




PLEMT: A Novel Pseudolikelihood Based EM Test For Homogeneity In Generalized Exponential Tilt Mixture Models

Chuan Hong, Yong Chen, Yang Ning, Shuang Wang, Hao Wu & Raymond J. Carroll


To cite this article: Chuan Hong, Yong Chen, Yang Ning, Shuang Wang, Hao Wu & Raymond J. Carroll (2017): PLEMT: A Novel Pseudolikelihood Based EM Test For Homogeneity In Generalized Exponential Tilt Mixture Models, Journal of the American Statistical Association, DOI: [10.1080/01621459.2017.1280405](https://doi.org/10.1080/01621459.2017.1280405)

To link to this article: <http://dx.doi.org/10.1080/01621459.2017.1280405>

 View supplementary material 

 Accepted author version posted online: 27 Feb 2017.

 Submit your article to this journal 

 Article views: 184

 View Crossmark data 

**PLEMT: A NOVEL PSEUDOLIKELIHOOD BASED EM TEST FOR
HOMOGENEITY IN GENERALIZED EXPONENTIAL TILT MIXTURE
MODELS**

CHUAN HONG

*Department of Biostatistics,
Harvard School of Public Health, Boston, MA, USA*

YONG CHEN*

*Department of Biostatistics and Epidemiology,
University of Pennsylvania, Philadelphia, PA, USA*

YANG NING

*Department of Statistical Science,
Cornell University, Ithaca, NY, USA*

SHUANG WANG

*Department of Biostatistics,
Mailman School of Public Health, Columbia University, New York, NY, USA*

HAO WU

*Department of Biostatistics and Bioinformatics,
Rollins School of Public Health, Emory University, Atlanta, GA, USA*

RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, College Station, TX, USA

Authors' Footnote: Chuan Hong is Post Doctoral Fellow, Department of Biostatistics, Harvard Univer-

*Correspondence to: Yong Chen, Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, 210 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, USA. E-mail: ychen123@mail.med.upenn.edu. TEL: (215) 746-8155

ACCEPTED MANUSCRIPT

sity School of Public Health, Boston, MA 02115, USA (chong@hsph.harvard.edu). Yong Chen is Assistant Professor, Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA 19104, USA (ychen123@mail.med.upenn.edu). Yong Chen was supported by grant number R03HS022900 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. Yang Ning is Assistant Professor, Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA (yn265@cornell.edu). Shuang Wang is Associate Professor, Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10027, USA (sw2206@columbia.edu). Hao Wu is Assistant Professor, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA (hao.wu@emory.edu). Raymond J. Carroll is Distinguished Professor of Statistics, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA (carroll@stat.tamu.edu).

Abstract

Motivated by analyses of DNA methylation data, we propose a semiparametric mixture model, namely the generalized exponential tilt mixture model, to account for heterogeneity between differentially methylated and non-differentially methylated subjects in the cancer group, and capture the differences in higher order moments (e.g. mean and variance) between subjects in cancer and normal groups. A pairwise pseudolikelihood is constructed to eliminate the unknown nuisance function. To circumvent boundary and non-identifiability problems as in parametric mixture models, we modify the pseudolikelihood by adding a penalty function. In addition, the test with simple asymptotic distribution has computational advantages compared with permutation-based test for high-dimensional genetic or epigenetic data. We propose a pseudolikelihood based expectation–maximization test, and show the proposed test follows a simple chi-squared limiting distribution. Simulation studies show that the proposed test controls Type I errors well and has better power compared to several current tests. In particular, the proposed test outperforms the commonly used tests under all simulation settings considered, especially when there are variance differences between two groups. The proposed test is applied to a real data set to identify differentially methylated sites between ovarian cancer subjects and normal subjects.

Key words: Asymptotics; Conditional likelihood; Non-regular problem; Penalized likelihood; Semiparametric mixture model

1. INTRODUCTION

DNA methylation plays an important role in the development of many types of cancer. To identify differentially methylated Cytosine-Phosphate-Guanine (CpG) sites between cancer and normal subjects is one of the central tasks to understand contributions of the DNA methylation process on cancer development. Usually, cancer subjects are more heterogeneous in terms of DNA methylation distribution as cancer subjects may have different subtypes of cancer, different stages of cancer, and different history of treatment (Mikeska et al., 2010). Thus, DNA methylation levels of some cancer subjects may follow one distribution and are differentially methylated compared to normal subjects, while the rest of cancer subjects may follow a similar distribution as that of normal subjects and are not differentially methylated. The epigenetic heterogeneity in cancer has gained tremendous interest lately (Brocks et al., 2014; Oakes et al., 2014; Easwaran et al., 2014). CpG sites with high variability among cancer samples can potentially be used as epigenetic biomarkers for determining the stage of cancer progression and designing personalized treatment. Most of the existing methods for DNA methylation data focus on testing for differences in means between the cancer and normal groups, which does not fully capture the differences in variances in DNA methylation data. There is evidence that there are not only differences in DNA methylation means but also differences in DNA methylation variations between the cancer and normal groups (Hansen et al., 2011; Gervin et al., 2011). A recently proposed method DiffVar (Phipson and Oshlack, 2014) tests the equality of variances in two groups by performing a t -test on the absolute or squared deviations of the methylation levels from the group mean. It is, however, restricted in comparing the variances and cannot detect mean differences in two groups. More importantly, most of the existing methods for DNA methylation data are distribution based methods, including the logit-normal mixture model (Siegmund et al., 2004), the beta mixture model (Houseman et al., 2008), the uniform-truncated-normal-uniform mixture model (Wang, 2011), and a GLM based method (Ahn and Wang, 2013). However, due to the heterogeneity in distributions of DNA methylation across loci (Huang et al., 2013), it is insufficient to assume a parametric distribution for all loci. Fitting site-specific parametric models may not be feasible for a large number of loci, and also leads to difficulties in model interpretations.

To relax parametric model assumptions, exponential tilt mixture models (ETMM) have been considered (Qin, 1999; Zou et al., 2002; Tan, 2009). Specifically, subjects under one condition are sampled from a population with the baseline density function $f(x)$, and subjects under the other condition are sampled from a mixture population with the density function $h(x)$. The density $h(\cdot)$ and the relationship between the densities $f(\cdot)$ and $g(\cdot)$ can be formulated as follows,

$$\begin{aligned} h(x) &= (1 - \lambda)f(x) + \lambda g(x); \\ \log\{g(x)/f(x)\} &= \alpha + \beta x, \end{aligned} \tag{1}$$

where $g(\cdot)$ is defined as the density function of the methylation levels for the subpopulation in the case group that are differentially methylated, λ is an unknown mixture proportion parameter, β is an unknown parameter and $\alpha = -\log\{\int_{-\infty}^{\infty} \exp(\beta x)f(x)dx\}$ is a normalizing constant for the density function $g(x)$. Note that $\beta = 0$ implies $\alpha = 0$. Model (1) contains many parametric models as special cases, such as the mixture of normal distributions with different means but equal variances, and the mixture of gamma distributions with different shape parameters but equal scale parameters. Under the ETMM assumption, testing for homogeneity between cancer and normal groups i.e., $f(\cdot) = h(\cdot)$, is equivalent to testing $\lambda = 0$ or $\beta = 0$.

It has been long recognized that testing for homogeneity in mixture models is a non-regular problem because the mixture proportion parameter λ lies on the boundary of its parameter space $[0, 1]$ and the parameters α and β are not identifiable when $\lambda = 0$. Thus, the asymptotic distributions of tests for homogeneity are usually rather complicated and possibly dependent on the parametric distributions assumed (Davies, 1977, 1987). Recently, Qin and Liang (2011) derived a original score test under model (1) with a simple limiting chi-squared distribution. More recently, Liu et al. (2012) proposed a novel modified empirical likelihood ratio test under model (1) and developed an efficient and intuitive expectation–maximization (EM) algorithm for computing the test statistic. Despite the current success on the test of homogeneity in ETMM, the aforementioned methods only allow a scalar parameter β , which excludes important parametric distributions such as normal distributions with unequal means and unequal variances, gamma distributions with different shape and scale parameters, and beta distributions. As recent studies have observed that

cancer tissues and some complex disease cases can also be characterized by an increased variability in DNA methylation patterns (Hansen et al., 2011; Issa, 2011; Teschendorff et al., 2012; Xu et al., 2013), tests that ignore this feature may lead to a substantial loss of power. We therefore extended both the score test by Qin and Liang (2011) and the modified empirical likelihood ratio test by Liu et al. (2012) by generalizing x to $\mathbf{k}(x) = (x, x^2)$ in equation (1) to account for differences in both means and variances. However, as the simulation results summarized in Section 2 of online supplementary materials suggested, both the extended score test and the extended modified empirical likelihood ratio test have inflated Type I errors. This suggests that an alternative approach should be considered.

