A Comparison of Parametric and Semi-Parametric Models for Microarray Data Analysis

Linda Chaba[a], John Odhiambo[a] and Bernard Omolo[b,*]

[a]*Strathmore Institute of Mathematical Sciences, Strathmore University, Ole Sangale Road, Nairobi, Kenya*

[b]*Division of Mathematics and Computer Science, University of South Carolina-Upstate 800 University Way, Spartanburg, South Carolina, USA*

Abstract: Microarray technology has revolutionized genomic studies by enabling the study of differential expression of thousands of genes simultaneously. Parametric, nonparametric and semi-parametric statistical methods have been proposed for gene selection within the last sixteen years. In an effort to find the "gold standard", the performance of some common parametric and nonparametric methods have been compared in terms of power to select differentially expressed genes and other desirable properties. However, no such comparisons have been conducted between parametric and semi-parametric models. In this study, we compared a semi-parametric model based on copulas with a parametric model (the quantitative trait analysis or QTA model) in terms of power and the ability to control the Type I error rate. In addition, we proposed a simple algorithm for choosing an optimal copula. The two approaches were applied to a publicly available melanoma cell lines dataset for validation. Both methods performed well in terms of power but the copula approach was notably the better. In terms of the Type I error rate control, the two methods were comparable. More methods for selecting an optimal copula for gene expression data need to be developed, as the proposed procedure is limited to copulas that permit both negative and positive dependence only.

Keywords: Copula, Goodness-of-fit, Melanoma, Microarray, Power, Type I error.

1. BACKGROUND

Microarray technology has revolutionized genomic studies by enabling the study of differential expression of thousands of genes simultaneously. The main objective in microarray experiments is to identify a panel of genes that are associated with a disease outcome or trait. A number of statistical methods have been proposed for gene selection within the last sixteen years. These include parametric [1-7], semi-parametric [8-11] and non-parametric [12-15] methods. Most of these methods concentrate on finding differentially expressed genes (DEGs) when the gene expression data is measured in two conditions.

When the relationship between two continuous variables is of interest, the dependence parameter in the predictive model becomes of interest as well. The most common dependence parameter used in such cases is the correlation coefficient. In differential gene expression analysis, the use of the correlation coefficient is implemented through a parametric method known as the quantitative trait analysis (QTA) model. The QTA model assumes that the variables are correlated and that the residuals are (approximately) normally distributed. An alternative method that uses the concept of the dependence parameter is the copula model. The copula model is a semi-parametric model, in the sense that no assumption is made on the distribution of the marginals but the dependence parameter is assumed to come from a parametric family.

In this paper, we compared a copula-based semi-parametric model with a parametric model (the QTA model), in terms of power and control of the Type I error rate. In addition, we proposed a simple procedure for choosing an optimal copula for gene expression data. The rest of the paper is organized as follows. In Section 2, we review the QTA method for gene selection. In Section 3, we discuss the copula model. Here, we also propose a simple approach for selecting an optimal copula for modeling gene expression data. The comparison of the two methods based on simulated datasets is presented in Section 4, and an application of the two methods to a real dataset is provided in Section 5. Section 6 provides a brief conclusion of the study.

2. MATERIALS AND METHODS

2.1. Quantitative Trait Analysis (QTA) Method

This approach finds genes that are significantly correlated with a quantitative outcome such as age. It uses the Pearson's correlation or the Spearman's (rank) correlation coefficient as a measure of dependence to compute p-values.

*Address correspondence to this author at the Division of Mathematics & Computer Science, University of South Carolina-Upstate, 800 University Way, Spartanburg, South Carolina, USA; Tel: +1 864-503-5362; Fax: +1 864-503-5930; E-mail: bomolo@uscupstate.edu

E-ISSN: 1929-6029/17

Let X_{ij} be the expression level for the i-th gene from the j-th sample and y_j be the covariate for the j-th sample. The (linear) model of analysis can be expressed as

$$X_{ij} = \beta_{i0} + \beta_{i1} y_j + \varepsilon_{ij},\ i = 1,2,...,G,\ j = 1,2,..,n. \quad (1)$$

Here, we assume that

$$\varepsilon_{ij} \sim N(0,\sigma_i^2),\ i = 1,...,G,\ j = 1,...,n \quad (2)$$

and that the y_j values are fixed (not random). β_{i0} and β_{i1} represent the regression coefficients specific to gene i. For testing the significance of correlation for the i-th gene ($H_{0i}: \beta_{i1} = 0$ vs $H_{1i}: \beta_{i1} \neq 0$), we use the statistic T_i, defined as

$$T_i = \frac{\hat{\beta}_{i1}}{SE(\hat{\beta}_{i1})}, \quad (3)$$

where $SE(\hat{\beta}_{i1})$ is the standard error of $\hat{\beta}_{i1}$. T_i has the t-distribution with $n-2$ degrees of freedom. We reject H_{0i} if $|t_i| < t_{\alpha/2}$, $0 < \alpha < 1$. Equivalently, one can test for ρ_i, the correlation coefficient. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in})'$ and $\mathbf{y} = (y_1, y_2, ...y_n)'$. Given the observations (x_{ij}, y_j), the Pearson's correlation coefficient for the i-th gene, r_i, is defined as

$$r_i = \frac{n\sum_{j=1}^{n} x_{ij} y_j - \sum_{j=1}^{n} x_{ij} \sum_{j=1}^{n} y_j}{\sqrt{n\sum_{j=1}^{n} x_{ij}^2 - (\sum_{j=1}^{n} x_{ij})^2}\sqrt{n\sum_{j=1}^{n} y_j^2 - (\sum_{j=1}^{n} y_j)^2}}. \quad (4)$$

