Gene expression

Accurate and efficient estimation of small *P*-values with the cross-entropy method: applications in genomic data analysis

Yang Shi^{1,2,3,4,*}, Mengqiao Wang², Weiping Shi⁵, Ji-Hyun Lee⁶, Huining Kang^{3,*} and Hui Jiang^{4,7,8,*}

¹Division of Biostatistics and Data Science, Department of Population Health Sciences, Medical College of Georgia, Augusta University, Augusta, Georgia 30912, USA, ²Department of Epidemiology and Biostatistics, West China School of Public Health, Sichuan University, Chengdu, Sichuan 610041, China, ³Biostatistics Shared Resource, University of New Mexico Comprehensive Cancer Center and Department of Internal Medicine, University of New Mexico, Albuquerque, New Mexico 87131, USA, ⁴Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA, ⁵College of Mathematics, Jilin University, Changchun, Jilin 130012, China, ⁶Division of Quantitative Sciences, University of Florida Health Cancer Center and Department of Biostatistics, University of Florida, Gainesville, Florida 32610, USA, ⁷Center for Computational Medicine and Bioinformatics and ⁸University of Michigan Rogel Cancer Center, University of Michigan, Ann Arbor, Michigan 48109, USA

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on May 15, 2018; revised on November 6, 2018; editorial decision on December 3, 2018; accepted on December 5, 2018

Abstract

Motivation: Small *P*-values are often required to be accurately estimated in large-scale genomic studies for the adjustment of multiple hypothesis tests and the ranking of genomic features based on their statistical significance. For those complicated test statistics whose cumulative distribution functions are analytically intractable, existing methods usually do not work well with small *P*-values due to lack of accuracy or computational restrictions. We propose a general approach for accurately and efficiently estimating small *P*-values for a broad range of complicated test statistics based on the principle of the cross-entropy method and Markov chain Monte Carlo sampling techniques.

Results: We evaluate the performance of the proposed algorithm through simulations and demonstrate its application to three real-world examples in genomic studies. The results show that our approach can accurately evaluate small to extremely small *P*-values (e.g. 10^{-6} to 10^{-100}). The proposed algorithm is helpful for the improvement of some existing test procedures and the development of new test procedures in genomic studies.

Availability and implementation: R programs for implementing the algorithm and reproducing the results are available at: <https://github.com/shilab2017/MCMC-CE-codes>.

Contact: yshi@augusta.edu or hukang@salud.unm.edu or jianghui@umich.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

P-value is the most widely used metric to assess the statistical significance of genomic features in large-scale genomic studies such as genome-wide association studies (GWAS) and high-throughput

differential gene expression analysis. In those studies, very small *P*-values are often required to be accurately calculated, because: (i) A large number of tests are often performed in those studies and most of the methods used for multiple comparison adjustment in genomic

studies, such as the Bonferroni correction for family-wise error rate and the Benjamini-Hochberg procedure for controlling false discovery rate (Benjamini and Hochberg, 1995), work directly on the P -values associated with the genomic features. Therefore, it is essential to accurately evaluate P -values at very small scales so that those procedures are reliable. (ii) In practice, it is desirable to rank the significant genomic features by their P -values (often together with their effect sizes) so that the researchers can prioritize and follow up with those significant genomic features for further biological studies, which also requires that the small P -values associated with those features to be accurately estimated. In the literature, it is not uncommon to see that very small P -values associated with the most significant genomic features to the order of less than 10^{-100} are reported [e.g. (Burton et al., 2007)]. More examples can be found in (Bangalore et al., 2009)].

1.1 Problem formulation

The problem addressed in this work is how to estimate small P -values for a group of complicated test statistics whose cumulative distribution functions (CDF) are analytically intractable. Specifically, the question can be formulated as follows: the goal is to estimate the P -value defined as

$$P = \Pr[T(\mathbf{Y}) \geq q | H_0], \quad (1)$$

where $\mathbf{Y}$ is the data or transformed data that follow some probability distribution (e.g. multivariate normal distribution), $T(\mathbf{Y})$ is the test statistic which is a function of $\mathbf{Y}$, q is the test statistic calculated based on the observed data that can be either a scalar or a vector, and H_0 means that the probability is obtained under the null hypothesis, which will be dropped for simplicity hereafter. In most commonly used test procedures (e.g. two-sample t -test), the P -values are obtained by deriving the exact or asymptotic distributions of $T(\mathbf{Y})$ under H_0 . However, the problem we often encounter is that $T(\mathbf{Y})$ is complicated whose CDF under H_0 cannot be derived analytically, and existing approaches do not work for estimating very small P -values either due to lack of accuracy or unaffordable computational burden.

We illustrate this problem with the following three real-world examples in genomic studies.

Example 1: Gene set/pathway enrichment analysis. Here the goal is to test the significance of the association between some clinical outcomes of interest and the global expression pattern of a gene set or pathway (for brevity, gene set will be used hereafter), where gene sets are pre-specified groups of genes according to the biological functions or genomic locations of the genes. For a study with n independent subjects and a gene set with k genes, Goeman et al. proposed to fit the following model,

$$E(\mathbf{y}) = g^{-1}(\mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta}), \quad (2)$$

where $\mathbf{y}$ is the $n \times 1$ outcome vector, $\mathbf{Z}$ is an $n \times k$ expression matrix of the k genes in n subjects, $\mathbf{X}$ is an $n \times m$ matrix for the m covariates that needs to be adjusted, g is the canonical link function for the distribution of $\mathbf{y}$ (e.g. g is the identity function for normally distributed data or the logit function for binomially distributed data), and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the corresponding vectors of coefficients. The association between $\mathbf{Z}$ and outcome $\mathbf{y}$ can be assessed by testing the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ using the following test statistic

$$Q_1 = n^{-1}(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{Z} \mathbf{Z}^T (\mathbf{y} - \boldsymbol{\mu}), \quad (3)$$

where $\boldsymbol{\mu}$ is the expectation of $\mathbf{y}$ under H_0 (Goeman et al., 2004, 2011). A similar approach is also proposed in (Liu et al., 2007), where the matrix $\mathbf{Z} \mathbf{Z}^T$ in Eq. (3) is replaced by a kernel function to account for the interaction of genes in the same gene set.