In this paper, we generalize the one-parameter ETMM to a multi-parameter ETMM, namely the generalized exponential tilt mixture model (GETMM), which aims to capture the differences in higher order moments between two distributions. Specifically, the right handside of equation (1) is extended to a general form of $\alpha + \beta^T \mathbf{k}(x)$, so that the multi-parameter ETMM includes many parametric models, such as the normal mixture model with unequal variances, the gamma mixture model, and the beta mixture model. Rather than estimating the baseline density function $f(\cdot)$ with the empirical likelihood procedure as in Qin and Liang (2011) and Liu et al. (2012), we construct a novel pseudolikelihood based on a conditioning procedure, which eliminates the baseline density function $f(\cdot)$ and avoids its estimation. To handle the non-regularity problems (i.e. boundary and non-identifiability problems), we construct a penalized pseudolikelihood where the impacts of the tuning parameter are studied. Finally, we propose an EM algorithm based test for computational efficiency and stability, which can be shown to follow a simple chi-squared limiting distribution.

The contributions of this work are three-fold. First, we develop a semiparametric model that captures the differences in higher moments between distributions. Second, we construct a novel penalized pseudolikelihood, where the unknown baseline density function $f(\cdot)$ is eliminated and the non-regularity problems are circumvented. Third, we propose an EM algorithm based test with a simple chi-squared limiting distribution, which is computationally efficient and stable. The pseudolikelihood EM algorithm has been proposed for handling spatial data (Varin et al., 2005), hidden Markov model (Gao and Song, 2011), and family data with multistage sampling (Choi and

Briollais, 2011). The convergence property of the EM algorithm is established by Gao and Song (2011). Unlike these existing results, the estimated parameters at each iteration of EM algorithm rather than the estimated stationary point are used to construct the proposed test.

This paper is organized as follows. Section 2 describes the penalized pseudolikelihood based EM test (hereafter referred to as the PLEMT test). The asymptotic null distribution and the local asymptotic power for the PLEMT test are provided in Section 3. Simulation studies comparing Type I errors and power of the PLEMT test with existing tests are summarized in Section 4. A real data application to DNA methylation data for ovarian cancer is given in Section 5 followed by a brief discussion in Section 6. Proofs are relegated to Appendices in online supplementary materials.

2. STATISTICAL METHODOLOGY

We propose the following two-group generalized exponential tilt mixture model (GETMM). We present our model in the setting of modeling DNA methylation data for the concreteness of interpretation, while noting that it can be generally applied to any two-group testing problem for homogeneity. At the ℓ^{th} CpG site, let $u_{\ell 1}, \dots, u_{\ell n_0}$ be independent, identically distributed (i.i.d.) DNA methylation levels in the normal group with distribution $f_{\ell}(u)$, where n_0 is the number of normal subjects. It is believed that in the cancer group, only a proportion of subjects are methylated differentially compared to those in the normal group, known as non-homogeneity or heterogeneity in methylation among cancer subjects (Kalari, 2010). Moreover, the effect of differential methylation may appear as changes in variation, in addition to potential a shift in means (Hansen et al., 2011; Gervin et al., 2011; Teschendorff et al., 2012; Xu et al., 2013). To account for such features of DNA methylation data, we assume that the i.i.d subjects $v_{\ell 1}, \dots, v_{\ell n_1}$ in the cancer group follow a mixture distribution with the density $h_{\ell}(v)$ as follows

$$h_{\ell}(v) = (1 - \lambda_{\ell})f_{\ell}(v) + \lambda_{\ell}g_{\ell}(v),$$

where n_1 is the number of cancer subjects, λ_{ℓ} is an unknown mixture proportion parameter ($0 \leq \lambda_{\ell} \leq 1$), and the density functions $f_{\ell}(v)$ and $g_{\ell}(v)$ are related through a multi-parameter exponential

tilt

$$\log\{g_\ell(v)/f_\ell(v)\} = \alpha_\ell + \boldsymbol{\beta}_\ell^T \mathbf{k}(v). \quad (2)$$

Here $\boldsymbol{\beta}_\ell = (\beta_{\ell 1}, \dots, \beta_{\ell d})^T$ is a d -dimensional vector of unknown parameters, $\mathbf{k}(v) = \{k_1(v), \dots, k_d(v)\}^T$ is a vector of pre-specified functions of v , and

$\alpha_\ell = -\log \left[\int \exp\{\boldsymbol{\beta}_\ell^T \mathbf{k}(v)\} f_\ell(v) dv \right]$ is a normalizing constant. It is easy to see that $\boldsymbol{\beta}_\ell = \mathbf{0}$ implies $\alpha_\ell = 0$. For simplicity of notation, we hereafter suppress the site index ℓ . We acknowledge that the baseline density function $f_\ell(\cdot)$ can be site-specific and is left completely unspecified. The parameter λ_ℓ can also be site-specific. Note that the GETMM includes many parametric mixture models. When $\mathbf{k}(v) = v$, the GETMM reduces to the one-parameter ETMM described in equation (1); when $\mathbf{k}(v) = (v, v^2)$, the GETMM includes the normal mixture model with unequal variances; when $\mathbf{k}(v) = \{\log(v), \log(1 - v)\}$, the GETMM includes the beta mixture model. Both parametric models have been used to model DNA methylation data (Siegmond et al., 2004; Houseman et al., 2008).

Since the majority of differences between the cancer and normal groups may be contained in means and variances of methylation levels, we consider the GETMM with two parameters for model parsimony. While the GETMM with more than two parameters may better capture the differences in higher moments such as skewness and kurtosis, the corresponding tests may be underpowered due to the larger degree of freedom. In addition, the theoretical development of GETMM with more than two parameters is similar. Specifically, we let $\mathbf{k}(v) = (v, v^2)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)$. Under this model, testing for homogeneity between two groups is equivalent to testing

$$H_0 : \lambda = 0 \text{ or } \beta_1 = \beta_2 = 0.$$

Maximizing the likelihood function generally involves estimating the baseline density function $f(\cdot)$, typically by an empirical likelihood procedure (Owen, 1988). Here, we construct a pairwise pseudolikelihood, which eliminates $f(\cdot)$ by conditioning on order statistics. Specifically, the advantage of the conditioning procedure is to avoid the estimation of $f(\cdot)$. We consider a pair of observations from the two groups, i.e., u_i from the normal group and v_j from the cancer group. The conditional density of (u_i, v_j) given their order statistics $t^{(1)} = \min(u_i, v_j)$ and $t^{(2)} = \max(u_i, v_j)$

can be calculated as,

$$\text{pr}(u_i, v_j | t^{(1)}, t^{(2)}) = \left\{ 1 + R(u_i, v_j; \lambda, \alpha, \beta_1, \beta_2) \right\}^{-1}, \quad (3)$$

where

$$R(u_i, v_j; \lambda, \alpha, \beta_1, \beta_2) = \frac{(1 - \lambda) + \lambda \exp(\alpha + \beta_1 u_i + \beta_2 u_i^2)}{(1 - \lambda) + \lambda \exp(\alpha + \beta_1 v_j + \beta_2 v_j^2)}.$$

The derivation of equation (3) is provided in Appendix A of online supplementary materials. The baseline density function $f(\cdot)$ is eliminated through this conditioning procedure. This idea of conditioning was originally proposed by [Kalbfleisch \(1978\)](#) for rank tests and permutation tests in regression problems, and later revitalized by [Liang and Qin \(2000\)](#) in regression analyses under biased sampling.

For each pair of observations (u_i, v_j) , we can calculate the pairwise conditional density. We then multiply all these densities together and obtain the following log pseudolikelihood function for all observations,

$$\mathcal{L}_p(\lambda, \alpha, \beta_1, \beta_2) = \frac{2}{n} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} -\log \left\{ 1 + R(u_i, v_j; \lambda, \alpha, \beta_1, \beta_2) \right\},$$

where $n = n_0 + n_1$. We note that the sum of log conditional densities is standardized by the total number of individuals rather than the number of summands. This is owing to the projection theory in U-statistics ([Lehmann and D'Abrera, 1975](#)). As we will show later, such modification of the loglikelihood is necessary so that the proposed test has a simple χ^2 limiting distribution with 2 degrees of freedom.