For testing $H_{0i}: \rho_i = 0$ vs. $H_{1i}: \rho_i \neq 0$, we use the statistic T_i^*,

$$T_i^* = \frac{r_i\sqrt{(n-2)}}{\sqrt{1-r_i^2}}, \quad (5)$$

which also has a t-distribution with $n-2$ degrees of freedom. Here r_i is the estimator of ρ_i. H_{0i} is rejected if $|t_i^*| < t_{\alpha/2}$, $0 < \alpha < 1$. With a simple algebraic manipulation, it can be shown that (3) and (5) are equivalent and so the latter was employed in this study.

There are two ways of controlling the number of false positives in the QTA method. The first approach is based on the p-values computed from the parametric t-tests. Here, a stringent *p*-value threshold (say $p < 0.001$), is used in controlling the number of false positives. The second approach uses the multivariate permutation tests [16]. The multivariate permutation tests are based on permutations of the covariate. For each permutation, the parametric test statistics are re-computed to determine a p-value for each gene. The genes are ordered by their p-values computed for each permutation, with genes having the smallest p-values appearing at the top of the list. For a pre-selected p-value threshold, the distribution of the number of genes that would have p-values smaller than that threshold is computed. That is the distribution of the number of false discoveries, since genes that are significant for random permutations are false discoveries. The algorithm selects a threshold p-value so that the number of false discoveries is not greater than that specified by the user C percent ($C\%$) of the time, where C denotes the desired confidence level [17]. The QTA method is implemented by the BRB-ArrayTools software [17].

The QTA approach estimates the false discovery rate (FDR) using the Benjamini and Horchberg's approach [18]. For the i-th gene, the estimated FDR is given by

$$\widehat{FDR}_i = \frac{G \times p_i}{i}, \quad (6)$$

where p_i is the univariate *p*-value for the i-th most significant genes and G is the number of genes tested.

2.2. Copula Method

A copula is a bivariate distribution with uniform marginals. By Sklar's theorem [19], for any distribution function, F, with marginals F_1 and F_2, there exists a copula, C, such that

$$F(x_1, x_2) = C[F_1(x_1), F_2(x_2); \theta], \quad (7)$$

for $(x_1, x_2)'$ in the support of F, with dependence parameter θ. This result can be easily extended to multivariate distributions, to yield Sklar's theorem in *m*-dimensions, which we now state (without proof):

Let F be an *m*-dimensional distribution function with margins $F_1(x_1),...,F_m(x_m)$. Then there exists an *m*-copula, C, such that for all $\mathbf{x} = (x_1, x_2, ..., x_m)' \in \bar{\mathbb{R}}^m$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m)' \in \Theta \subset \mathbb{R}^m$,

$$F(x_1, ..., x_m) = C\left[F_1(x_1), ..., F_m(x_m); \boldsymbol{\theta}\right]. \quad (8)$$

If $F_1, F_2, ..., F_m$ are all continuous, then C is unique and can be expressed as

$$C(u_1,u_2,...,u_m;\boldsymbol{\theta}) = F(F_1^{-1}(u_1),F_2^{-1}(u_2),...,F_m^{-1}(u_m)), \quad (9)$$

for any $\mathbf{u} = (u_1,u_2,...,u_m)' \in [0,1]^m$.

Upon differentiation, (8) becomes

$$f(x_1,x_2,...,x_m) = \frac{\partial^m C(F_1(x_1),...,F_m(x_m);\boldsymbol{\theta})}{\partial F_1(x_1)...\partial F_m(x_m)} \prod_{i=1}^{m} \frac{\mathrm{d}F_i(x_i)}{\mathrm{d}x_i}$$

$$= c(F_1(x_1),...,F_m(x_m);\boldsymbol{\theta}) \prod_{i=1}^{m} f_i(x_i). \quad (10)$$

Here, f, c and f_i are the densities for F, C and F_i, respectively. Let $u_i = F_i(x_i)$. Then (10) becomes

$$f(x_1,x_2,...,x_m) = c(u_1,u_2,...,u_m;\boldsymbol{\theta}) \prod_{i=1}^{m} f_i(x_i). \quad (11)$$