Example 2: GWAS – joint testing a group of genetic markers in a genomic region. To increase the power of GWAS, approaches for joint testing a group of genetic markers (SNPs) in a genomic region instead of testing individual genetic markers are developed. Wu et al. proposed an approach under similar framework as in Example 1, and the following model is fit

$$E(\mathbf{y}) = g^{-1}(\mathbf{X}\boldsymbol{\alpha} + \mathbf{G}\boldsymbol{\beta}), \quad (4)$$

where $\mathbf{y}$ is the $n \times 1$ phenotype vector, $\mathbf{G}$ is an $n \times s$ genotype matrix for the s SNPs in the genomic region that need to be tested, $\mathbf{X}$ is an $n \times m$ matrix for the m covariates that needs to be adjusted, g is the canonical link function for the distribution of $\mathbf{y}$, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the vectors of coefficients. The association between $\mathbf{G}$ and phenotype $\mathbf{y}$ can be assessed by testing the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ using the following test statistic

$$Q_2 = (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{G} \mathbf{W} \mathbf{G}^T (\mathbf{y} - \boldsymbol{\mu}), \quad (5)$$

where $\boldsymbol{\mu}$ is the expectation of $\mathbf{y}$ under H_0 and $\mathbf{W}$ is a diagonal matrix containing the weights of the s SNPs (Wu et al., 2011). If the weight of each SNP is 1, then Q_2 is the same as Q_1 in Example 1 up to a constant. Similarly, the matrix $\mathbf{G} \mathbf{W} \mathbf{G}^T$ can be replaced by a kernel matrix to account for the epistatic effects of the SNPs (Wu et al., 2011).

Example 3: Ratio statistics in differential gene expression analysis. Consider the differential expression analysis comparing two groups of gene expression data. For a gene g to be tested, let $\mathbf{x}_{g1}$ and $\mathbf{x}_{g2}$ be the vectors of positive normalized gene expression values respectively for the two groups with sample sizes n_1 and n_2 . The following ratio statistic (also known as fold change or proportion statistic) has been proposed to test the differential expression of g between the two groups (Segal et al., 2017),

$$L = \max(y_1/y_2, y_2/y_1), \quad (6)$$

where y_1 and y_2 are the respective sample means of the two groups. Note that the P -value computed based on (6) is the two-sided P -value based on the test statistic, y_1/y_2 (Segal et al., 2017). Without loss of generality and for the ease of derivation, we assume $y_1 \geq y_2$ and use the test statistic,

$$L = y_1/y_2, \quad (7)$$

in the following discussions. Other approaches for testing differential gene expression based on the ratio statistics are also proposed in (Bergemann and Wilson, 2011; Chen et al., 2002; Newton et al., 2001).

The exact or asymptotic P -values with the test statistics in the above three examples can be expressed as the general form (1). To see this, for Examples 1 and 2, it can be shown that the test statistics Q_1 and Q_2 can be written in the following quadratic form

$$Q = T(\mathbf{Y}) = \mathbf{Y}^T \mathbf{D} \mathbf{Y}, \quad (8)$$

where $\mathbf{Y}$ follows a multivariate normal (MVN) distribution either exactly normal if the outcome or phenotype data $\mathbf{y}$ is assumed to follow normal distributions or asymptotically normal if $\mathbf{y}$ is assumed to follow binomial distributions, and $\mathbf{D}$ is a diagonal matrix containing the positive eigenvalues of $\mathbf{Z} \mathbf{Z}^T$ in Q_1 or $\mathbf{G} \mathbf{W} \mathbf{G}^T$ in Q_2 . See (Duchesne and De Micheaux, 2010) for matrix calculations about how Q_1 and Q_2 can be expressed as Q . Therefore, the P -values with the test statistics Q_1 and Q_2 can be expressed in the form $\Pr[T(\mathbf{Y}) \geq q_1]$, where q_1 is the test statistic calculated based on the observed data. For Example 3, let q_2 be the ratio statistic calculated based on the observed data, then the P -value with the ratio statistic L can be further expressed as

$$P = \Pr(y_1/y_2 \geq q_2 | y_1, y_2 > 0) = \Pr(\mathbf{C} \mathbf{Y} > 0), \quad (9)$$

where $\mathbf{C} = \begin{pmatrix} 0 & 1 \\ 1 & -q_2 \end{pmatrix}$ and $\mathbf{Y} = (y_1, y_2)^T$. Based on the central limit theorem, y_1 and y_2 respectively follow $N(\mu, \sigma^2/n_1)$ and $N(\mu, \sigma^2/n_2)$

either exactly if the expression data are assumed to be normal or asymptotically if the expression data are not normal, where μ and σ^2 are the population mean and variance, respectively, under the null hypothesis that there is no differential expression between the two groups.

1.2 Related works

In the literature, the quadratic form statistics (8), also known as a linear combination or weighted sum of chi-squared random variables (Bausch, 2013), are used for testing the associations between genomic features and the outcomes or phenotypes under several settings in genomic studies. A few methods are proposed specifically to calculate the tail probabilities for this form of statistic (Davies, 1980; Farebrother, 1984; Imhof, 1961; Liu *et al.*, 2009). See Duchesne and De Micheaux (2010) for comparisons of them and Bausch (2013) for a review. As commented in (Bausch, 2013), most of those existing methods do not work well when P is very small due to lack of accuracy or computational restrictions. In addition, those approaches are aimed specifically for the quadratic form Q and cannot be generalized to other types of test statistics, which have a relative narrow scope of applications.