Under the null hypothesis, $\lambda = 0$ lies at the boundary of its parameter space, which leads to a boundary problem ([Self and Liang, 1987](#); [Chen and Liang, 2010](#)). Furthermore, the null hypothesis holds for $\lambda = 0$ regardless of the values of α , β_1 and β_2 , and holds for $\beta_1 = \beta_2 = 0$ regardless of the value of λ . This implies that the parameter $(\lambda, \alpha, \beta_1, \beta_2)$ is not identifiable under H_0 , which results in complicated asymptotic properties of the pseudolikelihood ratio function ([Chen and Chen, 2001](#); [Zhu and Zhang, 2004](#)). To deal with the non-identifiability problem, [Qin and Liang \(2011\)](#) fixed the value of λ at 1, so that the other parameters (α and β) are identifiable and can be estimated by maximizing the empirical likelihood. The score test statistic was then constructed. However, the choice of the value for the fixed λ is arbitrary, and the performance of the score test depends on the

choice of λ . Alternatively, [Liu et al. \(2012\)](#) proposed an empirical likelihood function by adding penalty on λ . Rather than using the empirical likelihood, we propose the following penalized pseudolikelihood function to avoid the boundary and identifiability problems

$$\mathcal{L}_{pp}(\lambda, \alpha, \beta_1, \beta_2) = \mathcal{L}_p(\lambda, \alpha, \beta_1, \beta_2) + C \log(\lambda),$$

where C is a positive number. The penalty is heavy when λ is close to 0 and less so when λ approaches 1. The parameter C is a multiplicative factor of such penalty and is often termed as the tuning parameter. By using the penalized pseudolikelihood, the parameter λ is bounded away from 0, and the null hypothesis is then reduced to $\beta_1 = \beta_2 = 0$. That is, β_1 and β_2 in the penalized pseudolikelihood function is asymptotically identifiable. When $\mathbf{k}(x) = x$, it has been recommended in the literature that C can be taken as 1, and the existing tests based on penalizing λ are not sensitive to the choice of C ([Chen and Chen, 2001](#); [Fu et al., 2006a](#)). The GETMM we considered involves both mean and variance, e.g., $\mathbf{k}(x) = (x, x^2)$, and the recommendation on the choice of C may be different. In our simulation studies, we investigate the impact of C on the performance of the proposed test. [Di and Liang \(2011\)](#) have investigated the impacts of tuning parameter C on type I errors and power. They suggested to use the smallest C that still provides the correct type I error rate. However, the admixture model considered in [Di and Liang \(2011\)](#) is different from the model considered in this paper. Sensitivity analyses on the choice of C have been conducted and summarized in the Supplementary Materials. In general, $C = 20$ is recommended based on various scenarios considered in our sensitivity analyses.

In general, the parameter λ and the parameters $(\alpha, \beta_1, \beta_2)$ are highly intertwined with each other as in many mixture models ([Bandeem-Roche et al., 1997](#)). This can lead to numerical problems such as multiple local maxima when maximizing the pseudolikelihood. Our simulation results summarized in Table 7 of the Supplementary Materials show that the pseudolikelihood method that simultaneously maximizes over all parameters faces the problems of unstable results and inflated Type I errors. In our proposed PLEMT test, we use a set of different initial values of λ to avoid being trapped at a local maximum.

The key idea of the PLEMT procedure is that maximizing the penalized pseudolikelihood with

respect to $(\alpha, \beta_1, \beta_2)$ for a fixed λ is more stable than maximizing over all parameters simultaneously. In DNA methylation data, consider DNA methylation levels v_1, \dots, v_{n_1} in the cancer group are sampled from a mixture of $f(\cdot)$ and $g(\cdot)$. That is, v_1, \dots, v_{n_1} are i.i.d. with a density function $h(v) = (1 - \lambda)f(v) + \lambda g(v)$. The information on whether the DNA methylation level of a particular subject in the cancer group is from the subpopulation $f(\cdot)$ or $g(\cdot)$ is considered as missing data.

Here we describe the algorithm to calculate the PLEMT test statistic. We first choose a finite set of $\Lambda = \{\lambda_1, \dots, \lambda_Z\}$, where Z is the number of points in the grid (e.g., $\Lambda = \{0.1, \dots, 1\}$). We then choose the number of iterations S of the EM algorithm. Although the PLEMT test is motivated by the EM algorithm, the calculation of the test statistic does not require the convergence of the EM algorithm, which has been shown in Appendix D of online supplementary materials. Generally, only a few steps of iterations is needed. This feature offers great computational advantages in analyses of high dimensional data. Sensitivity analyses in the Supplementary Materials suggest that the performance of the proposed PLEMT test is not sensitive to the choices of S and Z . Here we choose $Z = 10$, and $S = 3$. At the z^{th} grid value λ_z , we set the initial value $\lambda_z^{(1)} = \lambda_z$, and calculate the initial values of $(\alpha_z^{(1)}, \beta_{z1}^{(1)}, \beta_{z2}^{(1)})$ by maximizing $\mathcal{L}_{pp}(\lambda_z^{(1)}, \alpha, \beta_1, \beta_2)$. We carry out the following EM algorithm for $S - 1$ times.

At the E step of the EM algorithm, we calculate the posterior probability of the j^{th} cancer subject being differentially methylated given v_j and $(\lambda_z^{(s)}, \alpha_z^{(s)}, \beta_{z1}^{(s)}, \beta_{z2}^{(s)})$ as,

$$\omega_{jz}^{(s)} = \frac{\lambda_z^{(s)} g(v_j)}{(1 - \lambda_z^{(s)})f(v_j) + \lambda_z^{(s)} g(v_j)} = \frac{\lambda_z^{(s)} \exp(\alpha_z^{(s)} + \beta_{z1}^{(s)} v_j + \beta_{z2}^{(s)} v_j^2)}{1 - \lambda_z^{(s)} + \lambda_z^{(s)} \exp(\alpha_z^{(s)} + \beta_{z1}^{(s)} v_j + \beta_{z2}^{(s)} v_j^2)}.$$

We then calculate the expected complete likelihood given the data and the current parameter estimates, which only involves the parameter λ

$$\sum_{j=1}^{n_1} (1 - \omega_{jz}^{(s)}) \log(1 - \lambda) + \sum_{j=1}^{n_1} \omega_{jz}^{(s)} \log(\lambda) + C \log(\lambda).$$

At the M step, we update λ and other parameters (α , β_1 , and β_2). Specifically,

$$\begin{aligned}\lambda_z^{(s+1)} &= \operatorname{argmax}_{\lambda} \sum_{j=1}^{n_1} (1 - \omega_{jz}^{(s)}) \log(1 - \lambda) + \sum_{j=1}^{n_1} \omega_{jz}^{(s)} \log(\lambda) + C \log(\lambda) \\ &= \frac{\sum_{j=1}^{n_1} \omega_{jz}^{(s)} + C}{n_1 + C},\end{aligned}\tag{4}$$

$$(\alpha_z^{(s+1)}, \beta_{z1}^{(s+1)}, \beta_{z2}^{(s+1)}) = \operatorname{argmax}_{\alpha, \beta_1, \beta_2} \mathcal{L}_{pp}(\lambda_z^{(s+1)}, \alpha, \beta_1, \beta_2).$$

Equation (4) suggests an intuitive explanation of the penalty term $C \log(\lambda)$ in $\mathcal{L}_{pp}(\lambda, \alpha, \beta_1, \beta_2)$. Specifically, the proportion of the subgroup with differential methylation among cases, $\lambda_z^{(s+1)}$, is calculated as the average of posterior probabilities of being in this subgroup, $\omega_{jz}^{(s)}$, plus C pseudo observations (known to be in this subgroup). This idea of pseudo-observation adjustment was originally proposed by Jiahua Chen and his colleagues. The pseudolikelihood method that maximizing over all parameters simultaneously faces the non-convergence and computational problems. Our simulation result summarized in the Supplementary Materials shows that the parameter estimates are sensitive to the choice of initial values when maximizing over all parameters simultaneously, leading to unstable results. By fixing λ at this step, we can get more stable results than maximizing over all parameters simultaneously. We use the general-purpose optimization procedure in R (R Development Core Team, 2009) to obtain $(\alpha_z^{(s+1)}, \beta_{z1}^{(s+1)}, \beta_{z2}^{(s+1)})$. At the z^{th} grid value, we define $M_n(\lambda_z^{(S)}) = 2 \{ \mathcal{L}_{pp}(\lambda_z^{(S)}, \alpha_z^{(S)}, \beta_{z1}^{(S)}, \beta_{z2}^{(S)}) - \mathcal{L}_{pp}(1, 0, 0, 0) \}$. We then define our PLEMT test statistic as follows,

$$\text{PLEMT} = \max \{ M_n(\lambda_z^{(S)}), z = 1, \dots, Z \}.$$

Details of the derivation of the PLEMT test statistic are provided in Appendix D of online supplementary materials.

3. ASYMPTOTIC RESULTS

This section provides asymptotic results for the PLEMT test statistic under the null H_0 and under the local alternatives.

We assume the following regularity conditions.

(C1). The parameter sets Ω_α and Ω_β for α and β are compact.

(C2). The distributions of u_i and v_j have common support and are not degenerate to a point measure.

(C3). The ratio $n_1/n \rightarrow \rho$, as $n \rightarrow \infty$, where $0 < \rho < 1$. The variance $\sigma^2 = \text{var}(u_i) < \infty$ and for some $t > M$, $\int u^2 \exp(t|u|)f(u)du < \infty$, where M is a small positive constant.