Now, consider a random sample $\{(X_{1j},...,X_{mj}) : j = 1,2,...,n\}$ from the distribution $F(x_1,...,x_m)$. One can fit a copula model by estimating the dependence parameters using the maximum likelihood approach. In practice, it is more convenient to work with the logarithm of a likelihood function because it simplifies subsequent mathematical analysis. Since the logarithm is a monotonically increasing function, maximizing the log of a function is the same as maximizing the function itself. The log-likelihood is given as

$$\ell_n(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log c(F_1(x_{1j}),...,F_m(x_{mj});\boldsymbol{\theta}) + \sum_{j=1}^{n}\sum_{i=1}^{m} \log(f_i(x_{ij})). \quad (12)$$

Since the marginals are unknown, each $F_i(x_i)$ may be replaced with its marginal estimator $\hat{F}_i(x_i)$ to obtain $\hat{\theta}_i$. This approach is referred to as the canonical maximum likelihood estimation (CMLE) method [20]. Here, $\hat{F}_i(x_i)$ is given by

$$\hat{F}_i(x_i) = \frac{n}{n+1}\frac{1}{n}\sum_{j=1}^{n} I\left(X_{ij} \le x_i\right), \quad (13)$$

where I is the indicator function. Rescaling the empirical distribution by $\frac{n}{n+1}$ avoids the potential unboundedness of $\log(c(F_1(x_{1j}),...,F_m(x_{mj});\boldsymbol{\theta}))$, as some of the $F_i(x_{ij})$'s tend to be one [20]. The corresponding pseudo-loglikelihood is given as

$$\ell_n^*(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log c(\hat{F}_1(x_{1j}),...,\hat{F}_m(x_{mj});\boldsymbol{\theta}) + \sum_{j=1}^{n}\sum_{i=1}^{m} \log(f_i(x_{ij})), \quad (14)$$

and the estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} \approx \underset{\boldsymbol{\theta}\in\Theta}{argmax} \sum_{j=1}^{n} \log c(\hat{F}_1(x_{1j}),...,\hat{F}_m(x_{mj});\boldsymbol{\theta}), \quad (15)$$

since the last summand in (14) does not depend on $\boldsymbol{\theta}$. Under suitable regularity conditions, $\hat{\boldsymbol{\theta}}$ is consistent and is asymptotically normal [20]. In general, multivariate models do not have closed form estimators and so numerical methods are used in the estimation process [21].

2.3. Copula Model for Differential Gene Expression

We were interested in the pairwise correlation between each gene's expression profile and a quantitative outcome. Therefore, the copula of interest was the bivariate copula ($m = 2$). Suppose a microarray experiment consists of n samples and G genes. Let $\mathbf{x}_i = (x_{1i},...,x_{ni})'$ be the gene expression profile for gene i and $\mathbf{y} = (y_1,...,y_n)'$ be a vector of the covariate of interest (quantitative trait). We wanted to find K genes that are correlated with $\mathbf{y}$, $0 < K < G$. That is, we were interested in determining whether, for each gene i, X_i and Y are independent or not. The test for independence, thus, becomes testing for the null hypothesis

$$H_{0i} : Y \perp X_i \quad (X_i \text{ and } Y \text{ are independent}), \quad (16)$$

against the alternative hypothesis

$$H_{1i} : Y \not\perp X_i \quad (X_i \text{ and } Y \text{ are independent}). \quad (17)$$

For multiple genes, (16) is tested simultaneously and so the hypothesis of interest becomes

$$H_0 : Y \perp X_i \text{ for all i} = \bigcap_{i=1}^{G} H_{0i} \quad (18)$$

vs.

$$H_1 : Y \not\perp X_i \text{ for some i} = \bigcup_{i=1}^{G} H_{1i}. \quad (19)$$

In terms of copulas, assume that for each gene i, the joint distribution of Y and X_i is generated by a parametric copula $C(u_1,u_2;\theta_i)$ such that

$$H_i(y,x_i) = C[F(y),F_i(x_i);\theta_i], \quad (20)$$

where $H_i(y,x_i)$, $F(y)$ and $F_i(x_i)$ are the CDFs of (Y,X_i), Y and X_i respectively. Here $u_1 = F(y)$, $u_2 = F_i(x_i)$ and θ_i is the dependence parameter. Equation (18) and (19) now become

$$H_0 : \bigcap_{i=1}^{G} \left[C(u_1, u_2; \theta_i) = u_1 u_2 \text{ for all } (u_1, u_2)^T \in [0,1]^2 \right], \quad (21)$$

vs.

$$H_1 : \bigcup_{i=1}^{G} \left[C(u_1, u_2; \theta_i) \neq u_1 u_2 \text{ for some } (u_1, u_2)^T \in [0,1]^2 \right]. \quad (22)$$

A normal copula, for instance, attains independence when θ_i = 0. In this case, the global hypothesis to test for the dependence in terms of θ_i is expressed as

$$H_0 : \bigcap_{i=1}^{G} (\theta_i = 0) \text{ vs. } H_1 : \bigcup_{i=1}^{G} (\theta_i \neq 0). \quad (23)$$

2.4. Hypothesis Testing

To test (23), we needed the distribution of $\hat{\theta}_i$ under the null hypothesis. Rather than assume a parametric distribution for the null hypothesis, we used a permutation resampling-based approach [22]. For a given nominal level α, a gene is differentially expressed if its *p*-value is less than α. To adjust for multiple comparisons, the FDR approach [23] was used. The global null hypothesis (23) is rejected if at least one of its components (H_{0i}) is rejected, based on the estimated FDR values.