Alternative to the above methods, a more general type of approaches is the utilization of Monte Carlo (MC) sampling methods, where a large number of MC random samples can be generated under H_0 either via simulations from the distribution of $\mathbf{Y}$ under H_0 (Lin, 2005) or through permutation or bootstrap of the observed data (note that permutation and bootstrap are special cases of MC methods, where the former samples the observed data without replacement and the latter samples with replacement), and then repeatedly calculate the test statistics using those MC samples and estimate P as the proportion of the test statistics based on the MC samples that are greater than or equal to the one based on the observed data. Yet if we want to accurately estimate very small P -values, this type of brute-force MC methods will need enormous computational effort.

In this paper, we propose a general approach for accurately and efficiently estimating small to extremely small P -values for any test statistic that can be expressed in the form (1). The basis of our approach contains two components. The first component is the cross-entropy (CE) method, which is originated from the concept of CE in information theory and has been widely used for rare event simulation in the operations research field (Rubinstein and Kroese, 2004). The second component is Markov chain Monte Carlo (MCMC) sampling techniques. Therefore, we refer our approach as MCMC-CE algorithm hereafter. The rest of this paper is organized as follows. In the next section, we give a general introduction of the CE method and MCMC techniques used in our approach, and then present our algorithm for estimating small P -values. In Section 3, we evaluate the performance of the proposed algorithm by comparing it with several existing approaches through simulations and demonstrate its applications with three real genomic datasets. Discussions are given in Section 4.

2 Materials and methods

2.1 The CE method

Our goal is to calculate the P -value as expressed in Eq. (1), which can be further written as

$$P = \Pr[T(\mathbf{Y}) \geq q] = E_{\theta_0}[I\{T(\mathbf{Y}) \geq q\}], \quad (10)$$

where the subscript θ_0 denotes the parameter vector of the probability distribution that $\mathbf{Y}$ follows under H_0 [e.g. it is an MVN distribution in the above three examples. We use $f(\cdot; \theta_0)$ to denote this distribution hereafter], and the expectation is taken with respect to $f(\cdot; \theta_0)$ with $I(\cdot)$ as the indicator function.

As discussed above, when P is small, using the brute-force MC method is computationally inefficient. The CE method is a general approach for the efficient estimation of small probabilities in MC simulations, which we briefly introduce below following the monograph on CE method (Rubinstein and Kroese, 2004). The technique used in the CE method is importance sampling (IS). Let $g(\cdot)$ be the proposal density function used in IS, then the expectation in Eq. (10) can be re-expressed as

$$p = \int I\{T(\mathbf{Y}) \geq q\} \frac{f(\mathbf{Y}; \theta_0)}{g(\mathbf{Y})} g(\mathbf{Y}) d\mathbf{Y} = E_g \left[I\{T(\mathbf{Y}) \geq q\} \frac{f(\mathbf{Y}; \theta_0)}{g(\mathbf{Y})} \right], \quad (11)$$

where the subscript g denotes that the expectation is taken with respect to $g(\cdot)$ now. Then P can be estimated by the MC counterpart (also known as stochastic counterpart) of (11),

$$\hat{p} = \frac{1}{N} \sum_{i=1}^N I\{T(\mathbf{y}_i) \geq q\} \frac{f(\mathbf{y}_i; \theta_0)}{g(\mathbf{y}_i)}, \quad (12)$$

where $\mathbf{y}_i$'s, $i = 1, \dots, N$ are random samples drawn from $g(\cdot)$ now. There is an optimal proposal density under which the IS estimator (12) has zero MC sampling variance (Rubinstein and Kroese, 2004), which is given by

$$g^*(\mathbf{Y}) = \frac{I\{T(\mathbf{Y}) \geq q\} f(\mathbf{Y}; \theta_0)}{P} \quad (13)$$

However, g^* cannot be directly used as the proposal density for estimating P in IS, since it contains the unknown probability P that is the quantity we want to calculate. The CE method provides a general solution to finding a proposal density $f(\cdot; \theta)$ which is close to the optimal proposal density g^* within the same distribution family as $f(\cdot; \theta_0)$ in the sense that the Kullback–Leibler divergence [also known as the Kullback–Leibler cross-entropy or cross-entropy (Rubinstein and Kroese, 2004)] between $f(\cdot; \theta)$ and g^* is minimized:

$$D(g^*(\cdot), f(\cdot; \theta)) := \int g^*(\mathbf{Y}) \ln \frac{g^*(\mathbf{Y})}{f(\mathbf{Y}; \theta)} d\mathbf{Y} = \int g^*(\mathbf{Y}) \ln g^*(\mathbf{Y}) d\mathbf{Y} - \int g^*(\mathbf{Y}) \ln f(\mathbf{Y}; \theta) d\mathbf{Y}, \quad (14)$$

where $f(\cdot; \theta)$ is another distribution within the same family as $f(\cdot; \theta_0)$, and we will call $f(\cdot; \theta)$ the CE-optimal proposal density below. Since the first term on the right-hand side of the second equality in Eq. (14) does not contain θ , the parameter θ that minimizes $D(g^*(\cdot), f(\cdot; \theta))$ should maximize the second term. Hence, the problem of finding the CE-optimal proposal density $f(\cdot; \theta)$ turns into an optimization problem of finding θ that maximizes the second term, $\int g^*(\mathbf{Y}) \ln f(\mathbf{Y}; \theta) d\mathbf{Y}$. Originally, Rubinstein *et al.* developed an adaptive algorithm to solve this optimization problem, which is referred as ‘multi-level CE method’ in the literature (Rubinstein and Kroese, 2004). We review that algorithm and discuss its limitations in detail in Supplementary Text Section S1.