The compactness of the parameter spaces in assumption (C1) is commonly adopted in the statistical literature. Such an assumption may be relaxed by imposing the uniform boundedness assumption on the baseline density; for example, see Li et al. (2009). The assumption (C2) is to guarantee the GETMM is identifiable. The assumption (C3) is a reasonable technical condition for applying the uniform law of large numbers.

Theorem 1 *Under the regularity conditions C1–C3, and the null hypothesis H_0 , the PLEMT test statistic converges weakly to χ_2^2 as $n \rightarrow \infty$.*

The proof of Theorem 1 is given in Appendix D of online supplementary materials.

Evaluation of test statistics under alternative hypotheses is crucial for sample size calculation and experimental design. In practice, the most interesting situations for alternatives are those close to the null hypothesis. For mathematical convenience, statisticians typically focus on the local asymptotic power of test statistics. Here we provide a useful result about the asymptotic power of the PLEMT test under a sequence of local alternatives. Specifically, for any $0 < \lambda_0 < 1$, density function $f_0(\cdot)$ and a fixed τ_0 , consider a sequence of alternatives:

$$H_a : \lambda = \lambda_0, f(\cdot) = f_0(\cdot), \beta = n^{-1/2}\tau_0,$$

where τ_0 is a vector. Under H_a , we do not explicitly specify the local alternative for α , because α is a function of β and $f_0(\cdot)$. With LeCam's third lemma (Van der Vaart, 2000), we can establish the following results.

Theorem 2 *Under the alternatives H_a , the limiting distribution of the PLEMT test statistic is $\chi_2^2\{\lambda_0^2(1-\rho)\rho\tau_0^T\Sigma\tau_0\}$, where $\chi_2^2(c)$ denotes the noncentral chi-squared distribution with 2 degrees of freedom and non-centrality parameter c .*

The proof is given in Appendix E of online supplementary materials. An important observation is, given a total number of subjects n , the asymptotic power is maximized when the design is balanced (i.e., $\rho = 0.5$ or $n_0 = n_1$).

4. SIMULATION STUDIES

We conduct simulation studies to evaluate the finite sample performance of the proposed PLEMT test comparing with seven existing tests, namely the score test based on the one-parameter ETMM, the modified empirical likelihood ratio test, the Wald test based on logistic regression, the t -test, the Wilcoxon test, the F test of equality of variances and the Kolmogorov-Smirnov test.

We consider a variety of parametric models. Specifically, for each simulation setting, the data are generated from the mixture model (2) with one of the following choices of density functions $f(\cdot)$ and $g(\cdot)$.

Model A (Normal model). Let $f(\cdot)$ and $g(\cdot)$ be the density functions of Normal (μ_1, σ_1^2) and Normal (μ_2, σ_2^2) , respectively. Then

$$\log\{g(x)/f(x)\} = \frac{1}{2}(\log \sigma_1^2 - \log \sigma_2^2) + \frac{\sigma_2^2 \mu_1^2 - \sigma_1^2 \mu_2^2}{2\sigma_1^2 \sigma_2^2} + \frac{\sigma_1^2 \mu_2 - \sigma_2^2 \mu_1}{\sigma_1^2 \sigma_2^2} x + \frac{\sigma_2^2 - \sigma_1^2}{\sigma_1^2 \sigma_2^2} x^2.$$

Model B (Beta model). Let $f(\cdot)$ and $g(\cdot)$ be the density functions of two beta distributions with shape parameters (a_1, b_1) and (a_2, b_2) , respectively. Then

$$\log\{g(x)/f(x)\} = \log \left\{ \frac{B(a_1, b_1)}{B(a_2, b_2)} \right\} + (a_2 - a_1) \log x + (b_2 - b_1) \log(1 - x),$$

where $B(\cdot, \cdot)$ is the beta function.

Model C (Gamma model). Let $f(\cdot)$ and $g(\cdot)$ be the density functions of Gamma (m_1, θ_1) and Gamma (m_2, θ_2) with shape parameters m_1, m_2 and scale parameters $\theta_1, \theta_2 > 0$. Then

$$\log \frac{g(x)}{f(x)} = \log \left\{ \frac{\Gamma(m_1)}{\Gamma(m_2)} \right\} + m_1 \log(\theta_1) - m_2 \log(\theta_2) + (m_2 - m_1) \log x + \left(\frac{1}{\theta_1} - \frac{1}{\theta_2} \right) x.$$

Note that Models A–C belong to the GETMM. In equation (2), Models A–C have $\mathbf{k}(x) = (x, x^2)$, $\{\log(x), \log(1 - x)\}$, and $\{\log(x), x\}$, respectively. To evaluate the robustness of the proposed test to model misspecifications, we consider two additional models (Models D–E) when the expo-

nential tilt model assumption is not satisfied. It is expected that misspecified models may lead to incorrect Type I errors for tests based on the GETMM assumption.

Model D (Negative binomial model). Let $f(\cdot)$ and $g(\cdot)$ be the density functions of the Negative Binomial (NB) distribution (r_1, p_1) and NB (r_2, p_2) , where r_1 and r_2 are the numbers of failures until the experiment is stopped, and p_1 and p_2 are success probabilities in each experiment. Then

$$\log \frac{g(x)}{f(x)} = r_2 \log\left(\frac{p_2}{1-p_2}\right) - r_1 \log\left(\frac{p_1}{1-p_1}\right) + x \log\left(\frac{1-p_2}{1-p_1}\right) + \log\binom{x-1}{r_2-1} - \log\binom{x-1}{r_1-1}.$$

Model E (t distribution). Let $f(\cdot)$ and $g(\cdot)$ be the density functions of the t distributions (ncp_1, df_1) and (ncp_2, df_2) , where ncp_1 and ncp_2 are the noncentrality parameters, and df_1 and df_2 are degrees of freedom.

We compare Type I errors and power of the tests. For power comparisons, we conduct simulation studies under Scenario I where $f(\cdot)$ and $g(\cdot)$ are different in means only, Scenario II where $f(\cdot)$ and $g(\cdot)$ are different in variances only, and Scenario III where both means and variances of $f(\cdot)$ and $g(\cdot)$ are different. For each of the power scenarios, we consider settings where λ takes different values from 0 to 1. The rejection rates based on 5000 simulations are used to estimate Type I errors and the rejection rates based on 1000 simulations are used to estimate the power. We consider a sample size setting with 100 subjects in each group.

Table 1 summarizes the Type I errors of the eight tests under comparison for five models (A–E). Under both non-misspecification scenario (Models A–C) and misspecification scenario (Models D–E), the Type I errors of the PLEMT test, the EST test, the MELRT test, the t -test, the Wilcoxon test and the Logistic regression test are relatively close to the corresponding nominal levels given the moderate sample size ($n = 200$). The proposed test has slightly inflated type I errors under model misspecifications compared to t -test, Wilcoxon test and logistic regression test. It is not surprising to see that the Type I errors of the F test are inflated under the Gamma, Negative binomial and the t models, because the F test is known to be sensitive to non-normality. The Kolmogorov-Smirnov test yields conservative Type I errors under all settings, especially in the negative binomial model.

Since the Type I errors of the F test are inflated, we only compare the power of the remaining

seven tests. Figure 1 plots the power curves of the seven tests under GETMMs (Models A–C). As λ increases, the power of all tests increases. For power Scenario I with mean differences only, the PLEMT test is slightly more powerful than the Kolmogorov-Smirnov test, and has slightly lower but comparable power than the rest five tests. This is because the PLEMT test has two degrees of freedom, while the other five tests have only one degree of freedom. The slight loss of power for the PLEMT test is due to the extra one degree freedom when there are only mean differences. For the power Scenario II with variance differences only, the PLEMT test is much more powerful than all the other tests as expected since they do not fully account for variance differences and the mixture structure of the data. More specifically, the MELRT test has about 17% – 46% less power than that of the PLEMT test, the Kolmogorov-Smirnov test has about 40% – 90% less power than that of the PLEMT test, whereas the other four tests have essentially no power beyond Type I errors. This is consistent with the simulation results in [Liu et al. \(2012\)](#) that the MELRT test is more powerful than the EST test. When both mean and variance are different, the PLEMT test remains to be the most powerful one. The degree of the power loss of the other six tests depends on the proportion of the variance difference in the overall mean and variance differences. Under the setting we considered, the MELRT test is the second most powerful test with a power about 5% – 25% lower than that of the PLEMT test. The other four tests are grossly underpowered.

Similar patterns are observed in Figure 2 for misspecified models (Models D and E). The PLEMT test has a slightly lower power than the other tests under power Scenario I, while it is the most powerful test under power Scenarios II and III. One interesting phenomena is that there is a power gain for the Wilcoxon test under Scenario II and III for the misspecified models comparing with that under GETMMs, while the Wilcoxon test has essentially no power beyond Type I errors under GETMMs. This is because the Wilcoxon test may capture some of the variance differences when the distributions are heavily skewed, but cannot capture these differences when the distributions are symmetric.

