\lambda}_{MHDE}$ are quite similar.
- The α-IF of $\hat{\lambda}_{MLE}$ is generally unbounded while that of $\hat{\lambda}$, $\hat{\lambda}_L$ and $\hat{\lambda}_{MHDE}$ seems bounded when the outlying observation increases in both directions for mixture of normals and in the right direction for mixture of Poissons.
- The bad performance of $\hat{\lambda}_{MLE}$ is mostly for when the outlying observation is bigger than 10. When the outlying observation is less than 10, the performance of $\hat{\lambda}_{MLE}$ is generally ok and is similar to that of other three estimators.
- When $\hat{\lambda}$, $\hat{\lambda}_L$ and $\hat{\lambda}_{MHDE}$ are compared, $\hat{\lambda}_L$ behaves the worst in terms of having largest α-IF for mixture of normals and $\hat{\lambda}$ behaves the worst for mixture of Poissons.
- In summary, $\hat{\lambda}_{MHDE}$ has the best robustness, followed by $\hat{\lambda}$ and then $\hat{\lambda}_L$, and $\hat{\lambda}_{MLE}$ doesn't have robustness against outliers.

K-S tests

To test the validity of model (4.3) with $r(x) = x$, we propose to use the test statistic

$$KS = N^{1/2} \sup_t \left| \hat{F}(t) - \tilde{F}(t) \right|, \qquad (6.1)$$

where $\hat{F}$ is the empirical distribution and $\tilde{F}$ is either the MLE or an estimator of F based on MHDE of $\theta = (\lambda, \alpha, \beta)^\top$ under model assumption (4.3).

Recall that, with $p_i = dF(T_i)$ and $\rho = n/N$, the MLE of p_i is

$$\hat{p}_i = \frac{1}{N\left[1 + \rho\hat{\lambda}(e^{\hat{\alpha}+\hat{\beta}T_i} - 1)\right]}, \quad i =, 1, \ldots, N. \qquad (6.2)$$

Now an estimator $\tilde{F}$ of F under model (4.3) is given by

$$\tilde{F}(t) = \sum_{i=1}^{N} \hat{p}_i I(T_i \leq t) = \frac{1}{N} \sum_{i=1}^{N} \frac{I(T_i \leq t)}{1 - \rho\hat{\lambda} + \rho\hat{\lambda} e^{(\hat{\alpha} + \hat{\beta} T_i)}}. \quad (6.3)$$

If the $\hat{\theta}$ in (6.2) and (6.3) is the MLE $\hat{\theta}_{MLE}$, then the resulting $\tilde{F}_{MLE}$ is the actual MLE of F under (4.3) and we denote the corresponding test statistic in (6.1) as KS_{MLE}.

Intuitively, we can also use $\hat{\theta}_{MHDE}$ for $\hat{\theta}$, then we denote the resulting $\tilde{F}$ in (6.3) and KS in (6.1) as $\tilde{F}_{MHDE}$ and KS_{MHDE} respectively.

We use bootstrap procedure to find the approximated distributions and critical values for hypothesis testing.

To generate bootstrapping data, we randomly select independent samples X_i^*'s from $d\tilde{F}(x)$ and Y_i^*'s from $(1-\hat{\lambda}+\hat{\lambda}e^{\hat{\alpha}+\hat{\beta}x})d\tilde{F}(x)$, where $\hat{\theta}$ and $\tilde{F}$ are either the MLEs $\hat{\theta}_{MLE}$ and $\tilde{F}_{MLE}$ or the MHDEs $\hat{\theta}_{MHDE}$ and $\tilde{F}_{MHDE}$ respectively.

We generate 1000 bootstrapping samples.

- We consider model (4.3) with $r(x) = (x, x^2)^\top$ as the collection of all possible models under consideration. Then we test whether the reduced model (4.3) with $r(x) = x$ is the actual true model or not.
- For demonstration purpose, we only consider mixture of normals $H(x) = (1-\lambda)F(x) + \lambda G(x)$ with $F \sim N(0,1)$ and $G \sim N(\mu, \sigma^2)$. Then $f(x)$ and $h(x)$ are related by

$$h_\theta(x) \;=:\; h(x) \;=\; \left(1 - \lambda + \lambda e^{\alpha+\beta x+\gamma x^2}\right) f(x), \qquad (6.4)$$

where

$$\alpha = -\frac{1}{2}\left(\log\sigma^2 + \frac{\mu^2}{\sigma^2}\right), \quad \beta = \frac{\mu}{\sigma^2}, \quad \gamma = \frac{1}{2}\left(1 - \frac{1}{\sigma^2}\right). \qquad (6.5)$$

Simulation study

- Note that (6.4) is a special case of (4.3) with $r(x) = (x, x^2)^\top$.
- If $\sigma = 1$, then $\gamma = 0$ and thus model (4.3) holds with $r(x) = x$. So testing the validity of model (4.3) with $r(x) = x$ is equivalent to testing the null hypothesis $H_0 : \gamma = 0$ under model (6.4).
- In our simulation study, we consider $\gamma = 0$, -0.9 and -1.5, $\lambda = 0.35$ and 0.65, and sample sizes $m = n = 30$ and $m = n = 100$.
- For simplicity, we just fix $\mu = 1$ and as a result $\sigma = 1$, 0.6 and 0.5 for $\gamma = 0$, -0.9 and -1.5 respectively.