One of the major limitations of the multi-level CE algorithm is that it is unreliable in high-dimensional settings (i.e. when the dimension of θ is large). With recent progress in MCMC sampling techniques, here we apply and implement an improved version of CE method based on the theoretical work in (Chan and Kroese, 2012) that combines the CE criterion (14) and MCMC techniques. Observe that the second term on the right-hand side of the second equality in (14) can be written as

$$\int g^*(\mathbf{Y}) \ln f(\mathbf{Y}; \theta) d\mathbf{Y} = E_{g^*}[\ln f(\mathbf{Y}; \theta)], \quad (15)$$

where the subscript g^* means that the expectation is taken with respect to the optimal proposal density g^* . Therefore, if we can draw

random samples from g^* (in the next section, we will discuss the techniques to that end), the parameter θ that minimizes $D(g^*(\cdot), f(\cdot; \theta))$ can be found by the maximization of the expectation in (15). By replacing the expectation in (15) with its MC counterpart, θ can be found by solving

$$\operatorname{argmax}_{\theta} \frac{1}{N} \sum_{l=1}^N [\ln f(y_l; \theta)], \quad (16)$$

where y_l 's, $l = 1, \dots, N$ are random samples drawn from g^* . Problem (16) is a regular maximum likelihood estimation problem, which can be solved either analytically or numerically by widely used approaches such as Newton-Raphson or the Expectation-Maximization algorithms.

2.2 Sampling from the optimal proposal density

Here we discuss the algorithms for sampling from g^* . From (13), observe that g^* is a truncated distribution with $f(\cdot; \theta_0)$ restricted by the constraint $T(\mathbf{Y}) \geq q$. With the recent progress in MCMC techniques, several algorithms are developed for sampling from a truncated distribution in the form of g^* and Table 1 summarizes four of them. The Gibbs sampler is a classical MCMC method for sampling from truncated distributions that consecutively draws samples from a sequence of conditional distributions (Geweke, 1991; Kotecha and Djuric, 1999). The hit-and-run sampler belongs to the class of line samplers, which reduces the problem of sampling from a multivariate constrained distribution to that of sampling from a univariate truncated distribution (Chen and Schmeiser, 1993; Kroese et al., 2011). The Hamiltonian and Lagrangian Monte Carlo samplers are two more recently developed powerful tools for sampling from many complicated distributions, which respectively use the principles of the Hamiltonian and Lagrangian dynamics in physics (Lan et al., 2015; Pakman and Paninski, 2014). In our empirical comparisons, we find that the Hamiltonian and Lagrangian Monte Carlo samplers are more efficient for sampling from a truncated distribution with quadratic constraints such as Examples 1 and 2, while the Gibbs sampler is more efficient for sampling from a truncated distribution with linear constraints such as Example 3 (not shown here). More comparisons of those algorithms can be found in (Kroese et al., 2011; Lan et al., 2015; Pakman and Paninski, 2014).

2.3 The MCMC-CE algorithm for calculating small P -values

Combining the above discussions, our algorithm for estimating the small P -value $p = \Pr[T(\mathbf{Y}) \geq q]$ contains two steps: in the first step we draw random samples from g^* and solve the maximization problem (16) to obtain the CE-optimal proposal density $f(\cdot; \theta)$, and in the second step we estimate P using standard IS with $f(\cdot; \theta)$ as the proposal density. The algorithm is summarized as follows:

Table 1. Algorithms for sampling from the optimal proposal distribution

Algorithm	Reference
Gibbs sampler	Geweke (1991) and Kotecha and Djuric (1999)
Hit-and-run sampler	Chen and Schmeiser (1993) and Kroese et al. (2011)
Hamiltonian Monte Carlo sampler	Brubaker et al. (2012) and Pakman and Paninski, (2014)
Lagrangian Monte Carlo sampler	Lan et al. (2015)

Main Algorithm (MCMC-CE method for estimating small P -values)

A. Parameter updating step:

1. Draw N random samples $y_1, \dots, y_N$ from g^* using an efficient MCMC sampling algorithm (as shown in Table 1).
2. Solve the maximization problem (16) and obtain the CE-optimal proposal density $f(\cdot; \theta)$.

B. Estimating step:

Draw M random samples $y_1, \dots, y_M$ from $f(\cdot; \theta)$. Estimate P as

$$\hat{p} = \frac{1}{M} \sum_{l=1}^M I\{T(y_l) \geq q\} \frac{f(y_l; \theta_0)}{f(y_l; \theta)}.$$

3 Results

3.1 Simulation studies

We perform simulations to evaluate the performance of MCMC-CE. In the first part of the simulations, we investigate the estimation accuracy and computational efficiency of MCMC-CE via simulations from random variables whose tail probabilities can be calculated analytically. Our simulations focus on calculating $\Pr[T(\mathbf{Y}) \geq q]$, where $\mathbf{Y}$ is an MVN random variable. Specifically, we use the following two types of random variables whose true tail probabilities are available in most statistical packages, so that we can evaluate the errors and variations of MCMC-CE: (1) chi-squared random variables, which can be expressed as a special case of quadratic functions of MVN random variables, and (2) standard Cauchy random variable, which is the ratio of two independent standard normal random variables and can be expressed as a special case of linear functions of MVN random variables. The details are given in Supplementary Text Section S2.1 and the results are shown in Supplementary Tables S1–S5, which are briefly summarized below.

