



Published in final edited form as:

Commun Stat Theory Methods. 2017 ; 46(18): 9092–9100. doi:10.1080/03610926.2016.1205612.

Testing homogeneity in semiparametric mixture case-control models

Chong-Zhi Di,

Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M2-B500, Seattle, WA 98109, USA

Kwun Chuen Gary Chan,

University of Washington, Seattle WA, USA

Cheng Zheng, and

University of Wisconsin, Milwaukee WI, USA

Kung-Yee Liang

National Yang-Ming University, Taiwan

Abstract

Parametric and semiparametric mixture models have been widely used in applications from many areas, and it is often of interest to test homogeneity in these models. However, hypothesis testing is nonstandard due to the fact that several regularity conditions do not hold under the null hypothesis. We consider a semiparametric mixture case-control model, in the sense that the density ratio of two distributions is assumed to be of an exponential form, while the baseline density is unspecified. This model was first considered by Qin and Liang (2011, biometrics), and they proposed a modified score statistic for testing homogeneity. In this paper, we consider alternative testing procedures based on supremum statistics, which could improve power against certain types of alternatives. We demonstrate the connection and comparison among proposed and existing these approaches. In addition, we provide a unified theoretical justification of the supremum test and other existing test statistics from an empirical likelihood perspective. The finite sample performance of the supremum test statistics were evaluated in simulation studies.

1 Introduction

Parametric mixture and admixture models have been widely used in public health and biomedical studies to account for possible heterogeneity in the population. In particular, a two-component admixture model has density function of the form,

$$(1 - \lambda)f(x; \theta_0) + \lambda f(x; \theta), \quad (1)$$

where λ is the proportion of the second component, $f(\cdot; \boldsymbol{\theta})$ is a family of parametric distribution, and $\boldsymbol{\theta}_0$ a fixed value that corresponds to the first component. For example, in genetic linkage analysis, admixture models with binomial kernels have been used to account for genetic heterogeneity (Smith, 1963). In such models, it is of interest to test null hypothesis of homogeneity, $H_0: \lambda = 0$ or $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, which is equivalent to no effects (no linkage in the genetic linkage example). This hypothesis testing problem is nonstandard, due to nonidentifiability and the fact that $\lambda = 0$ is on the boundary of its parameter space under the null. Several testing approaches have been proposed for this model (Smith, 1963; Shoukri and Lathrop, 1993; Lemdani and Pons, 1995; Liang and Rathouz, 1999; Fu et al., 2006). Recently, Di and Liang (2011) systematically studied the asymptotics of the likelihood ratio test statistics (LRT) for general parametric admixture models, and provided insights on connections and comparisons of the LRT versus of other testing procedures.

Recently, Qin and Liang (2011) considered a semiparametric mixture case-control model,

$$h(x) = (1 - \lambda)f(x) + \lambda g(x),$$

$$g(x) = \exp(\alpha + \beta x)f(x),$$

where $f(x)$, $g(x)$ and $h(x)$ are density functions. This model can be viewed as a semiparametric extension to the parametric admixture model (1). Two sets of independent datasets are generated from this model. The first sample, $\{x_1, \dots, x_{n_0}\}$, comes from the

“control” group with density $f(x)$, while the second sample, $\{y_1, \dots, y_{n_1}\}$, contains

contaminated cases, i.e., a mixture of realizations from $f(x)$ and $g(x)$ with proportions $1 - \lambda$ and λ , respectively. The density functions $g(x)$ and $f(x)$ are assumed to follow an exponential tilted density ratio model as above. Note that the parameters α , β and the density $f(x)$ must satisfy a constraint

$$\int \exp(\alpha + \beta x)f(x)dx = 1, \quad (2)$$

or equivalently $E_f\{\exp(\beta x)\} = \int \exp(\beta x)f(x)dx = \exp(-\alpha)$, where E_f denotes expectation with respect to the distribution $f(\cdot)$. Constraint (2) holds because $g(x)$ is a probability density function. The goal is to test homogeneity of two groups, and thus the null hypothesis can be represented as $H_0: \lambda = 0$ or $\alpha = \beta = 0$, which is also nonstandard due to loss of identifiability and boundary issue under the null. As Qin and Liang (2011) pointed out, the motivation for this family of mixture case-control model comes from a wide variety of applications, including case-control studies with contaminated controls, clinical trials with “non-responders”, genetic linkage and association analysis, infectious epidemiological studies, etc. They then developed a version of score test for homogeneity (henceforth denoted as “QL”). However, hypothesis testing for this semiparametric model are not well studied to the best of our knowledge.