2.5. Copula Algorithm for Identifying DEGs

Our copula-based algorithm for finding DEGs can be summarised as follows:

1. Estimate θ_i using the CMLE method. In the CMLE approach, no assumption is made on the marginal distribution. The marginal distribution for each gene, $F_i(x_i)$, and a quantitative outcome, $F(y)$, are replaced with their estimators $\hat{F}_i(x_i)$ and $\hat{F}(y)$, respectively, to obtain $\hat{\theta}_i$.

$$\hat{\theta}_i \approx \underset{\theta_i \in \Theta}{argmax} \sum_{j=1}^{n} \log c(\hat{F}_i(x_{ij}), \hat{F}(y_j); \theta_i). \quad (24)$$

A detailed explanation of the CMLE method is provided in the Supplementary Materials (B).

2. Find gene-specific *p*-values (unadjusted *p*-values) using the permutation based resampling method. See Supplementary Materials (C) for details.

3. Apply the FDR approach to control for Type I error. See Supplementary Materials (D) for details.

4. A gene is differentially expressed if its estimated FDR (estimated *q*-value) is less than some specified value $\alpha \in$ [0,1].

An R code for implementing the algorithm is available from the authors upon request.

2.5.1. Which Copula to Use?

Having an appropriate copula in copula modelling is very crucial. To date, no study has been conducted on choosing the best copula model for gene expression data analysis. Whenever a copula model has been applied to gene expression data, the choice has been arbitrary. Some authors have chosen copulas based on how convenient they were for their analyses [11, 24]. Others have chosen copulas based on the magnitude of the likelihood of the copulas (e.g. [25]).

Several tests have been proposed for the copula specification. The most commonly used are the goodness-of-fit tests [26-30]. Goodness-of-fit tests are based on a direct comparison of the dependence implied by the copula with the dependence observed in the data.

In most empirical applications, the unique copula C is assumed to come from a parametric family $C_0 = \{C_\theta, \theta \in \Theta\}$ with $\Theta \subset R$. In goodness-of-fit testing for copula models, the hypothesis of interest is given by $H_0 : C \subset C_0$, i.e. that the copula C belongs to a pre-determined parametric family C_0. For testing H_0, the marginal distributions are treated as nuisance parameters and are replaced by their empirical distribution functions, $\hat{F}_i(x_i)$, as defined in (13) [31].

Copulas can also be selected according to their ranks based on some criteria. The most commonly used criteria are the Akaike Information Criteria (AIC) [32] and the Bayesian Information Criteria (BIC) [33]. These are defined as follows:

$$AIC = -2 \sum_{j=1}^{n} \ln[c(u_{1j}, u_{2j}); \theta] + 2K. \quad (25)$$

$$BIC = -2 \sum_{j=1}^{n} \ln[c(u_{1j}, u_{2j}); \theta] + K \ln(n). \quad (26)$$

Here, $u_{ij} = F_i(x_{ij}), i = 1,2$, and $K = 1$ for the one-parameter copulas. Similarly, $K = 2$ for the two-parameter copulas. The copula with the least AIC or least BIC is chosen to be the best. Kim *et al*. [34] used the AIC approach to assess the goodness-of-fit of their proposed copula-based method, the survival truncated Farlie-Gumbel-Morgenstern (FGM) type modification copulas.

With several copulas to choose from in empirical applications, one needs an appropriate one for the data at hand. The statistical features of the data should guide the selection of the copulas. For example, gene expression profiles can be positively or negatively associated with a quantitative outcome. Therefore, naturally, copulas that can capture both the negative and the positive dependence such as the Normal copula, the Student-t copula and the Frank copula should be superior to the Gumbel and Clayton copulas, which do not permit negative dependence. To this end, we recommend the following procedure:

1. Perform copula model selection from a list of candidate parametric copulas on all the pairs (a quantitative outcome and each gene expression profile). This helps in determining the closest parametric copula family from the list of candidate copulas.

2. Record the proportion of pairs that are fitted by the parametric copulas.

3. The copula that fits most of the pairs is assumed for the whole analysis.

We considered two copulas: the Normal copula and the Frank copula, since they permit both positive and negative dependence. We performed model selection based on the AIC and the BIC, using the melanoma cell lines dataset. The Student-t copula is close to Normal copula, hence was not considered. The copula that fitted the highest proportion of the pairs was adopted for the comparison of the two gene selection methods. The goodness-of-fit-test for the two copulas was also performed, using the Cramer-Von Mises (CVM) function [31].