In the first experiment, we use the following four χ_m^2 's (denotes a chi-squared random variable with m degrees of freedom): χ_3^2 , χ_{20}^2 , χ_{50}^2 and χ_{100}^2 , where the dimension increases from 5 to 100. With each of them, we generate a sequence of true P -values on the order from 10^{-6} to 10^{-100} and compare the results of MCMC-CE and several other methods specific for calculating the tail probabilities of the quadratic form statistic (8), including Davies' method (Davies, 1980), Farebrother's method (Farebrother, 1984) and Imhof's method (Imhof, 1961). The details of this experiment are given in Supplementary Text Section S2.1 and the results are shown in Supplementary Tables S1–S4. This experiment shows that MCMC-CE can accurately calculate P to the order of 10^{-100} with less than 5% relative errors in all the four settings, while none of other methods in the comparisons works well when P is smaller than 10^{-16} . Figure 1 shows a graphical demonstration of the concordance between the P -values estimated using MCMC-CE and the true P -values for chi-squared random variables with different degrees of freedom.

In the second experiment, we use MCMC-CE to calculate the small tail probabilities of the standard Cauchy random variable and compare the results with the true values. The details of this experiment are given in Supplementary Text Section S2.1 and the results are shown in Supplementary Table S5. The experiment shows that MCMC-CE can accurately calculate P -values to the order of 10^{-100} with less than 3% relative errors.

In the second part of the simulations, we perform sensitivity analysis to examine the estimation accuracy and computational

efficiency of MCMC-CE with different numbers of random samples in the parameter updating and estimating steps (i.e. N and M in the Main Algorithm), where we use MCMC-CE to estimate the P -values on the orders from 10^{-6} to 10^{-100} based on four chi-squared random variables, χ^2_5 , χ^2_{20} , χ^2_{50} and χ^2_{100} with combinations of different N s and M s. The details are given in [Supplementary Text Section S2.2](#) and the results are presented in [Supplementary Table S6](#). The results show that the variations of MCMC-CE (assessed using the metric standardized root-mean-square error, SRMSE) are reduced by increasing N and M . When N grows large enough with a fixed M , the variations no longer decrease, which indicates that the parameter updating step is stabilized. Once the parameter updating step is stabilized with a large enough N , the SRMSE decreases roughly proportionally with the square root of M . The variations can be well-controlled (less than 5%) with large enough N and M even when the dimensionality of the parameter space is very large and the P -value is extremely small, and the computation time is affordable on a typical desktop computer ([Supplementary Table S6](#)).

3.2 Application to genomic data analysis

We apply MCMC-CE to the estimation of small P -values in three real-world examples from genomic studies.

Example 1: Gene set/pathway enrichment analysis. We apply MCMC-CE to a dataset containing gene expression measurements and clinical variables of melanoma patients, which is part of The Cancer Genome Atlas (TCGA) project and publicly available from TCGA data portal: <https://portal.gdc.cancer.gov/>. Particularly, the dataset contains the expression levels of 20 531 genes from 355 melanoma patients measured by RNA-Seq, and we are interested in testing which gene sets are associated with the clinical variable of interest, Breslow thickness. The gene set annotations are extracted from the Gene Ontology Consortium ([Ashburner et al., 2000](#); [Gene Ontology Consortium, 2017](#)), where 22 211 gene sets are curated. For each gene set, model (2) is fitted with log-transformed Breslow thickness as the outcome variable and gender and

age as the adjusted covariates. Since the computational time will be overwhelming if all the 22 211 gene sets are to be tested using MCMC-CE, we use the following screening test to filter out those less-significant gene sets: we calculate the approximated P -values for all the gene sets using the method implementing in the *globaltest* package ([Goeman et al., 2004](#)), where the distribution of the test statistic Q_1 as defined in Eq. (3) is approximated by a scaled chi-square distribution. Based on this screening test, there are 35 gene sets with the approximated P -values less than 10^{-8} , and we use MCMC-CE to accurately calculate the P -values associated with those 35 gene sets. For each gene set, the algorithm is run 100 times to obtain the variations of the estimated P -value, and $N=10\,000$ random samples in the parameter updating step and $M=10\,000$ random samples in the estimating step are used.

[Supplementary Table S7](#) presents the results for all the 35 gene sets and [Table 2](#) presents the top 10 most significant ones. The results show that MCMC-CE can efficiently calculate extremely small P -values to the order of 10^{-54} ([Table 2](#)). For comparison, two other approaches for calculating the P -values for the quadratic form statistic (8), the Davies' and Farebrother's methods, are also applied in this example. Similar to the observations in the simulations, neither of those two methods works when the P -values are smaller than 10^{-16} , and they agree with the results from MCMC-CE in the cases where the P -values are not too small ([Supplementary Table S7](#)).

Example 2: GWAS – joint testing a group of genetic markers in a genomic region. We demonstrate the application of MCMC-CE in testing groups of SNPs in GWAS. The dataset used is collected in a GWAS performed in the population of about 2000 heterogeneous stock mice phenotyped for over 100 traits ([Valdar et al., 2006](#)), which is public available at: <https://wp.cs.ucl.ac.uk/outbredmice/heterogeneous-stock-mice/>. We are interested in testing which regions of SNPs are associated with the serum concentration of high-density lipoprotein (HDL). After pre-processing, the dataset contains 1640 subjects with complete HDL data and 10 990 SNPs with complete genotype data. For simplicity, we group each adjacent 20 SNPs based on their genomic locations as one region, which results in 549 groups of SNPs. For each group of SNPs, model (4) is fitted with HDL as the phenotype variable and gender and weight of the mice as the adjusted covariates, and MCMC-CE is used to calculate the P -value with the test statistic (5). Similar to Example 1, the algorithm is run 100 times to obtain the variations of the estimated P -value, and $N=10\,000$ and $M=10\,000$ random samples are used. [Supplementary Table S8](#) presents the 70 groups of SNPs that are significantly associated HDL with P -values less than 10^{-8} and [Table 3](#) shows the top 10 most

Fig. 1. Concordance between the true tail probabilities of the chi-squared random variables and the results from MCMC-CE. In each figure panel, the solid line represents the true tail probabilities and the dots represent the ones estimated from MCMC-CE. The detailed results are presented in [Supplementary Tables S1–S4](#)

Table 2. Top ten gene sets significantly associated with Breslow thickness ranked by their P -values.