In this paper, we generalize the likelihood ratio or supremum statistics to the semiparametric mixture model. We also demonstrate the connection and comparison of different approaches in parametric and semiparametric mixture models. In addition, we provide a unified theoretical justification of the supremum test as well as the QL approach from an empirical likelihood perspective.

2 Methods for testing homogeneity

Depending on whether one is willing to make distributional assumptions, the mixture model can be parametric or semiparametric. If one assumes that the density $f(x)$ is known, then the model becomes a parametric admixture model (Di and Liang, 2011), which has been used for genetic linkage analysis with binomial kernel distributions (Smith, 1963; Lemdani and Pons, 1995; Liang and Rathouz, 1999; Fu et al., 2006). If one leaves $f(x)$ unspecified, the model is semiparametric. In this section, we will first review existing approaches for the parametric model, before introducing the proposed procedures for the semiparametric mixture model.

2.1 Parametric setting: $f(x)$ is specified

When $f(x)$ is known, the first dataset $\{x_1, \dots, x_{n_0}\}$ does not contribute to the likelihood function, and only the second dataset $\{y_1, \dots, y_{n_1}\}$ contributes to the inference for λ , α and β .

The testing problem for this parametric admixture model is nonstandard. We briefly review existing approaches, which provide insights for extension to the semiparametric setting.

When β is known/fixed ($\beta = 0$)—The null hypothesis can be represented by $H_0 : \lambda = 0$ only. The parameter value $\lambda = 0$ is on the boundary of its parameter space $[0, 1]$, but one can obtain the asymptotics of the likelihood ratio test using theory of Self and Liang (1987),

$$LRT(\beta) = 2 \left\{ \sup_{\lambda \in [0, 1]} l_n(\lambda, \beta) - l_n^0 \right\} = \{W_n^+(\beta)\}^2 + o_p(1) \rightarrow 0.5\chi_1^2 + 0.5\chi_1^2, \quad (3)$$

$$\text{where } W_n(\beta) = \frac{\sum_{j=1}^{n_1} \{\exp(\alpha + \beta y_j) - 1\}}{\sqrt{\sum_{j=1}^{n_1} \{\exp(\alpha + \beta y_j) - 1\}^2}} = \frac{\sum_{j=1}^{n_1} \{\exp(\beta y_j) - \exp(-\alpha)\}}{\sqrt{\sum_{j=1}^{n_1} \{\exp(\beta y_j) - \exp(-\alpha)\}^2}}.$$

Note $W_n(\beta)$ is equivalent to the standardized score statistic for λ evaluated at $\lambda = 0$.

When β is unknown—Identifiability is lost under the null, as any value of β yields the same null distribution when $\lambda = 0$. Two types of approaches were proposed to circumvent nonidentifiability: taking the supremum over a range of β or substituting β by a suitable estimate. For the former, we demonstrated in our previous work that the likelihood ratio test statistic for parametric mixture models (denoted M^{par}) has the following asymptotic property Di and Liang (2011),

$$M^{par} = \sup_{\beta \in B} LRT(\beta) = \sup_{\beta \in B} \{W_n^+(\beta)\}^2 + o_p(1), \quad (4)$$

where $W_n^+(\beta) = \max\{W_n(\beta), 0\}$ is the positive part of $W_n(\beta)$. Thus, the asymptotic distribution of the likelihood ratio test is the supremum of a squared Gaussian process. Alternatively, Liang and Rathouz (1999) derived a modified score test statistic, $W_n^+(\hat{\beta}_0)$, where $\hat{\beta}_0$ is the constrained MLE of β while fixing λ , e.g., taking $\lambda = 1$. Di and Liang (2011) discussed connections and power comparisons between these approaches. In terms of power, the modified score statistic performs well under some scenarios, but may lose power to detect certain type of alternatives, e.g., with small λ and large β . On the other hand, the likelihood ratio or supremum statistic is omnibus, as it is shown to be powerful for all types of alternatives.