- We use 500 total number of replications. Within each replication, we use totally 1000 bootstrapping samples to estimate the distribution and critical value of the test statistics KS_{MLE} and KS_{MHDE}.
- Note that $\gamma = 0$ means model (4.3) with $r(x) = x$ is correct and thus the correspondingly calculated values in the following table are the estimated significance levels. When $\gamma \neq 0$, model (4.3) with $r(x) = x$ is not correct and thus the correspondingly calculated values are the estimated powers at that value of γ.

Table 7: Estimated significance level and power of KS_{MLE} and KS_{MHDE}.

λ	γ	Significance level	$m = n = 30$		$m = n = 100$	
			KS_{MLE}	KS_{MHDE}	KS_{MLE}	KS_{MHDE}
0.35	0	0.10	0.040	0.104	0.156	0.186
		0.05	0.030	0.014	0.122	0.084
		0.01	0.002	0.000	0.002	0.002
	−0.9	0.10	0.950	0.860	0.956	0.870
		0.05	0.904	0.802	0.910	0.710
		0.01	0.734	0.410	0.578	0.184
	−1.5	0.10	0.948	0.966	0.958	0.998
		0.05	0.898	0.912	0.910	0.984
		0.01	0.716	0.580	0.536	0.846
0.65	0	0.10	0.036	0.388	0.096	0.136
		0.05	0.030	0.170	0.122	0.056
		0.01	0.008	0.010	0.002	0.006
	−0.9	0.10	0.970	0.910	0.894	0.928
		0.05	0.888	0.818	0.708	0.758
		0.01	0.464	0.282	0.158	0.120
	−1.5	0.10	0.956	0.876	0.990	0.990
		0.05	0.858	0.762	0.908	0.944
		0.01	0.424	0.302	0.174	0.396

- The two test statistics KS_{MLE} and KS_{MHDE} are quite competitive in terms of achieved significance level and power.
- The achieved levels of significance are quite close to the true levels for most of the cases except for the case of KS_{MHDE} with $\lambda = 0.65$ and $m = n = 30$.
- The powers of KS_{MHDE} become larger when γ is away from 0 except for the case with $\lambda = 0.65$ and $m = n = 30$.
- Surprisingly, the powers of KS_{MLE} become smaller when γ is away from 0 except for the case with $\lambda = 0.65$ and $m = n = 100$.
- As expected, when the significance level a decrease, both the observed significance level and power decrease.
- For both KS_{MLE} and KS_{MHDE}, the powers are generally high for significance levels $a = 0.10$ and 0.05.

Example 1: Grain data

Smith and Vounatsou (1997) analyzed a data where an autoradiography assay was used to determine the intracellular transfer of small molecules in mouse cells in culture. The assay was used to determine the proportion of cells in the test population which were exposed to radio active materials. The cells in control group were not exposed to radioactivity, but otherwise were similar in nature. Autoradiograph of the cells can determine the amount of radio active material in the cell by counting the number of grains, X. Now grains can appear in autoradiograph due to the presence of radioactive material or due to background fogging. Hence the proportion of cells with radio active material can only be revealed by comparing the distribution of grain counts in test sample and that in control sample.

Table 8: Frequency distribution for the test group and control group in the grain data.

Number of grains (X)	Frequency in recipients (test sample from mixture h)	Frequency in controls (control sample from f)
0	2	3
1	2	6
2	2	12
3	3	16
4	4	8
5	3	11
6	1	9
7	2	5
8	4	9
9	2	5
10	4	5
11	3	1
12	4	3
14	3	0
15	1	0
16	2	0
17	1	1
18	2	0
>19	49	0
Total	94	94

Table 9: Point and interval estimation of the proportion for the grain data.