GO term	P -value	S.D.	Time(Average)	Time(SD)
GO:0048880	2.29×10^{-54}	4.73×10^{-56}	2.45×10^{-1}	2.28×10^{-2}
GO:1900019	7.17×10^{-53}	1.10×10^{-54}	2.58×10^{-1}	2.36×10^{-2}
GO:1900020	7.20×10^{-53}	1.67×10^{-54}	2.63×10^{-1}	2.12×10^{-2}
GO:0045499	1.25×10^{-48}	2.12×10^{-50}	2.30×10^{-1}	1.99×10^{-2}
GO:0004415	1.50×10^{-45}	3.09×10^{-47}	2.54×10^{-1}	2.30×10^{-2}
GO:0016941	5.05×10^{-45}	7.49×10^{-47}	2.37×10^{-1}	2.41×10^{-2}
GO:0007168	5.19×10^{-45}	9.91×10^{-47}	2.43×10^{-1}	2.13×10^{-2}
GO:2000020	6.46×10^{-35}	2.08×10^{-36}	1.97×10^{-1}	1.71×10^{-2}
GO:2000018	7.21×10^{-35}	2.14×10^{-36}	1.87×10^{-1}	1.75×10^{-2}
GO:0045163	1.85×10^{-34}	3.61×10^{-36}	2.26×10^{-1}	1.98×10^{-2}

Note: P -value, the P -value estimated from MCMC-CE, which is the average of the 100 runs of the algorithm; S.D., standard deviation of the results of 100 runs. Time (Average) and Time (SD), the average and standard deviation of the real elapsed time in seconds for a single run of MCMC-CE on a single core of Intel Xeon X5550 2.67GHz CPU calculated based on 100 runs of the algorithm. See [Supplementary Table S7](#) for the detailed results.

Table 3. Top ten groups of SNPs significantly associated with HDL ranked by their P -values

SNP group index	P -value	S.D.	Time (Average)	Time (SD)
Group 42	5.93×10^{-38}	1.11×10^{-39}	2.00×10^{-1}	1.78×10^{-2}
Group 44	5.65×10^{-36}	3.80×10^{-37}	2.28×10^{-1}	1.91×10^{-2}
Group 200	1.51×10^{-19}	6.27×10^{-21}	1.76×10^{-1}	1.59×10^{-2}
Group 214	1.88×10^{-19}	1.78×10^{-20}	2.43×10^{-1}	1.81×10^{-2}
Group 43	6.06×10^{-19}	1.98×10^{-20}	1.66×10^{-1}	1.93×10^{-2}
Group 314	1.03×10^{-17}	7.38×10^{-18}	1.89×10^{-1}	1.77×10^{-2}
Group 528	7.61×10^{-17}	8.03×10^{-18}	2.06×10^{-1}	1.66×10^{-2}
Group 213	4.64×10^{-16}	9.71×10^{-18}	1.84×10^{-1}	1.65×10^{-2}
Group 276	6.09×10^{-16}	9.56×10^{-17}	2.12×10^{-1}	1.74×10^{-2}
Group 273	7.71×10^{-16}	1.13×10^{-17}	1.83×10^{-1}	1.87×10^{-2}

Note: The meanings of P -value, S.D., Time (Average) and Time (SD) are the same as Table 2. See Supplementary Table S8 for the detailed results.

significant ones. Those results demonstrate that MCMC-CE can efficiently calculate extremely small P -values to the order of 10^{-38} . Our analyses based on SNP groups are also in agreement with the tests for individual SNP reported in the original study (Valdar et al., 2006). We also applied the Davies' and Farebrother's methods in this example, and same as before, neither of them works when the P -values are smaller than 10^{-16} (Supplementary Table S8).

Example 3: Ratio statistic in differential gene expression analysis. We demonstrate the application of MCMC-CE in estimating small P -values based on the ratio statistic (6) in differential gene expression analysis and show how MCMC-CE can be used to assess the genome-wide significance after the adjustment of multiple comparisons. The dataset used is from a study on patients who were diagnosed with salivary adenoid cystic carcinoma and received radiation therapy, and the expression levels of 22 243 genes in the salivary gland tissues of those patients were measured by RNA-Seq. The details of the study can be found in Brayer et al. (2016) and the sequencing read data are available from the NCBI Sequence Read Archive using accession number SRP059557. The dataset used consists of 14 patients, where 8 were free of cancer and 6 were not at the end of the study, and here we are interested in testing genes differentially expressed between those two groups of patients. After filtering out lowly expressed genes, 11 390 genes are left and the gene count data are normalized using the trimmed mean of M -values method implemented in R package *edgeR* (Robinson et al., 2010) and log-transformed counts per million (CPM) values are used for our analysis.