2.2 Semiparametric setting: $f(x)$ is unspecified

When $f(x)$ is unknown, the mixture model is semiparametric. Under this setting, α is still a function of β , but the functional form depends on $f(\cdot)$ and thus is unknown. Testing homogeneity in this model is relatively new, except Qin and Liang (2011). Our aim is to consider the hypothesis testing problem in this semiparametric mixture model, propose alternative test statistics and compare it with existing approaches.

In the case of known $f(\cdot)$, the density function enters the test statistics only through $\exp(\alpha)$ implicitly. Thus, one natural way of extending the situation to unknown $f(\cdot)$ is to replace $\exp(\alpha)$ by its nonparametric estimate. Due to the constraint that $\exp(-\alpha) = \int \exp(\beta x) f(x) dx$, we propose to use its nonparametric estimate, $\frac{\sum_{i=1}^{n_0} \exp(\beta x_i)}{n_0}$ or

$\frac{\sum_{i=1}^{n_0} \exp(\beta x_i) + \sum_{j=1}^{n_1} \exp(\beta y_j)}{n_0 + n_1}$, in the test statistic $W_n(\beta)$. Under the null hypothesis, one can use the first sample, or the combined sample to estimate this quantity. We choose to use the latter for efficiency considerations.

When β is known/fixed—By replacing $\exp(-\alpha)$ with its nonparametric estimator, the

numerator of (3) is proportional to $\bar{U}_1 - \bar{U}_0$, where $\bar{U}_1 = \frac{\sum_{j=1}^{n_1} \exp(\beta y_j)}{n_1}$ and

$\bar{U}_0 = \frac{\sum_{i=1}^{n_0} \exp(\beta x_i)}{n_0}$. Therefore, we plan to use $\{\tilde{W}_n(\beta)\}^2$ as test statistic,

$$\tilde{W}_n(\beta) = \frac{\bar{U}_1 - \bar{U}_0}{\sqrt{S^2(\frac{1}{n_1} + \frac{1}{n_0})}}, \text{ where } S^2 = \frac{\sum_{j=1}^{n_1} \{\exp(\beta y_j) - \bar{U}_1\}^2 + \sum_{i=1}^{n_0} \{\exp(\beta x_i) - \bar{U}_0\}^2}{n_1 + n_0 - 2}$$

.It is interesting to note that the test statistic is equivalent to the square of a two- sample T- test on exponentially transformed data, so its limiting distribution is χ_1^2 for a fixed β . With a small β value ($\beta \rightarrow 0$), the test statistics is asymptotically equivalent to the two-sample T- test on the original data, and thus also asymptotically equivalent to Qin and Liang (2011).

When β is unknown—Naturally, when β is unknown, there is loss of identifiability under the null hypothesis $H_0 : \lambda = 0$ or $\alpha = \beta = 0$. One can extend two types of approaches from the parametric case, by taking the supremum of score statistics over a range of plausible β values or plugging in a suitable estimate of (α, β) . We consider the former approach by varying β over its parameter space B and taking the supremum, i.e., using

$M^{semi} = \sup_{\beta \in B} \{\tilde{W}_n(\beta)\}^2$ as a test statistic. It can be shown that the limiting distribution of M^{semi} is equivalent to the supremum of a squared Gaussian process, analogous to results for parametric mixture models.

Note that Qin and Liang (2011) adopted the latter approach, as an extension of Liang and Rathouz (1999) from parametric to semiparametric settings, and derived a modified score test for homogeneity. More specifically, they first fixed $\lambda = 1$ and obtained $(\tilde{\alpha}, \tilde{\beta})$ by fitting a logistic regression model. They then used the score statistic evaluated at $(\tilde{\alpha}, \tilde{\beta})$, which was equivalent to $\{\tilde{W}_n(\tilde{\beta})\}^2$, and showed its limiting distribution to be χ_1^2 . This QL approach was asymptotically equivalent to the two sample T-test, or $\{\tilde{W}_n(\beta)\}^2$ with small β . However, they demonstrated that the QL was often more powerful than the two sample T-test in finite samples.