In summary, our simulation studies suggest performance from the proposed PLEMT test compared to the existing ones. The proposed PLEMT test has well controlled Type I errors and substantial power gain when the variance differences need to be taken into account. The proposed

PLEMT test also shows some degree of robustness under model misspecifications. The PLEMT test is implemented as an R software package *robustETM*, which is attached as the Supplementary Materials.

5. APPLICATION TO DNA METHYLATION DATA OF OVARIAN CANCER

We apply the PLEMT test to the data from the United Kingdom Ovarian Cancer Population Study to select differentially methylated sites between ovarian cancer cases and age-matched healthy controls using the Illumina Infinium Human Methylation27 Beadchip (Teschendorff et al., 2010). The original data have 266 ovarian cancer cases with 131 pre-treatment cases and 135 post-treatment cases, and 274 age-matched healthy controls. Since age and having received treatment or not when blood samples are taken are factors known to affect DNA methylation levels, we choose to use the 131 ovarian cancer cases who gave their blood at the time of their diagnosis prior to treatment and with age-matched controls. We refer readers to Wang (2011) for the detailed quality control steps. We end up with 96 cancer subjects and 136 normal subjects with DNA methylation levels at 22951 sites. Because our simulation results suggest that the MELRT test is the second most powerful test accounting for the variance differences, we focus on the comparisons among the PLEMT test, the commonly used t -test, and the MELRT test in this real data application. We also only focus on the original DNA methylation levels instead of the logit transformed ones. Due to the presence of multiple hypothesis testing, the number of false positives may rapidly increase such that the scientific discoveries may become unreliable. To address this problem, we adapt the procedure in Storey (2002) to control the false discovery rate (FDR), which is defined as the number of false positives divided by the number of total discoveries. In particular, for each hypothesis test, we calculate a number called q -value, which is the minimum FDR that can be attained when the test is significant. It can be regarded as a hypothesis testing error measure for each test with respect to FDR (Storey, 2002). Here, we use the “ q value” package in R to calculate the q -value for each hypothesis test.

The proposed PLEMT test, the t -test and the MELRT test are then applied to the methylation data at these 22951 sites. Of the sites tested, 3112 sites have q -values < 0.05 using the PLEMT

test, 2699 sites have q -values < 0.05 using the t -test, and 2881 sites have q -values < 0.05 using the MELRT test. These numbers are cross-tabulated in Table 2. There are 2418 overlapping sites that have q -values < 0.05 using both the PLEMT test and the t -test, and 694 sites that are identified by the PLEMT test but not by the t -test. There are 2543 overlapping sites that have q -values < 0.05 using both the PLEMT test and the MELRT test, and 569 sites that were identified by the PLEMT test but not by the MELRT test.

We denote $\Delta = (m_1 - m_2)/sd_1$ as the standardized mean difference between cancer and normal subjects, and $r_{21} = sd_2/sd_1$ as the ratio of standard deviations between cancer and normal subjects, where m_1 and m_2 are the means of the normal subjects and cancer subjects, respectively, and sd_1 and sd_2 are the standard deviations of the normal subjects and cancer subjects, respectively. We further examine the distribution of Δ and r_{21} for the sites tested. The upper three panels of Figure 3 displays the distribution of Δ and r_{21} for the 694 sites that are identified by the PLEMT test but not the t -test, for the 281 sites that are identified by the t -test but not the PLEMT test, and for the 2418 overlapping sites that are identified by both the PLEMT test and the t -test, respectively. The lower three panels of Figure 3 displays the distributions of Δ and r_{21} for the 569 sites that are identified by the PLEMT test but not the MELRT test, for the 338 sites that are identified by the MELRT test but not the PLEMT test, and for the 2542 sites that are identified by both the PLEMT test and the MELRT test, respectively.

It is clear that the sites identified by the PLEMT test only but not the t -test have more significant variance differences than mean differences between the cancer and normal groups, which is the scenario the proposed test is designed for. In contrast, those sites that are identified by the t -test only have more significant mean differences than variance differences between the cancer and normal groups in general. For the overlapping 2418 sites that are identified by both the PLEMT test and the t -test, the majority have much larger differences in means than the sites identified by only one method. Thus these sites are relatively easier to be identified as all methods look for mean differences. Moreover, for those sites that have relatively small mean differences (but still have larger mean differences than sites that are identified by the PLEMT test only), they have large differences in variance in general. Thus the PLEMT test is able to identify them, although

simulation studies suggested a slightly lower power for the PLEMT test in such scenarios than the *t*-test. Similar patterns can be found when comparing the PLEMT test and the MELRT test. An interesting phenomenon is that for the 338 sites identified by the MELRT test only but not the PLEMT test (lower middle panel of Figure 3), there are 9 points between the clouds (highlighted in red), where the differences in both means and variances are very small. As shown later, our further examination suggests that those sites identified by the MELRT test may be false positive.

We examine the top 50 sites with most significant results from the PLEMT test among the 694 sites that are identified by the PLEMT test but not the *t*-test. By estimating the proportion of differentially methylated subjects in the cancer group compared to those in the normal group, we find the mixture feature at 16 out of the 50 sites (i.e., close to one-third) based on the estimated λ . For example, at site *cg26457013*, the proportion of differentially methylated subjects in the cancer group compared to those in the normal group is 22%, with estimated $(\beta_1, \beta_2) = (-86.79, -5.52)$. At site *cg11905589*, the proportion of differentially methylated subjects in the cancer group compared to those in the normal group is 78%, with estimated $(\beta_1, \beta_2) = (-10.87, -71.68)$. At both sites, there are sizable differences in means and variances. Specifically, we have $(\Delta, r_{21}) = (0.28, 0.76)$ at site *cg26457013*, and $(\Delta, r_{21}) = (0.11, 0.70)$ at site *cg11905589*. We further examine the 9 CpG sites between the clouds in the lower middle panel of Figure 3 with small differences in both means and variances that are identified by the MELRT test but not the PLEMT test, where we apply all eight tests investigated in the simulation studies on these 9 sites. The results suggest that some of the identified sites might have been false positive since all other six existing tests generate large *p*-values (results are included in Section 3 of online supplementary materials).

We compare the predictive power of significant CpG sites detected by PLEMT, MELRT and *t*-test. The top ranked CpG sites (by statistical significance) detected by the three methods are largely overlapped, because most of them show differences in average methylation levels. To emphasize the distinctions of different methods, we ignore the common ones and pick top 30 CpG sites uniquely identified by these three tests, and then use them as predictors to classify cancer and normal samples. We use Random Forest (Breiman, 2001) as the classification method, and compare the receiver operating characteristics (ROC) curves generated from 100 runs of 3-fold

cross validation. As shown in Figure 4, the area under the ROC curve of PLEMT is greater than that of MELRT and t -test, (AUC: 0.76, 0.66 and 0.64 for PLEMT, MELRT and t -test, respectively). This result indicates that compared with the other two methods, PLEMT identifies CpG sites that can better distinguish cancer patients from normal people.

We further investigate the 15 genes that the top 15 CpG sites reside among those CpG sites that were uniquely identified by the proposed PLEMT method only but not by the t -test or by the MELRT method. Out of the 15 genes, 9 of them have been reported to be associated with cancer. These include genes which are reported to be related with breast cancer: PMP22 (Winslow et al., 2013), AIM2 (Liu et al., 2015), colorectal cancer: CNGA3 (Shaikh et al., 2015), ovarian cancer HOXB8 (Stavnes et al., 2013), liver cancer HNF4A (Ning et al., 2010), pancreatic cancer PCDHB2 (Carter et al., 2010), and cancer progression STRN4 (Wong et al., 2014), CHCHD4 (Yang et al., 2012), and VHL (Kim and Kaelin, 2004), where the number in the parenthesis is the rank of p -value among the 15 genes.

In summary, the proposed PLEMT test has identified novel sites of potential interest that are missed by the commonly used t -test and the MELRT test, and has better predictive powers. Therefore, it can serve as a useful complement to the standard tests.

6. DISCUSSION

In this paper, we proposed a novel pseudolikelihood based EM test to identify differentially methylated loci. Specifically, we developed a semiparametric model to account for heterogeneity between differentially methylated subjects and non-differentially methylated subjects in the cancer group, and capture the differences in higher order moments (e.g. mean and variance) between subjects in the cancer and normal groups. We constructed a novel penalized pseudolikelihood to eliminate the unknown baseline density function and circumvent the non-regularity problems. We also proposed an EM algorithm based test for computational efficiency and stability, which follows a simple chi-squared limiting distribution. Through simulation studies we demonstrated the feasibility and power of the proposed test. The proposed test outperformed the existing tests especially when there is variance difference between two groups. We have also conducted sensitivity

analyses to empirically show that the results are not sensitive to the tuning parameters C , S and Z . Cross-validation procedures can be used here to obtain an optimal choice of tuning parameters. However, these procedures are usually computational expensive. Instead, C can be simply set at 20 for reasonable type I errors and power, as suggested from our sensitivity analyses.