The following methods for estimating the P -values based on the ratio statistic are included in our comparisons:

1. A brute-force MC method: for each gene, the two-sided P -value is computed as $\hat{p} = 2 \times \min[\frac{1}{M_{BF}} \sum_{l=1}^M I(\frac{y_{1l}}{y_{2l}} \geq q), \frac{1}{M_{BF}} \sum_{l=1}^M I(\frac{y_{1l}}{y_{2l}} \leq q)]$, where y_{1l} 's and y_{2l} 's, $l = 1, \dots, M_{BF}$ are MC samples drawn from the normal distributions described in Section 1.1, q is the statistic computed based on the observed data, and M_{BF} , the number of MC samples, is set as 10^5 .
2. MCMC-CE: based on the results of the above brute-force MC method, there are 150 genes with P -values less than 10^{-4} . We use MCMC-CE to accurately compute the P -values for those 150 genes, then we combine the results with those genes with P -values greater than 10^{-4} as the final results. Here $N = 10\,000$ random samples in the parameter updating step and $M = 10\,000$ random samples in the estimating step are used, and the algorithm is run 100 times for each P -value to assess the variations.
3. A permutation method: for each gene, the P -value is computed as $\hat{p} = \frac{1}{N_{perm}} \sum_{i=1}^N I(L_i \geq q)$, where L_i 's, $i = 1, \dots, N_{perm}$ are the

Table 4. Number of significantly differentially expressed genes identified by each method with different FDR thresholds

	FDR threshold					
	0.001	0.005	0.01	0.05	0.1	0.15
Brute-force MC	52	52	52	105	195	296
MCMC-CE	17	30	40	82	190	296
Perm0	31	31	31	31	31	57
Perm1	0	0	0	0	0	0
samr	7	7	7	22	56	226

test statistic as defined in (6) computed based on the permutations of the observed gene expression data and q is the same statistic computed based on the observed data. Here, the set of all permutations of the observed data can be enumerated (i.e. $N = 3003$). We also use an alternative formula to compute the P -value as $\hat{p} = \frac{1}{N_{perm}+1} [\sum_{i=1}^N I(L_i \geq q) + 1]$. The two formulas are referred as 'Perm0' and 'Perm1' below.

For all the above methods, we use the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control the FDR given the P -values computed by each method. As a comparison, we also run the differential expression analysis using *samr* package with its default settings (Tusher et al., 2001). Table 4 presents the numbers of significantly differentially expressed genes identified by each method with different FDR threshold values, and Supplementary Table S8 shows the detailed results. We can see that the brute-force MC method suffers the issue that the results with more stringent FDR thresholds are not reliable (Table 4, FDR threshold = 0.01, 0.005 and 0.001), as the small P -values with those most significant genes cannot be accurately estimated due to the limited number of MC samples. Those permutation-based approaches (Perm0, Perm1 and *samr*) suffer the same issue, though it should be noted that the null hypotheses tested by them are different from the brute-force MC and MCMC-CE. This example shows that accurate estimation of small P -values is useful for correctly evaluating the genome-wide significance. Of note, the smallest P -value estimated by MCMC-CE in this example is 1.11×10^{-15} (Supplementary Table S8). We also illustrate the application of MCMC-CE for differential expression analysis in a microarray dataset with larger sample size and more extreme P -values, which is given in Supplementary Text Section S3.

4 Discussion

In summary, we propose an algorithm, MCMC-CE, for accurate and efficient estimation of small P -values for test statistics that can be expressed in the form of Eq. (1) and demonstrate its applications in genomic data analyses. To apply MCMC-CE, the following requirements should be met: (i) The test statistic needs to be written as a function of $\mathbf{Y}$, and $\mathbf{Y}$ follows a certain distribution that belongs to a parametric family of distributions, as denoted by $f(\cdot; \cdot)$ in Section 2; (ii) It is feasible to generate random samples from $f(\cdot; \cdot)$; (iii) The density of $f(\cdot; \cdot)$ can be evaluated pointwisely. Although we demonstrate the application of MCMC-CE for the test statistics that are quadratic or linear functions of MVN random variables in the simulations and real-world examples, MCMC-CE can also work for more complicated test statistics that are not limited to the functions of MVN random variables (Chan and Kroese, 2012; Pakman and Paninski, 2013). Therefore, it can help researchers develop new test procedures in genomic studies and estimate small P -values for a broad range of test statistics.

Since both the parameter updating and estimating steps of MCMC-CE involve drawing random samples and the simulations also show that the variations of the algorithm tend to increase when the dimensionality of the parameter space grows and the P -values become smaller (Supplementary Text Section S2.1), here we discuss the sources of variations and provide some strategy for the choice of the numbers of random samples in the parameter updating step (i.e. N) and in the estimating step (i.e. M) for the practical usage of the algorithm based on the sensitivity analysis (Supplementary Text Section S2.2 and Supplementary Table S6). Basically, the variations are from the following sources:

1. The variations of random sampling in the parameter updating step, which can affect the estimation of the CE-optimal proposal density in Eq. (16). This portion of variations can be reduced by increasing N . When N is large enough with a fixed M , the overall variations no longer decrease, which indicates that the parameter updating step is stabilized (Supplementary Text Section S2.2 and Supplementary Table S6).
2. The variations of random sampling from the CE-optimal proposal density in the estimating step. This portion of variations can be reduced by increasing M . As shown in the simulations, once the parameter updating step is stabilized with large enough N , the standardized root-mean-square error decreases roughly proportionally with the square root of M (Supplementary Table S6). This agrees with the general Monte Carlo method, where the Monte Carlo sampling error decreases proportionally with the square root of the number of random samples (Kroese et al., 2011).
3. Since MCMC-CE uses the CE-optimal proposal density $f(\cdot; \theta)$ to approximate the optimal proposal density g^* , $f(\cdot; \theta)$ still does not behave exactly like g^* , which leads to the intrinsic variations of this method. This portion of variations gains its weight with the ‘curse of dimensionality’ (i.e. when the dimensionality of the parameter space grows), which can be seen by comparing the variations in the simulation results from χ^2_5 to χ^2_{100} (Supplementary Table S1–S4).