In finite samples, since the process $\{\tilde{W}_n(\beta)\}^2$ is often skewed, we suggest using permutation or bootstrap methods to calculate p values for both the supremum and QL approaches. In terms of power comparison, it is expected that the QL performs well under some scenarios, but may lose power to detect certain type of alternatives, e.g., with small λ and medium to large β . On the other hand, the supremum statistic is expected to be omnibus, with decent power to detect all types of alternatives.

Remark 1—In the parametric mixture model (Section 2.1), the term $\exp(-\alpha)$ is determined by the parametric distributional assumption and is a known function of β . In contrast, one needs to estimate $\exp(-\alpha)$ nonparametrically in constructing test statistics for the semiparametric mixture model. Compared to the parametric tests, the semiparametric testing procedures are more robust to mis-specifications, but can be less efficient if parametric assumptions are plausible.

Remark 2—In practice, if the parameter space of β is not compact, e.g., $\beta \in (-\infty, \infty)$, one can choose a compact sub-interval $\beta \subset R$ to take supremum on. In practice, it is a good idea to avoid choosing β with very large absolute values, because the exponentially transformed data may be too skewed and the test statistics may be dominated by a few outliers. If the data are standardized with mean 0 and variance 1, we find that maximizing over $\beta \in [-1, 1]$ or $[-2, 2]$ often work well from our empirical simulations.

3 The empirical likelihood perspective

In this section, we provide a unified empirical likelihood framework to derive various test statistics in the semiparametric mixture model. The likelihood function involves an infinite dimensional nuisance parameter $f(\cdot)$ in addition to parameters of interest λ , α and β . In the previous section, we derived test statistics from the usual likelihood, by replacing terms that implicitly involve $f(\cdot)$ with their nonparametric estimates. The empirical likelihood approach is an alternative inferential procedure for semiparametric models. As will be shown below, this approach provides an alternative theoretical justification for the QL approach and supremum statistics in Section 2 of this paper.

Let $F(\cdot)$ be the probability distribution function corresponding to $f(\cdot)$, $n = n_0 + n_1$, $\{z_1, \dots, z_n\} = \{x_1, \dots, x_{n_0}, y_1, \dots, y_{n_1}\}$ be the combined sample and $p_1 = dF(z_1), \dots, p_n = dF(z_n)$.

The notations are used in the empirical likelihood method of Qin and Zhang (1997) for the special case of $\lambda = 1$. We re-write the log-likelihood functions in terms of empirical likelihood weights p_1, \dots, p_n .

$$l(\lambda, \alpha, \beta) = \sum_{i=1}^{n_0} \log p_i + \sum_{j=1}^{n_1} \log \{(1 - \lambda)p_j + \lambda \exp(\alpha + \beta y_j)p_j\} = \sum_{i=1}^n \log p_i \quad (5)$$

$$+ \sum_{j=1}^{n_1} \log \{(1 - \lambda) + \lambda \exp(\alpha + \beta y_j)\}.$$

We maximize $l(\lambda, \alpha, \beta)$ subject to the following constraints:

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i \exp(\alpha + \beta z_i) = 0 \quad \text{and} \quad p_i \geq 0, i = 1, \dots, n. \quad (6)$$

The constraints are standard in empirical likelihood methods for logistic regression model (Qin and Zhang, 1997). Fixing (λ, α, β) and solving the constrained maximization problem gives

$$p_i = \frac{1}{n \{1 + \nu (\exp(\alpha + \beta z_i) - 1)\}} \quad (7)$$

where ν is a Lagrange multiplier which solves

$$-\sum_{i=1}^n \frac{\nu \exp(\alpha + \beta z_i)}{1 + \nu(\exp(\alpha + \beta z_i) - 1)} + \sum_{j=1}^{n_1} \frac{\lambda \exp(\alpha + \beta y_j)}{(1 - \lambda) + \lambda \exp(\alpha + \beta y_j)} = 0.$$