2.6. Data

2.6.1. Simulated Gene Expression Data

Let n and G denote the number of samples and genes, respectively. Further, let D denote the number of genes assumed to be truly differentially expressed. Then $(G-D)$ genes are assumed to be non-differentially expressed. The gene expression data matrix, $\mathbf{X}$, is a $G \times n$ matrix of log2-ratios. We can write $\mathbf{X}$ as $\mathbf{X}=(\mathbf{X_1},\mathbf{X_2})$, where $\mathbf{X_1}$ and $\mathbf{X_2}$ are $D \times n$ and $(G-D)\times n$ matrices, respectively. We set $D \in (50,100,200,300,400)$, $n=35$ and G to be 1000. We generated the $(1000-D)$ genes from the standard normal distribution. To generate the D genes, we used the standard normal distribution in conjunction with the Cholesky decomposition [35] of their correlation matrix as follows:

1. Generate an unstructured correlation matrix $\mathbf{\Omega}$. $\mathbf{\Omega}$ is a $(D+1)\times(D+1)$ matrix that has $(i,j)^{th}$ element given by $\omega_{i,j}=\text{corr}(x_i,x_j)$

2. Find the Cholesky factor, $\mathbf{A}$, of $\mathbf{\Omega}$ such that $\mathbf{\Omega}=\mathbf{AA}'$.

3. Let $\mathbf{z}_i \sim N(\mathbf{0},\mathrm{I}_n), i=1,2,...,(D+1)$.

4. $\mathbf{Z}=(\mathbf{z}_1,\mathbf{z}_2,...,\mathbf{z}_{D+1})'$

5. $\mathbf{X}_{D+1}=\mathbf{AZ}$

$\mathbf{X}_{D+1}$ is the gene expression matrix for D genes assumed to be differentially expressed and a covariate $\mathbf{y}$. $\mathbf{y}$ can take any of the $D+1$ row vectors from the matrix $\mathbf{X}_{D+1}$. $\mathbf{X}_1$ is therefore a submatrix of $\mathbf{X}_{D+1}$ with dimensions $D \times n$.

2.6.2. Real Datasets

For validation, a publicly available melanoma cell lines dataset was used. This dataset contained gene expression data (raw intensities) on 54 cell-lines (35 melanoma cell lines and 19 normal human melanocytes (NHMs), each with 45,015 probes. Only the melanoma cell lines were analyzed. The raw data was median-normalized and log2-transformed. Multiple probes were reduced to one per gene by using the most variable probe(set)–measured by interquartile range (IQR)–across arrays. Filtration and normalization of the gene expression data was implemented using BRB Array Tools software [17]. A gene was filtered out if less than 20% of its expression data values had at least 1.5-fold change in either direction from the genes median value. Genes with more than 50% missing data across all its samples were also filtered out. There were 3,860 genes available for subsequent analysis.

For the continuous outcome, we used three uncorrelated quantitative traits studied by Kaufmann *et al.* [36] and Kaufmann *et al.* [37], to quantify the biological process in melanoma progression. These quantitative traits are the G_1 checkpoint function, the G_2 checkpoint function and the chromosomal instability (CIN) index. The G_1 checkpoint regulates entry into synthesis phase (S-phase) based on internal and external conditions. If the conditions are not conducive, the G_1 checkpoint will not allow cells to enter the S-phase. On the other hand, the G_2 checkpoint is a position of control in the cell cycle that delays or arrests mitosis when damaged DNA cells are detected, thereby providing the opportunity for repair and stopping the proliferation of damaged cells. Kaufmann

et al. [36] quantified the G_1 checkpoint function by treating cells with 1.5 Gy IR (or sham treatment for control) and then labeling cells with BrdU for 2 hours beginning 6 hours post-treatment. They then calculated the fraction of the BrdU-labeled nuclei within the first half of the S-phase in irradiated cells expressed as a percentage of the equivalent fraction in the sham-treated controls. The G_2 checkpoint function was scored as a ratio of mitotic cells in 1.5 Gy ironizing radiation (IR)-treated cultures in comparison to their sham-treated control (i.e. IR to sham ratio). Chromosomal instability (CIN) is a type of genomic instability in which chromosomes are unstable. We used the CIN data described in [37]. Kaufmann *et al*. [37] quantified the CIN by using array comparative genomic hybridization to identify somatic copy number alterations (deletions and duplications). The CIN index was determined by summing all segments with non-diploid DNA content. These outcome data were obtained from Kaufmann's lab (UNC - Pathology and Lab Medicine).

2.7. Analysis

The copula and the QTA methods were applied to the simulated datasets. They were compared in terms of the power to detect DEGs and the ability to control Type I error rate. The DEGs were identified at different nominal levels: 0.01, 0.05, 0.1 and 0.2. Power was calculated as the ratio of the number of correctly identified differentially expressed genes, true positives (TP), to the total number of actual DEGs, *D*. Thus,

$$\text{Power} = \frac{TP}{D}. \tag{27}$$

Type I error rate (Error) was calculated as the ratio of the number of genes that were falsely declared differentially expressed, false positives (FP), to the number of genes that were identified as differentially expressed. Thus

$$\text{Error} = \frac{FP}{FP + TP}. \tag{28}$$

For validation, the two methods were applied to the real melanoma cell lines dataset to identify DEGs using the G_1 checkpoint function, the G_2 checkpoint function and the CIN index separately as the quantitative traits.