The proposed PLEMT test with pairwise conditioning procedure has the advantage of eliminating the nuisance baseline density function $f(\cdot)$. However, the baseline density function may be assumed known in other scenarios. For example, an accurate estimate of the baseline density function could be obtained when the data of large size of normal subjects are available. In this case, there is no need to use the proposed conditioning procedure to eliminate the nuisance baseline function. Instead, a penalized likelihood ratio test for admixture model can be used to test for homogeneity (Fu et al., 2006b; Di and Liang, 2011). As suggested by a referee, we have conducted simulation studies to compare the performance of the proposed PLEMT test ($f(\cdot)$ unknown) with the penalized likelihood ratio test for admixture model ($f(\cdot)$ known); see the additional simulation results in the Supplementary Materials. We found that the penalized likelihood ratio test for admixture model has more power compared with the proposed PLEMT test, especially when the proportion parameter is relatively small.

In this paper, the sample sizes in normal and cancer groups are set to be equal. In some cancer data sets, there may be more cancer samples than normal samples, or more normal samples than cancer samples for relatively rare/under-studied cancer. We have conducted additional simulation studies when the two groups are unbalanced. The results for type I errors and power comparisons are summarized in Table 7 in the Supplementary Materials. The balance of two groups has some impacts on type I errors and permutation methods may be needed to better control type I errors if two groups are highly unbalanced.

We compare the type I errors of the proposed PLEMT test using χ^2 distribution with the type I errors of the permutation-based PLEMT test. The type I errors of the proposed PLEMT test and the permutation-based PLEMT test are similar. Due to the time cost of the permutation-based test, we suggest the use of the proposed PLEMT test with a simple χ^2 asymptotic distribution when the sample size is sufficiently large.

In this paper, the pre-specified kernel function $k(v) = (v, v^2)$ is considered for illustration. The misspecification of the kernel function does not lead to inflated type I errors. However, better choice of the kernel function may yield higher power. We are currently working on incorporating the Box-Cox transformation into the generalized exponential tilt model for more model flexibilities.

The proposed method aims to detect methylation loci that are marginally different between the case and control groups, which plays the same role as the sure independent screening method for high dimensional feature selection (Fan and Lv, 2008). However, this marginal approach may neglect the methylation loci which are jointly associated with cancer development but are marginally uncorrelated. To address this problem in the framework of linear models and generalized linear models, Fan and Lv (2008) and Fan et al. (2009) proposed an iterative sure independent screening method, which iteratively adds a new feature into the current variables and then perform the sure screening step. Different from these existing approaches, our case group may contain misclassification, such that the cancer status cannot be directly modeled by a logistic regression. It is of interest to develop an iterative screening method to handle our methylation data.

For analysis of DNA methylation data in cancer research, age has been considered as a strong demographic risk factor (Christensen et al., 2009). Chen et al. (2013) proposed an age-adjusted nonparametric method to detect differentially methylated loci. Specifically, the rank-based Kruskal-Wallis test was conducted separately in different age groups, then a combined p-value was reported. This method focuses on the differences in medians, and is expected to be underpowered when there are differences in variances. Alternatively, Huang et al. (2013) proposed an age-adjusted nonparametric method to capture the differences in both means and variances, where the Neuhäuser's one-sided test (Neuhäuser, 2003) was conducted within each age group and a combined p-value was reported. Permutation was used to obtain the age-specific p-values. However, both methods are not easy to extend to more than one confounders and do not account for the heterogeneity in DNA methylation among cases. The proposed PLEMT test can be extended to regression models where multivariate covariates are simultaneously adjusted and heterogeneity in cases can be accounted for. Such an extension is currently under investigation and will be reported in the future.

Acknowledgement

Yong Chen was supported in part by National Institutes of Health grants R01-LM009012 and R01-AI116794. Hao Wu was partially supported by National Institute of Health R01GM122083. We appreciate the constructive comments from the editor and the anonymous reviewers that have substantially improved the quality of this paper.

References

- Ahn, S. and Wang, T. (2013). A powerful statistical method for identifying differentially methylated markers in complex diseases. In *Pacific Symposium on Biocomputing*. <http://www.ncbi.nlm.nih.gov/pubmed/23424113>. World Scientific.
- Bandeem-Roche, K., Miglioretti, D. L., Zeger, S., and Rathouz, P. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92:1375–1386.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- Brocks, D., Assenov, Y., Minner, S., Bogatyrova, O., Simon, R., Koop, C., Oakes, C., Zucknick, M., Lipka, D. B., Weischenfeldt, J., et al. (2014). Intratumor dna methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell reports*, 8(3):798–806.
- Carter, H., Samayoa, J., Hruban, R. H., and Karchin, R. (2010). Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (chasm). *Cancer biology & therapy*, 10(6):582–587.
- Chen, H. and Chen, J. (2001). The likelihood ratio test for homogeneity in finite mixture models. *Canadian Journal of Statistics*, 29:201–215.
- Chen, Y. and Liang, K.-Y. (2010). On the asymptotic behavior of the pseudolikelihood ratio test statistic with boundary problems. *Biometrika*, 97:603–620.
- Chen, Z., Huang, H., Liu, J., Ng, H. K. T. and Nadarajah, S., Huang, X., and Deng, Y. (2013). Detecting differentially methylated loci for illumina array methylation data based on human ovarian cancer data. *BMC medical genomics*, 6:S9.
- Choi, Y. and Briollais, L. (2011). An EM composite likelihood approach for multistage sampling of family data. *Statistica Sinica*, 21:231–253.

- Christensen, B. C., Houseman, E. A. and Marsit, C. J. Z. S. W. M. R. W. J. L., et al. (2009). Aging and environmental exposures alter tissue-specific dna methylation dependent upon cpg island context. *PLoS genetics*, 5:e1000602.
- Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64:247–254.
- Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74:33–43.
- Di, C.-Z. and Liang, K.-Y. (2011). Likelihood ratio testing for admixture models with application to genetic linkage analysis. *Biometrics*, 67(4):1249–1259.
- Easwaran, H., Tsai, H.-C., and Baylin, S. B. (2014). Cancer epigenetics: tumor heterogeneity, plasticity of stem-like states, and drug resistance. *Molecular cell*, 54(5):716–727.
- Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.
- Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, 10:2013–2038.
- Fu, Y., Chen, J., and Kalbfleisch, J. D. (2006a). Testing for homogeneity in genetic linkage analysis. *Statistica Sinica*, 16:805–823.
- Fu, Y., Chen, J., and Kalbfleisch, J. D. (2006b). Testing for homogeneity in genetic linkage analysis. *Statistica Sinica*, pages 805–823.
- Gao, X. and Song, P. X.-K. (2011). Composite likelihood EM algorithm with applications to multivariate hidden markov model. *Statistica Sinica*, 21:165–185.
- Gervin, K., Hammerø, M., Akselsen, H. E., Moe, R., Nygård, H., Brandt, I., Gjessing, H. K., Harris, J., Undlien, D. E., and Lyle, R. (2011). Extensive variation and low heritability of dna methylation identified in a twin study. *Genome research*, 21:1813–1821.

- Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O., Wen, B., Wu, H., Liu, Y., Diep, D., et al. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43:768–775.
- Houseman, E. A., C., C. B., Yeh, R. F., Marsit, C. J., Karagas, M. R., Wrensch, M., Nelson, H. H., and et al. (2008). Model-based clustering of dna methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics*, 9:365.
- Huang, H., Chen, Z., and Huang, X. (2013). Age-adjusted nonparametric detection of differential dna methylation with case–control designs. *BMC Bioinformatics*, 14:86.
- Issa, J.-P. (2011). Epigenetic variation and cellular darwinism. *Nature genetics*, 43(8):724–726.
- Kalari, S., . P. G. P. (2010). Identification of driver and passenger dna methylation in cancer by epigenomic analysis. *Advances in Genetics*, 70:277.
- Kalbfleisch, J. D. (1978). Non-parametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society. Series B*, pages 214–221.
- Kim, W. Y. and Kaelin, W. G. (2004). Role of vhl gene mutation in human cancer. *Journal of Clinical Oncology*, 22(24):4991–5004.
- Lehmann, E. L. and D’Abrera, H. J. M. (1975). Nonparametrics: Statistical methods based on ranks, holden-day inc. *San Francisco*, pages 300–315.
- Li, P., Chen, J., and Marriott, P. (2009). Non-finite fisher information and homogeneity: an EM approach. *Biometrika*, 96:411–426.
- Liang, K.-Y. and Qin, J. (2000). Regression analysis under non-standard situations: a pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society: Series B*, 62:773–786.
- Liu, Y., Li, P., and Fu, Y. (2012). Testing homogeneity in a semiparametric two-sample problem. *Journal of Probability and Statistics*, 2012.