Substituting (7) into (5) gives a profile likelihood

$$l_p(\lambda, \alpha, \beta) = -\sum_{i=1}^n \log[1 + \nu\{\exp(\alpha + \beta z_i) - 1\}] + \sum_{j=1}^{n_1} \log\{(1 - \lambda) + \lambda \exp(\alpha + \beta y_j)\}. \quad (8)$$

We now take a closer look at the Lagrange multiplier ν . Note that

$$\sum_{i=1}^n \frac{\exp(\alpha + \beta z_i)}{1 + \nu(\exp(\alpha + \beta z_i) - 1)} = n \cdot \sum_{i=1}^n p_i \exp(\alpha + \beta z_i) = n.$$

Therefore,

$$\nu = \frac{1}{n} \sum_{j=1}^{n_1} \frac{\lambda \exp(\alpha + \beta y_j)}{(1 - \lambda) + \lambda \exp(\alpha + \beta y_j)}.$$

The Lagrange multiplier ν is a function of parameters (λ, α, β) and data, but when $\lambda = 1$ the Lagrange multiplier does not depend on data. In fact, $\nu = n_1/n = \rho/(1 + \rho)$ for usual case control studies (Qin and Zhang, 1997), where $\rho = n_1/n_0$ is the case-control sample size ratio. The dependence of ν on data when $\lambda \neq 1$ makes the construction of a score test based on l_p complicated. To overcome the challenge, we substitute ν by its probability limit which depends only on ρ and λ . By the law of large numbers, ν converges to

$$\frac{n_1}{n} \int \frac{\lambda \exp(\alpha + \beta y)}{(1 - \lambda) + \lambda \exp(\alpha + \beta y)} h(\lambda) dy = \frac{n_1}{n} \int \lambda \exp(\alpha + \beta y) f(y) dy = \frac{n_1}{n} \lambda \int g(y) dy = \frac{\rho \lambda}{1 + \rho}.$$

We substitute ν by its probability limit in (8) and obtain the following pseudo-likelihood function up to constants,

$$l_s(\lambda, \alpha, \beta) = -\sum_{i=1}^n \log\{1 + \rho - \rho \lambda + \rho \lambda \exp(\alpha + \beta z_i)\} + \sum_{j=1}^{n_1} \log\{(1 - \lambda) + \lambda \exp(\alpha + \beta y_j)\}.$$

Using the pseudo-likelihood function $l_s(\lambda, \alpha, \beta)$ as an objective function, one can construct pseudo-score or pseudo-likelihood ratio test statistics. This approach actually provides a unified framework to develop various test statistics for semiparametric mixture models, in parallel to test statistics for parametric mixture models. For example, following the argument of Liang and Rathouz (1999), i.e., plugging in suitable estimates of α and β , one can obtain exactly the score test statistics of Qin and Liang (2011). If we fix $\lambda > 0$ and construct score

statistics with respect to β , we can obtain two-sample T-test of untransformed data, which is also asymptotically equivalent to the QL approach. On the other hand, following the similar supremum statistics argument of Di and Liang (2011), we can obtain test statistics that are equivalent to those proposed in Section 2.2. Thus, the empirical likelihood perspective provides a nice unified justification of all test statistics discussed above.

4 Numerical results

We conducted simulation studies to compare the supremum test versus the two-sample T-test and Qin & Liang's approach (denoted T and T^{QL} , respectively). For the proposed supremum test, we implemented two versions, one based on the parametric mixture model (assuming the parametric kernel distribution is known; denoted M^{par}) and the other based on the semiparametric mixture model (denoted M^{semi}).