Method	Estimate	95% confidence interval
Poisson mixture (Smith, Smith and Hooper, 1986)	0.77	0.00 – 0.91
Two-by-two table (Smith and Vounatsou, 1997)	0.20	0.00 – 1.00
Logistic power (Smith and Vounatsou, 1997)	0.61	0.58 – 0.64
Monotone logistic (Smith and Vounatsou, 1997)	0.74	0.61 – 1.00
Latent class (Smith and Vounatsou, 1997)	0.73	0.63 – 0.83
$\hat{\lambda}$ based on c.d.f.s	0.78	0.68 – 0.87
$\hat{\lambda}_L$ based on multinomial approximation	0.79	0.58 – 0.88
$\hat{\lambda}_{MLE}$ based on semiparametric MLE	0.75	0.61 – 0.88 (0.60 – 0.89)
$\hat{\lambda}_{MHDE}$ based on semiparametric MHDE	0.76	0.64 – 0.92 (0.65 – 0.88)

We use bootstrap method with 1000 bootstrapping samples to calculate 95% confidence intervals.

λ: the proportion of cells in the test population which were exposed to radio active materials.

- Our proposed methods give very similar point estimate of λ in comparison with current available methods in literature.
- Our proposed four estimators give reasonable confidence intervals strictly within the range $[0, 1]$ and with relatively smaller widths.
- The confidence interval in the parentheses for $\hat{\lambda}_{MLE}$ and $\hat{\lambda}_{MHDE}$ are calculated using the derived asymptotic covariance matrices. From the results we see that bootstrap approximation is quite accurate.

Example 2: Malaria data

The parasite densities in children with fever is formulated using a two-component mixture model, where one component represents the parasite densities in children without clinical malaria (f) and the other with clinical malaria (g). Parasite levels in children from the community are available and are used as a training sample, i.e. a sample that comes from the component of the mixture corresponding to children without clinical malaria (f) but who may have parasites. The mixing proportion λ represents the proportion of children whose fever is attributable to malaria.

This data were first described in Kitua et al. (1996). The data arose from repeated cross-sectional surveys of parasitaemia and fever among 426 children up to one year old resided in a village in Kilombero district in Tanzania. A subset of this data was analyzed by Vounatsou, Smith and Smith (1998) where they considered children aged between 6 and 9 months and two seasons: the wet season (January-June) during which the mosquito population, and hence exposure to malaria infection, is high, and the dry season (July-December) during which the mosquito population is lower. The original data were grouped into 10 categories and the parasite level refers to the midpoint of each category.

Table 10: Frequency distribution of parasite density for children aged between 6 and 9 months in the malaria data.

Category	Wet Season Parasite level	Frequency f	Frequency h	Dry Season Parasite level	Frequency f	Frequency h
1	0	43	60	0	43	42
2	3251	40	58	11370	68	116
3	9673	3	14	34029	8	30
4	16095	3	13	56689	2	16
5	22518	2	10	79348	0	7
6	28940	1	8	102008	0	7
7	35362	0	7	124668	0	6
8	41785	1	6	147327	0	2
9	48207	1	6	169987	0	3
10	225685	0	69	290634	0	16
Total		94	251		122	245

- We apply $\hat{\lambda}_L$ and $\hat{\lambda}_{MLE}$ to this data and compare them with the Bayesian approach proposed by Vounatsou, Smith and Smith (1998).
- Note that this is a discretized data, so kernel smoothing is not appropriate and thus $\hat{\lambda}$ and $\hat{\lambda}_{MHDE}$ are not applied.
- The numbers in parentheses are the estimated standard errors based on 500 bootstrapping samples.
- From the results we can see that both $\hat{\lambda}_L$ and $\hat{\lambda}_{MLE}$ give consistent estimates with that of the Bayesian approach.

Key References

- Beran, R. (1977). Minimum Hellinger distance estimators for parametric models. *Ann. Statist.* **5**, 445-463.
- Karunamuni, R.J. and Wu, J. (2009). Minimum Hellinger distance estimation in a nonparametric mixture model. *J. Statist. Plann. Inference* **139**, 1118-1133.
- Qin, J. and Zhang, B. (1997). A goodness of fit test for logistic regression models based on case-control data. *Biometrika* **84**, 609-618.
- Smith, T.A., Smith, A.G. and Hooper, M.L. (1986). Selection of a mouse embryonal carcinoma clone resistant to the inhibition of metabolic cooperation by retinoic acid. *Exp. Cell Res.* **165**, 417-430.
- Smith, T. and Vounatsou, P. (1997). Logistic regression and latent class models for estimating positives in diagnostic assays with poor resolution *Comm. Statist. Theory methods* **27**, 1677-1700.
- Wu, J., Karunamuni, R.J. and Zhang, B. (2010). Minimum Hellinger distance estimation in a two-sample semiparametric model. *J. Multivariate Anal.* **101**, 1102-1122.

Thank You!

Banff & Jasper

Banff & Jasper

Banff & Jasper

Banff & Jasper