3. RESULTS AND DISCUSSIONS

3.1. Copula Selection

Given the results in Table **1**, all the three methods applied for copula selection and goodness-of-fit testing suggest that the Normal copula fitted most of the pairs, and was therefore adopted for subsequent analysis.

Table 1: Copula Model Selection Based on the Three Methods

Model selection method	Normal copula	Frank copula
CVM	35.18%	23.18%
AIC	58.40%	41.60%
BIC	58.40%	41.60%

3.2. Simulation Results

Table **2** shows the number of genes declared to be differentially expressed for the copula and the QTA methods at different levels of FDR threshold. The results indicated that, in general, the copula method identified more DEGs than the QTA method. We note that the identified DEGs are likely to include both the truly DEGs and the false positives.

Power comparison results are shown in Table **3**. Both the copula and the QTA methods had sufficient power to detect DEGs with a power of 1 in most cases. A power of 1 means that the method is able to detect all the known DEGs. In cases where the power was different for the two methods, the copula method stood out as the better method. This is noted for $D =$ 50, 200 and 300.

Controlling Type I error here means having an empirical Type I error rate close to the nominal level of the test. The closer the empirical Type I error rate is to the set nominal level, the better the method in controlling it. Table **4** shows that both methods reasonably controlled Type I error at different nominal levels. There was evidence of both over and under estimation of the nominal levels by both methods, though the deviations were minimal. We note that the copula method consistently estimated the 0.01 nominal level for all the values of D except for $D = 50$, compared to the QTA method. We also see that the accuracy of controlling the Type I error rate for the copula method increases with the increase in the number of known DEGs. This means that, even with a large number of DEGs identified by the copula method, we could still trust the copula approach to properly control the Type I error rate. The closeness of the results for the two methods on the simulated datasets may be due to the fact the data was generated from a normal distribution. Note that a bivariate Gaussian copula with two normal marginals corresponds to a bivariate Gaussian distribution. As such, the copula

Table 2: Number of DEGs by the Copula and the QTA Methods at Different Estimated FDR Levels

D	Method	Estimated FDR threshold (α)			
		0.01	**0.05**	**0.1**	**0.2**
50	Copula	48	52	53	59
	QTA	50	53	57	63
100	Copula	100	106	110	126
	QTA	99	103	109	116
200	Copula	192	221	240	276
	QTA	146	204	221	241
300	Copula	302	314	326	378
	QTA	284	306	315	348
400	Copula	403	423	449	506
	QTA	405	415	429	466

Table 3: Power Comparison for the Copula Method and the QTA Method at Different Nominal Levels and Number of known DEGs

D	Method	Estimated FDR threshold (α)			
		0.01	**0.05**	**0.1**	**0.2**
50	Copula	0.96	1.00	1.00	1.00
	QTA	1.00	1.00	1.00	1.00
100	Copula	0.99	1.00	1.00	1.00
	QTA	0.99	1.00	1.00	1.00
200	Copula	0.95	1.00	1.00	1.00
	QTA	0.72	0.97	1.00	1.00
300	Copula	1.00	1.00	1.00	1.00
	QTA	0.94	0.99	1.00	1.00
400	Copula	1.00	1.00	1.00	1.00
	QTA	1.00	1.00	1.00	1.00

Table 4: Type I Error Rates of the Copula Method Compared to the QTA Method at Different Nominal Levels and Number of known DEGs

D	Method	Estimated FDR threshold (α)			
		0.01	**0.05**	**0.1**	**0.2**
50	Copula	0.00	0.04	0.06	0.15
	QTA	0.00	0.06	0.12	0.21
100	Copula	0.01	0.06	0.09	0.21
	QTA	0.00	0.03	0.08	0.14
200	Copula	0.01	0.10	0.17	0.28
	QTA	0.02	0.05	0.10	0.17
300	Copula	0.01	0.04	0.08	0.21
	QTA	0.00	0.03	0.05	0.14
400	Copula	0.01	0.05	0.11	0.21
	QTA	0.01	0.04	0.07	0.14