Table 1 displays the empirical type I error rates and statistical power when the true model is the Normal mixture model, where the null model is $N(0,1)$ and the alternative model is $(1 - \lambda) N(0, 1) + \lambda N(\mu, 1)$. For the supremum statistics M^{par} and M^{semi} , we take supremum over the range of $\mu \in [-2, 2]$. First, we compare three test statistics that are derived under the semiparametric mixture model and thus do not depend on parametric distributional assumptions, T , T^{QL} and M^{semi} . The supremum statistic M^{semi} is often comparable or more powerful than T^{QL} , with a power gain of around 10 – 15% when λ is small and μ is relatively large (e.g., $\lambda = 0.2$, $\mu = 2$, or $\lambda = 0.3$, $\mu = 2$). We also confirmed that T^{QL} is more powerful than T in this finite sample setting, although they are asymptotically equivalent. Next, we compare these test statistics to M^{par} , which is the likelihood ratio test under parametric distributional assumptions. One can see that M^{par} is substantially more powerful over all other tests, by margins as large as 30 – 50%. This implies that knowing the distributional form of the “control” group can greatly improve power for the case-control mixture models. We also conducted simulation studies under mixture models with t or exponential kernel distributions, and show the results in Table 2. In general, M^{semi} performs better when the distributions are symmetric, while T^{QL} is more powerful for skewed distributions.

We applied the proposed methods to test differentially expressed genes in a prostate cancer microarray study. The study obtained expressions of 5,153 genes from 103 tissue samples, 41 of which are normal and 62 of which are cancerous. The details of study was described in Lapointe et al. (2004). In this analysis, we used a subset of data with 2,000 randomly selected genes, which are publicly available (R package “TailRank”). Four tests, the two-sample T-test T , the moderated T-test T^{mod} , the QL approach T^{QL} and proposed supremum statistic M^{semi} , were applied to the data. In this application, M^{semi} identified more differentially expression genes than T , and T^{QL} , as demonstrated in Table 3.

To summarize, we evaluated finite sample performance of T , T^{QL} , M^{semi} and M^{par} . Among the semiparametric approaches, M^{semi} is more powerful when the kernel distribution is symmetric, while T^{QL} is more powerful when the kernel distribution is especially skewed. Both T^{QL} and M^{semi} are more powerful or on par with T in the scenarios considered. If one has prior knowledge of plausible parametric distributions and is willing to make such

assumptions, the parametric approach M^{par} can gain power substantially over semiparametric approaches. On the other hand, the semiparametric approaches do have the advantage of robustness against distributional assumptions.

5 Discussion

In this paper, we considered the problem of testing homogeneity in parametric and semiparametric mixture case-control models. We demonstrated the connection and comparison of different approaches in parametric or semiparametric mixture models. We also extended our previous work on parametric mixture models to the semiparametric mixture model, and proposed supremum test statistics. The proposed statistics are equivalent to the supremum of two sample T-tests on exponentially transformed data, and are shown to improve power from alternative procedures especially for symmetric distributions. Another contribution of this paper is that we developed a pseudo-likelihood of the semiparametric mixture model based on empirical likelihood perspective, which provides a unified theoretical framework to construct and connect various testing procedures.

In terms of future work, it will be interesting to extend this work to semiparametric mixture models for censored time-to-event data. For example, in a randomized clinical trials, some subjects in the treatment group might be "nonresponders" who do not respond to the treatment. Under this situation, the treatment group is composed of responders and nonresponders, so semiparametric mixture models are natural tools to make inference on treatment effects (Pikounis and Rao, 1994). Another direction is to study more general models where the exponential tilt density ratio assumption might not hold. For example, a two-component normal mixture model with unknown mean and variance parameters in each component does not satisfy the exponential tilt assumption. It will be of interest to develop testing procedures for a more general semiparametric family of mixture models.

References

- Di CZ, Liang KY. Likelihood ratio testing for admixture models with application to genetic linkage analysis. *Biometrics*. 2011; 67(4):1249–1259. [PubMed: 21385166]
- Fu Y, Chen J, Kalbfleisch JD. Testing for Homogeneity in Genetic Linkage Analysis. *Statistica Sinica*. 2006; 16:805–823.
- Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*. 2004; 101(3):811–816. [PubMed: 14711987]
- Lemdani M, Pons O. Tests for Genetic Linkage and Homogeneity. *Biometrics*. 1995; 51(3):1033–1041. [PubMed: 7548688]
- Liang KY, Rathouz PJ. Hypothesis testing under mixture models: application to genetic linkage analysis. *Biometrics*. 1999; 55(1):65–74. [PubMed: 11318180]
- Pikounis VB, Rao P. Linear rank tests for the nonresponders problem with censored data. *Biometrika*. 1994; 81(4):691–699.
- Qin J, Liang KY. Hypothesis testing in a mixture case-control model. *Biometrics*. 2011; 67(1):182–198. [PubMed: 20345497]
- Qin J, Zhang B. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*. 1997; 84(3):609–618.