- Liu, Z.-Y., Yi, J., and Liu, F.-E. (2015). The molecular mechanism of breast cancer cell apoptosis induction by absent in melanoma (aim2). *International journal of clinical and experimental medicine*, 8(9):14750.
- Mikeska, T., Candiloro, I. L., and Dobrovic, A. (2010). The implications of heterogeneous dna methylation for the accurate quantification of methylation. *Epigenomics*, 2(4):561–573.
- Neuhäuser, M. (2003). Exact tests for the analysis of case-control studies of genetic markers. *Human Heredity*, 54:151–156.
- Ning, B.-F., Ding, J., Yin, C., Zhong, W., Wu, K., Zeng, X., Yang, W., Chen, Y.-X., Zhang, J.-P., Zhang, X., et al. (2010). Hepatocyte nuclear factor 4 α suppresses the development of hepatocellular carcinoma. *Cancer research*, 70(19):7640–7651.
- Oakes, C. C., Claus, R., Gu, L., Assenov, Y., Hüllein, J., Zucknick, M., Bieg, M., Brocks, D., Bogatyrova, O., Schmidt, C. R., et al. (2014). Evolution of dna methylation is linked to genetic aberrations in chronic lymphocytic leukemia. *Cancer discovery*, 4(3):348–361.
- Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249.
- Phipson, B. and Oshlack, A. (2014). Diffvar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biology*, 15(9):465.
- Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *The Annals of Statistics*, 27:1368–1384.
- Qin, J. and Liang, K.-Y. (2011). Hypothesis testing in a mixture case-control model. *Biometrics*, 67:182–193.
- Self, S. G. and Liang, K.-Y. (1987). Large sample properties of the maximum likelihood estimator and the likelihood ratio test on the boundary of the parameter space. *Journal of the American Statistical Association*, 82:605–611.

- Shaikh, R. S., Reuter, P., Sisk, R. A., Kausar, T., Shahzad, M., Maqsood, M. I., Yousif, A., Ali, M., Riazuddin, S., Wissinger, B., et al. (2015). Homozygous missense variant in the human *cnga3* channel causes cone-rod dystrophy. *European Journal of Human Genetics*, 23(4):473–480.
- Siegmund, K. D., Laird, P. W., and Laird-Offringa, I. A. (2004). A comparison of cluster analysis methods using dna methylation data. *Bioinformatics*, 20:1896–1904.
- Stavnes, H. T., Holth, A., Don, T., Kærn, J., Vaksman, O., Reich, R., Trope, C. G., and Davidson, B. (2013). *Hoxb8* expression in ovarian serous carcinoma effusions is associated with shorter survival. *Gynecologic oncology*, 129(2):358–363.
- Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.
- Tan, Z. (2009). A note on profile likelihood for exponential tilt mixture models. *Biometrika*, 96:229–236.
- Teschendorff, A. E., Jones, A., Fiegl, H., Sargent, A., Zhuang, J. J., Kitchener, H. C., and Widschwendter, M. (2012). Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Medicine*, 4(3):24.
- Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., and et al. (2010). Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research*, 20:440–446.
- Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.
- Varin, C., Høst, G., and Skare, Ø. (2005). Pairwise likelihood inference in spatial generalized linear mixed models. *Computational statistics & data analysis*, 49(4):1173–1191.
- Wang, S. (2011). Method to detect differentially methylated loci with case-control designs using illumina arrays. *Genetic epidemiology*, 35:686–694.
- Winslow, S., Leandersson, K., and Larsson, C. (2013). Regulation of *pmp22* mrna by *g3bp1* affects cell proliferation in breast cancer cells. *Molecular cancer*, 12(1):1.

- Wong, M., Hyodo, T., Asano, E., Funasaka, K., Miyahara, R., Hirooka, Y., Goto, H., Hamaguchi, M., and Senga, T. (2014). Silencing of *strn4* suppresses the malignant characteristics of cancer cells. *Cancer science*, 105(12):1526–1532.
- Xu, Z., Bolick, S. C., DeRoo, L. A., Weinberg, C. R., Sandler, D. P., and Taylor, J. A. (2013). Epigenome-wide association study of breast cancer using prospectively collected sister study samples. *Journal of the National Cancer Institute*, 105(10):694–700.
- Yang, J., Staples, O., Thomas, L. W., Briston, T., Robson, M., Poon, E., Simões, M. L., El-Emir, E., Buffa, F. M., Ahmed, A., et al. (2012). Human *chchd4* mitochondrial proteins regulate cellular oxygen consumption rate and metabolism and provide a critical role in hypoxia signaling and tumor progression. *The Journal of clinical investigation*, 122(2):600–611.
- Zhu, H. and Zhang, H. (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B*, 66:3–16.
- Zou, F., Fine, J. P., and Yandell, B. S. (2002). On empirical likelihood for a semiparametric mixture model. *Biometrika*, 89:61–75.

Table 1: Type I error (%) comparisons of the PLEMT test, the score test based on empirical likelihood (EST), the modified empirical likelihood ratio test (MELRT), the t -test, the Wilcoxon test, the Logistic regression test, the F test of equality of variances and the Kolmogorov-Smirnov (KS) test, at 0.05 and 0.1 significant levels for Normal, Beta, Gamma, Negative binomial and t models (Models A-E).

Model	level (%)	PLEMT	EST	MELRT	t-test	Wilcoxon	Logistic	F	KS
Non-misspecified models									
A: Normal	5.0	5.2	5.7	5.8	4.9	4.8	4.8	4.8	3.3
	10.0	10.1	10.9	11.3	10.2	9.9	10.0	9.6	7.7
B: Beta	5.0	5.0	5.1	5.1	4.7	4.7	4.3	5.3	3.7
	10.0	10.3	9.6	9.5	9.1	8.9	8.9	10.3	7.5
C: Gamma	5.0	5.3	6.2	6.3	5.5	5.4	5.2	10.7	3.8
	10.0	10.5	11.5	11.7	10.6	10.1	10.2	17.5	8.0
Misspecified models									
D: Negative binomial	5.0	5.5	5.7	5.7	5.0	4.8	4.8	7.9	1.5
	10.0	11.8	11.1	11.3	10.4	10.6	10.0	14.6	3.3
E: t	5.0	5.6	6.5	6.3	5.4	4.8	4.8	27.9	3.8
	10.0	11.0	11.2	12.1	9.8	10.1	9.3	36.0	7.9

Table 2: 2×2 tables for the number of sites identified by the PLEMT test vs. t -test, and the PLEMT test vs. the MELRT test.

		PLEMT				PLEMT	
		q -value < 0.05	≥ 0.05			q -value < 0.05	≥ 0.05
t -test	q -value < 0.05	2418	281	MELRT	< 0.05	2543	338
	q -value ≥ 0.05	694	19558		≥ 0.05	569	19501

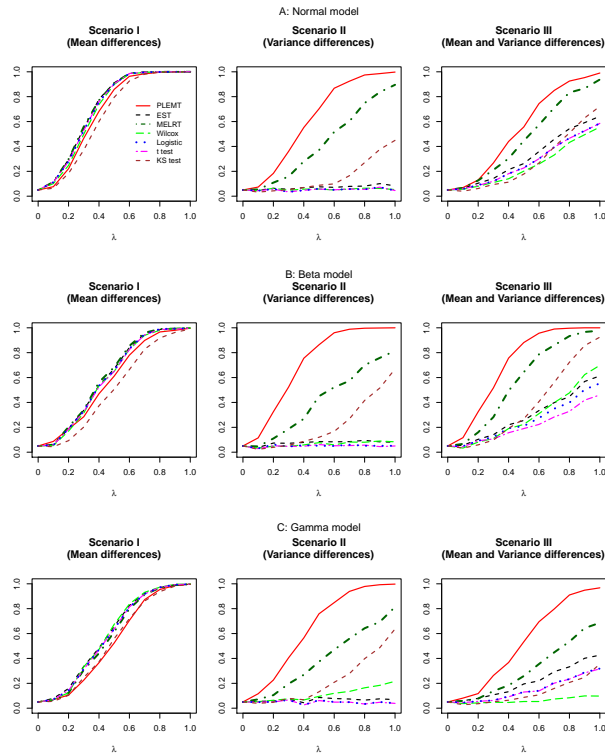


Figure 1: Power of the PLEMT test, the score test based on empirical likelihood (EST), the modified empirical likelihood ratio test (MELRT), the t -test, the Wilcoxon test, the Logistic regression test, and the Kolmogorov-Smirnov (KS) test for Normal, Beta, and Gamma models as a function of mixture proportion λ when the numbers of observations in two groups are $n_0 = n_1 = 100$.