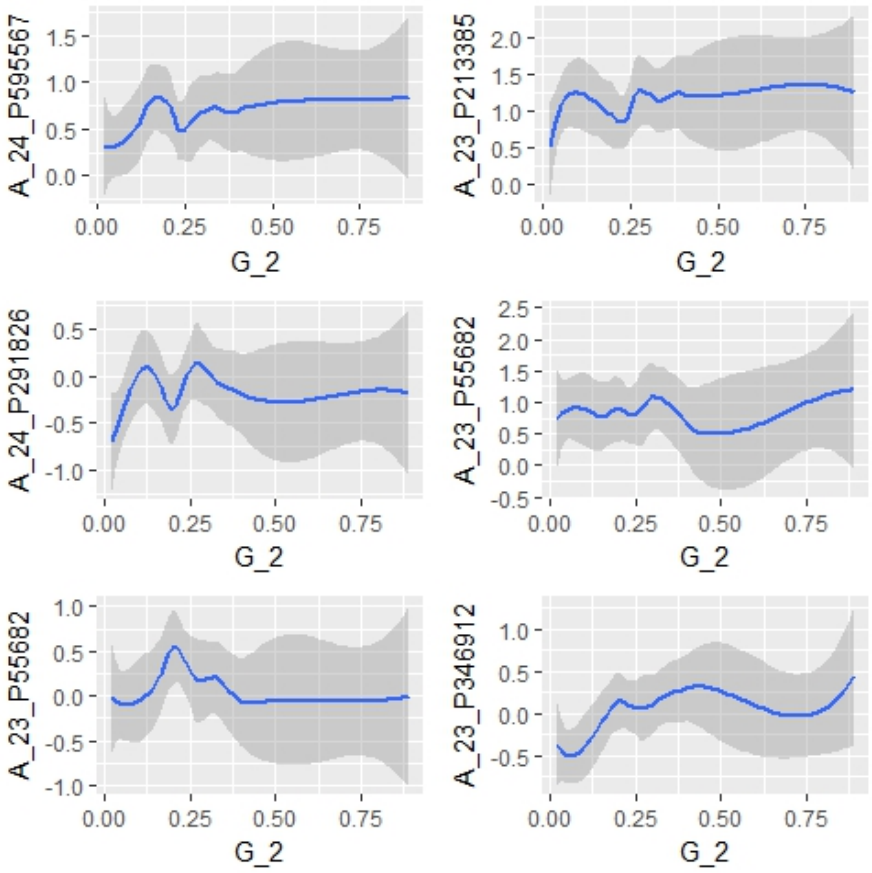

Figure 1: Expression levels of a few genes as a function of the quantitative outcome (G_2). Gene expressions are associated with the G_2 in a nonlinear manner.

parameter reduces to the linear correlation coefficient (Pearson's correlation coefficient). The QTA method calculates its *p*-values based on the correlation coefficients.

3.3. Application to Melanoma Datasets

The assumptions of the QTA method do not always hold, especially for gene expression data. In Figure **1**, none of the randomly selected genes showed a linear relationship with the G_2 checkpoint function.

For the QTA method, we applied the Spearman's correlation coefficient method. The DEGs were selected based on the FDR values calculated from the parametric *p*-values. Table **5** shows the number of DEGs identified by the copula and the QTA methods, based on the melanoma cell lines data. Three uncorrelated continuous outcomes were used. Table **5** indicates that in general, the copula method is more powerful in terms of the identification of DEGs. This validates the simulation results.

The inference procedure for the linear regression in the QTA method is based on the normality assumption of the residuals. The covariate in the regression model is also assumed to be fixed. With these assumptions, the regression model is only valid if the assumption of linearity holds. In case of a random covariate, the estimates of the regression coefficients will be biased. This is not a problem with the copula approach, since no distributional assumptions are required as long as the marginal are continuous. The correlation coefficient applied in the QTA approach measures the overall strength of the association between variables but does not give information about how the association varies across the distribution. It assumes a constant correlation throughout the distribution. In contrast, the copula method looks at where the association is strongest in the distribution. Both the QTA and the copula approach use the permutation approach to calculate p-values. The calculated p-values are then used to generate the FDRs. For the QTA method, it is also possible to calculate p-values based on the well-known correlation coefficients, e.g. the Pearson's correlation coefficient.

4. CONCLUSIONS

This study presents a comparison of a semi-parametric method based on the copula model with a

Table 5: Number of DEGs Identified Using the Copula Method and the QTA Method on the Melanoma Cell Lines Dataset. G_1 Checkpoint Function (G_1), G_2 Checkpoint Function (G_2) and CIN are Used as Quantitative Traits

Quantitative trait	Method	Estimated FDR threshold (α)			
		0.01	0.05	0.1	0.2
G_2	Copula	9	9	9	25
	QTA	0	0	4	56
G_1	Copula	0	0	0	21
	QTA	0	0	0	0
CIN	Copula	0	0	0	21
	QTA	0	0	0	0

parametric method (the QTA method) for finding differentially expressed genes when the outcome is continuous in nature. Both methods performed well in power comparison but the copula approach was notably the better. In terms of the Type I error rate control, the two methods were comparable. We also proposed a simple way of choosing a copula for gene expression studies. This approach was however limited to the copulas that permitted both negative and positive dependence only, and therefore better methods need to be developed.

It is reasonable to conclude that, based on the current study, semi-parametric models outperform their parametric counterparts in noisy high-dimensional data settings like in microarray studies. Here, however, we were limited to the QTA model, but other parametric models do exist (e.g. Bayesian models, etc.). It would therefore be interesting to see how semi-parametric models perform when compared to Bayesian models, in particular, in a future study.

ACKNOWLEDGEMENTS

We are indebted to Dr. William K. Kaufmann for the continuous outcome data used in this study. This work was partially supported by a grant from the Simons Foundation (# 282714 to BO). LC was supported by a scholarship from the African Union.

SUPPLEMENTARY MATERIALS

The supplementary materials can be downloaded from the journal website along with the article.