- Self S, Liang K. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*. 1987;605–610.
- Shoukri M, Lathrop G. Statistical testing of genetic linkage under heterogeneity. *Biometrics*. 1993; 49(1):151–161. [PubMed: 8513100]
- Smith CAB. Testing for heterogeneity of recombination fraction values in human genetics. *Annals of Human Genetics*. 1963; 27(2):175–182. [PubMed: 14081488]

Table 1

Simulated type I error rates and statistical power for Normal mixture models ($\mu_0 = \mu_I = 50$). The number of simulations is 10,000. The true distributions are $F: N(0,1)$ and $H: (1 - \lambda)N(0,1) + \lambda N(\mu, 1)$.

λ	μ	α	T	T ^{QL}	M ^{semi}	M ^{par}
0	0	0.05	0.051	0.050	0.050	0.051
0	0	0.01	0.009	0.009	0.011	0.013
0.1	2	0.05	0.150	0.179	0.188	0.579
0.1	2	0.01	0.042	0.060	0.053	0.398
0.2	2	0.05	0.408	0.463	0.521	0.923
0.2	2	0.01	0.185	0.243	0.249	0.838
0.3	2	0.05	0.703	0.756	0.822	0.992
0.3	2	0.01	0.446	0.532	0.579	0.978
0.5	2	0.05	0.981	0.988	0.995	1.000
0.5	2	0.01	0.921	0.945	0.969	1.000
0.3	1	0.05	0.290	0.304	0.313	0.616
0.3	1	0.01	0.113	0.124	0.134	0.412
0.5	1	0.05	0.646	0.665	0.666	0.937
0.5	1	0.01	0.399	0.420	0.432	0.838
0.7	1	0.05	0.912	0.917	0.910	0.997
0.7	1	0.01	0.756	0.765	0.776	0.984

Table 2

Simulated type I error rates and statistical power when the exponential tilt assumption does not hold: mixtures of t or exponential distributions ($n_0 = 50, n_1 = 50$). The number of simulations is 10,000. The true distributions are $F: \kappa(2)$ and $H: (1 - \lambda)\kappa(2) + \lambda \{ \kappa(2) + 2 \}$ for the t -distribution case, and $F: Exp(1)$ and $H: ((1 - \lambda)Exp(1) + \lambda Exp(0.5))$ for the exponential distribution case.

<i>T</i> -mixture					Exponential mixture				
λ	α	T	T ^{QL}	M ^{semi}	λ	α	T	T ^{QL}	M ^{semi}
0	0.05	0.048	0.048	0.049	0	0.05	0.049	0.052	0.049
0	0.01	0.009	0.009	0.009	0	0.01	0.011	0.009	0.011
0.2	0.05	0.141	0.140	0.170	0.2	0.05	0.123	0.224	0.184
0.2	0.01	0.046	0.047	0.053	0.2	0.01	0.036	0.090	0.065
0.3	0.05	0.254	0.250	0.298	0.3	0.05	0.201	0.335	0.282
0.3	0.01	0.111	0.106	0.130	0.3	0.01	0.069	0.154	0.118
0.5	0.05	0.356	0.361	0.402	0.3	0.05	0.448	0.600	0.532
0.5	0.01	0.158	0.154	0.171	0.3	0.01	0.210	0.359	0.298
0.7	0.05	0.569	0.571	0.602	0.3	0.05	0.681	0.798	0.742
0.7	0.01	0.253	0.247	0.271	0.3	0.01	0.422	0.581	0.508

Table 3

Numbers of identified differentially expression genes based on LaPoint *et al.* (2004) prostate cancer microarray data.

α	T	T ^{mod}	T ^{QL}	M ^{semi}
0.05	625	619	624	655
0.01	429	413	427	449