In practice, sensitivity analysis like Supplementary Table S6 is recommended for the choice of optimal N and M when applying MCMC-CE. A general strategy is to first find the optimal N by examining the variations across different values of N with a fixed M . Once the optimal N is determined, we can then increase M to achieve a desired level of accuracy.

Acknowledgements

We thank Drs. Maureen Sartor and Xiaoquan Wen at University of Michigan for reading and helpful discussions on Section 2, which is part of YS’s doctoral dissertation (Shi, 2016). The computational resources used in this work were provided by the University of New Mexico Center for Advanced Research Computing and Augusta University Medical College of Georgia, Department of Population Health Sciences.

Funding

This work was partly supported by the following grants: one startup grant from Augusta University Medical College of Georgia (YS), two startup research grants from Sichuan University supported by the Fundamental Research Funds for the Central Universities of China (20822041B4009 of YS and 20822041A4202 of MW), the National Natural Science Foundation of China Grant J1310022 (WS), NIH grant 5P30CA118100 (YS, JL and HK) and NIH grant P30CA046592 (HJ).

Conflict of Interest: none declared.

References

- Ashburner, M. et al. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
- Bangalore, S.S. et al. (2009) How accurate are the extremely small P -values used in genomic research: an evaluation of numerical libraries. *Comput. Stat. Data Anal.*, **53**, 2446–2452.
- Bausch, J. (2013) On the efficient calculation of a linear combination of chi-square random variables with an application in counting string vacua. *J. Phys. A Math. Theor.*, **46**, 505202.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, **289**–300.
- Bergmann, T.L. and Wilson, J. (2011) Proportion statistics to detect differentially expressed genes: a comparison with log-ratio statistics. *BMC Bioinformatics*, **12**, 228.
- Brayer, K.J. et al. (2016) Recurrent fusions in MYB and MYBL1 define a common, transcription factor-driven oncogenic pathway in salivary gland adenoid cystic carcinoma. *Cancer Disc.*, **6**, 176–187.
- Brubaker, M.A. et al. (2012) A Family of MCMC Methods on Implicitly Defined Manifolds. In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. The Society for Artificial Intelligence and Statistics, La Palma, Canary Islands, Spain, PMLR, pp. 161–172.
- Burton, P.R. et al. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
- Chan, J.C. and Kroese, D.P. (2012) Improved cross-entropy method for estimation. *Stat. Comput.*, **22**, 1031–1040.
- Chen, M.-H. and Schmeiser, B. (1993) Performance of the Gibbs, hit-and-run, and Metropolis samplers. *J. Comput. Graph. Stat.*, **2**, 251–272.
- Chen, Y. et al. (2002) Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics*, **18**, 1207–1215.
- Davies, R.B. (1980) Algorithm AS 155: the distribution of a linear combination of χ^2 2 random variables. *J. R. Stat. Soc. Ser. C (Appl. Stat.)*, **29**, 323–333.
- Duchesne, P. and De Micheaux, P.L. (2010) Computing the distribution of quadratic forms: further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Comput. Stat. Data Anal.*, **54**, 858–862.
- Farebrother, R. (1984) Algorithm AS 204: the distribution of a positive linear combination of χ^2 2 random variables. *J. R. Stat. Soc. Ser. C (Appl. Stat.)*, **33**, 332–339.
- Gene Ontology Consortium (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331–D338.
- Geweke, J. (1991) Efficient simulation from the multivariate normal and student- t distributions subject to linear constraints and the evaluation of constraint probabilities. In: *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. CiteSeer, pp. 571–578.
- Goeman, J.J. et al. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.
- Goeman, J.J. et al. (2011) Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika*, **98**, 381–390.
- Imhof, J. (1961) Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 419–426.
- Kotecha, J.H. and Djuric, P.M. (1999) Gibbs sampling approach for generation of truncated multivariate gaussian random variables. In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999. IEEE, pp. 1757–1760.
- Kroese, D.P. et al. (2011) *Handbook of Monte Carlo Methods*. John Wiley and Sons, New York.
- Lan, S. et al. (2015) Markov Chain Monte Carlo from Lagrangian Dynamics. *J. Comput. Graph. Stat. Jt. Publ. Am. Stat. Assoc. Inst. Math. Stat. Interface Found. N. Am.*, **24**, 357–378.
- Lin, D.Y. (2005) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, **21**, 781–787.
- Liu, D. et al. (2007) Semiparametric Regression of multidimensional genetic pathway data: least – Squares Kernel Machines and Linear Mixed Models. *Biometrics*, **63**, 1079–1088.

- Liu, H. et al. (2009) A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comput. Stat. Data Anal.*, **53**, 853–856.
- Newton, M.A. et al. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.
- Pakman, A. and Paninski, L. (2013) Auxiliary-variable exact Hamiltonian Monte Carlo samplers for binary distributions. In: *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Neural Information Processing Systems Foundation, Inc., pp. 2490–2498. Curran Associates, Inc., Lake Tahoe, Nevada, USA.
- Pakman, A. and Paninski, L. (2014) Exact hamiltonian monte carlo for truncated multivariate gaussians. *J. Comput. Graph. Stat.*, **23**, 518–542.
- Robinson, M.D. et al. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
- Rubinstein, R.Y. and Kroese, D.P. (2004) *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer-Verlag, New York.
- Segal, B.D. et al. (2017) Fast approximation of small p-values in permutation tests by partitioning the permutations. *Biometrics*, **74**, 196–206.
- Shi, Y. (2016) Statistical and computational methods for differential expression analysis in high-throughput gene expression data, Ph.D. Dissertation, University of Michigan.
- Tusher, V.G. et al. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, **98**, 5116–5121.
- Valdar, W. et al. (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.*, **38**, 879–887.
- Wu, M.C. et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.