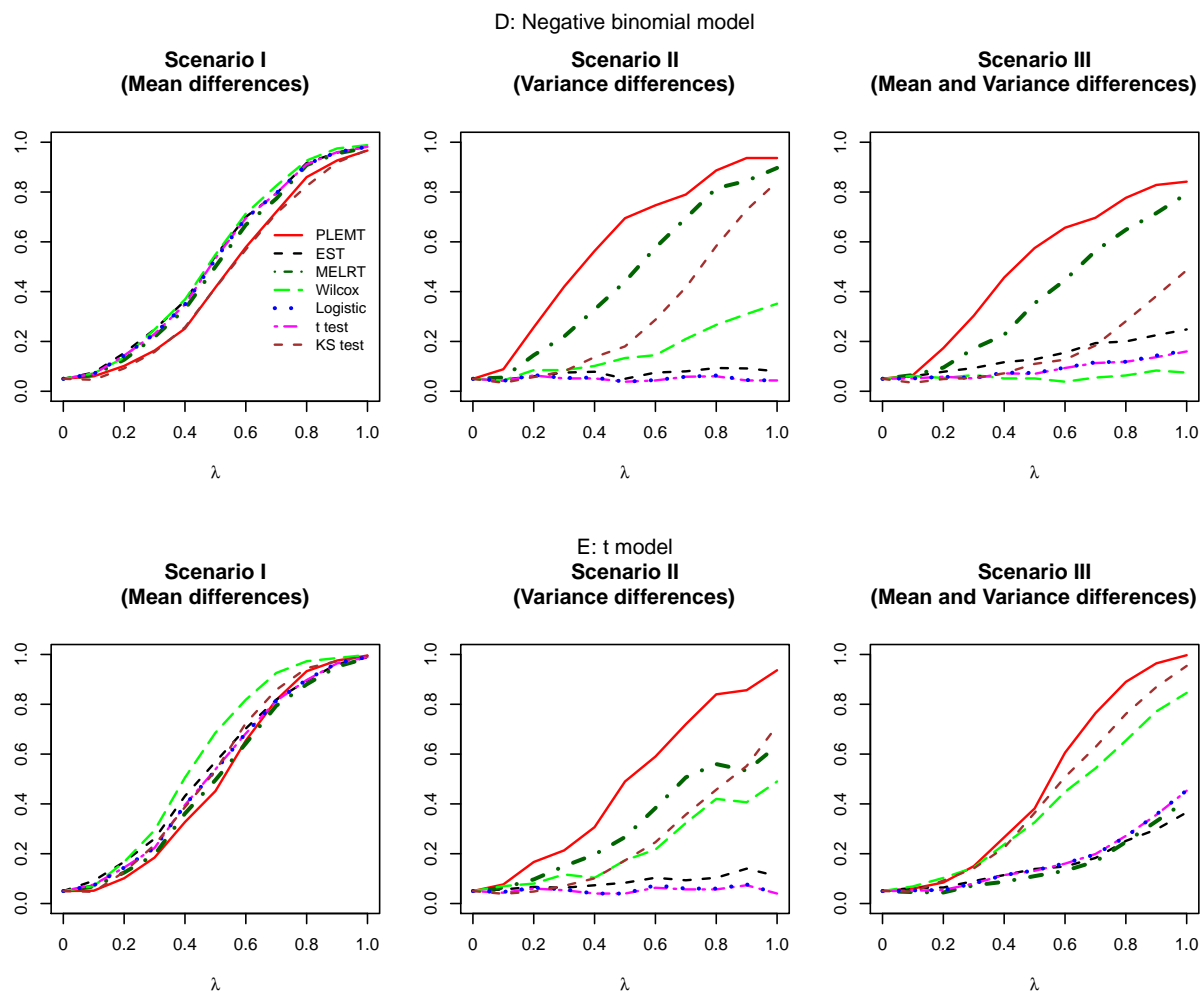


Figure 2: Power of the PLEMT test, the score test based on empirical likelihood (EST), the modified empirical likelihood ratio test (MELRT), the t -test, the Wilcoxon test, the Logistic regression test, and the Kolmogorov-Smirnov (KS) test for Negative binomial and T models as a function of mixture proportion λ when the numbers of observations in two groups are $n_0 = n_1 = 100$.

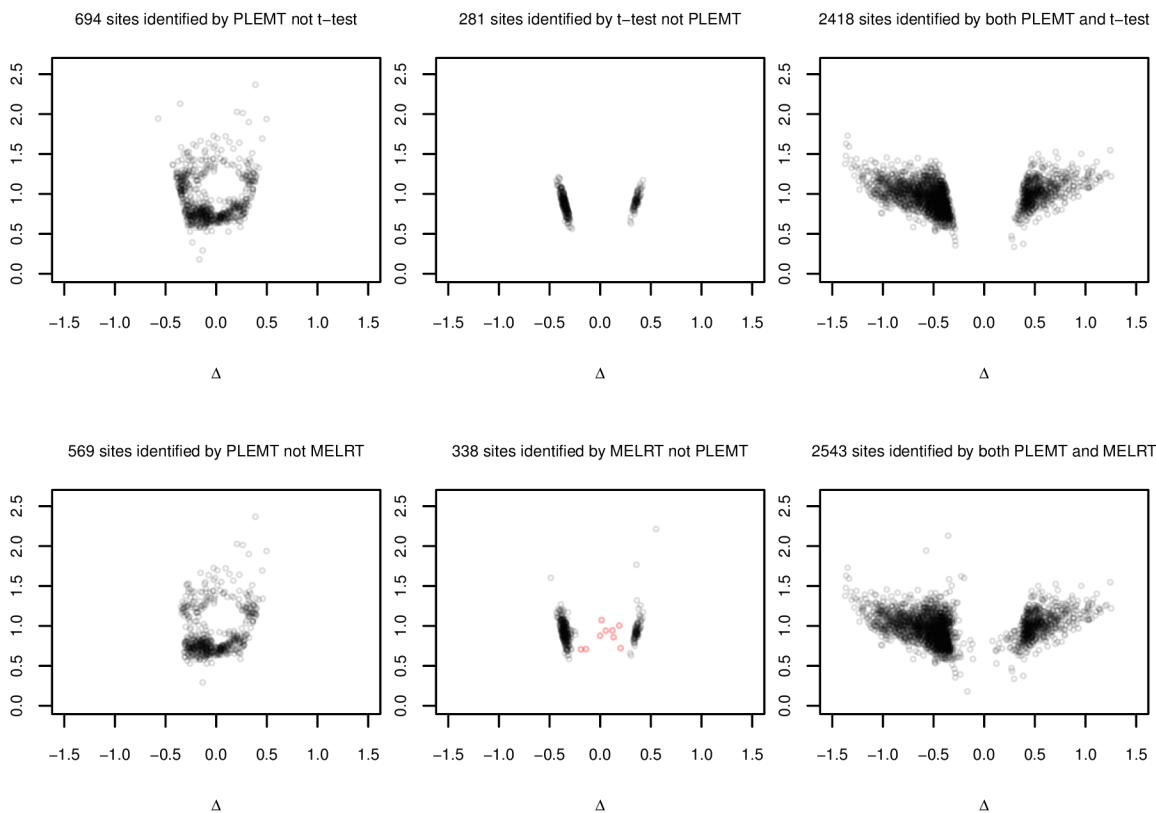


Figure 3: Upper panels: distributions of the standardized mean difference Δ and the ratio of standard deviations r_{21} for the 694, 281, 2418 sites that are identified by the PLEMT test but not the t -test, by the t -test but not the PLEMT test, by both the PLEMT test and the t -test, respectively; Lower panels: distributions of Δ and r_{21} for the 569, 338, 2543 sites that are identified by the PLEMT test but not the MELRT test, by the MELRT test but not the PLEMT test, by both the PLEMT test and the MELRT test. The 9 CpG sites between the clouds in the lower middle panel are highlighted in red.

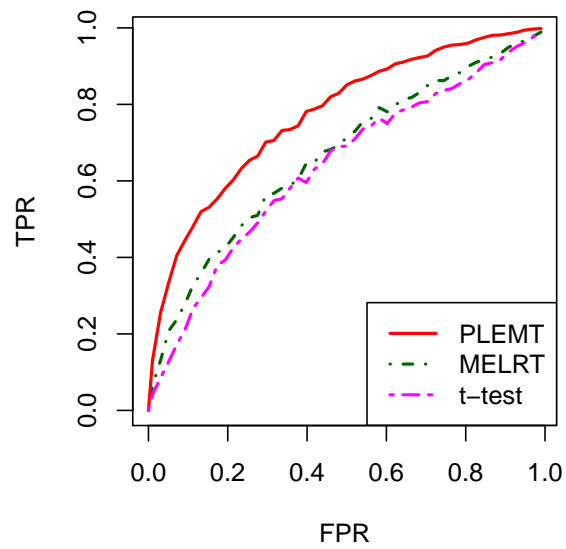


Figure 4: Receiver operating characteristics (ROC) curves for the PLEMT test, the MELRT test and t -test using Random Forest as the classification method.