REFERENCES

[1] Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 2001; 17: 509-519. https://doi.org/10.1093/bioinformatics/17.6.509

[2] Newton MA, Kendziorski CM, Richmond CS, Blattner FR. On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. J Comput Biol 2001; 8: 37-52. https://doi.org/10.1089/106652701300099074

[3] Ibrahim JG, Chen MH, Gray RJ. Bayesian Models for Gene Expression With DNA Microarray Data. J Am Stat Assoc 2002; 97: 88-99. https://doi.org/10.1198/016214502753479257

[4] Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: a Bayesian variable selection approach. Bioinformatics 2003; 19(1): 90-97. https://doi.org/10.1093/bioinformatics/19.1.90

[5] Kendziorski CM, Newton MA, Lan H, Gould MN. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Stat Med 2003; 22: 3899-3914. https://doi.org/10.1002/sim.1548

[6] Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004; 3: Article 3. https://doi.org/10.2202/1544-6115.1027

[7] Scharpf RB, Tjelmeland H, Parmigiani G, Nobel AB. A Bayesian Model for Cross-Study Differential Gene Expression. J Am Stat Assoc 2009; 104: 1295-1310. https://doi.org/10.1198/jasa.2009.ap07611

[8] Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, *et al.* Delineation of prognostic biomarkers in prostate cancer. Nature 2001; 412(6849): 822-826. https://doi.org/10.1038/35090585

[9] Wigle DA, Jurisica I, Radulovich N, Pintilie M, Rossant J, Liu N, *et al.* Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. Cancer Res 2002; 62: 3005-3008.

[10] Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics 2004; 5: 155-176. https://doi.org/10.1093/biostatistics/5.2.155

[11] Owzar K, Jung SH, Sen PK. A Copula Approach for Detecting Prognostic Genes Associated With Survival Outcome in Microarray Studies. Biometrics 2007; 63: 1089-1098. https://doi.org/10.1111/j.1541-0420.2007.00802.x

[12] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001; 98: 5116-5121. https://doi.org/10.1073/pnas.091062498

[13] Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 2001; 96: 1151-1160. https://doi.org/10.1198/016214501753382129

[14] Le CT, Pan W, Lin J. A mixture model approach to detecting differentially expressed genes with microarray data. Funct Integr Genomics 2003; 3: 117-124. https://doi.org/10.1007/s10142-003-0085-7

[15] Pan W. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. Bioinformatics 2003; 19: 1333-1340. https://doi.org/10.1093/bioinformatics/btg167

[16] Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. J Stat Plan Inference 2004; 124: 379-398. https://doi.org/10.1016/S0378-3758(03)00211-8

[17] Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y. Analysis of gene expression data using BRB-Array Tools. Cancer Inform 2007; 3: 11.

[18] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol 1995; 57: 289-300.

[19] Sklar. Fonctions de r'epartition 'a n dimensions et leures marges. Publications de l'Institut de Statistique de L'Universit'e de Paris 1959; 8: 229-231.

[20] Genest C, Ghoudi K, Rivest LP. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 1995; 82: 543-552. https://doi.org/10.1093/biomet/82.3.543

[21] Joe H. Asymptotic efficiency of the two-stage estimation method for copula-based models. J Multivar Anal 2005; 94: 401-419. https://doi.org/10.1016/j.jmva.2004.06.003

[22] Westfall PH, Young SS. Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons 1993; vol. 279.

[23] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 2003; 100: 9440-9445. https://doi.org/10.1073/pnas.1530509100

[24] Kim JM, Jung YS, Sungur EA, Han KH, Park C, Sohn I. A copula method for modeling directional dependence of genes. BMC Bioinformatics 2008; 9: 225. https://doi.org/10.1186/1471-2105-9-225

[25] Yuan A, Chen G, Zhou ZC, Bonney G, Rotimi C. Gene Copy Number Analysis for Family Data Using Semiparametric Copula Model. Bioinform Biol Insights 2008; 2: 343-355.

[26] Fermanian JD. Goodness-of-fit tests for copulas. J Multivar Anal 2005; 95: 119-152. https://doi.org/10.1016/j.jmva.2004.07.004

[27] Wang A. Goodness-of-fit tests for Archimedean copula models. Stat Sin 2010; 20: 441.

[28] Genest C, Quessy JF, Remillard B. Goodness-of-fit Procedures for Copula Models Based on the Probability Integral Transformation. Scand Stat Theory Appl 2006; 33: 337-366. https://doi.org/10.1111/j.1467-9469.2006.00470.x

[29] Dobri J, Schmid F. A goodness of fit test for copulas based on Rosenblatt's transformation. Comput Stat Data Anal 2007; 51: 4633-4642. https://doi.org/10.1016/j.csda.2006.08.012

[30] Berg D. Copula goodness-of-fit testing: an overview and power comparison. Euro J Financ 2009; 15: 675-701. https://doi.org/10.1080/13518470802697428

[31] Genest C, Remillard B, Beaudoin D. Goodness-of-fit tests for copulas: A review and a power study. Insur Math Econ 2009; 44: 199-213. https://doi.org/10.1016/j.insmatheco.2007.10.005

[32] Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr 1974; 19: 716-723. https://doi.org/10.1109/TAC.1974.1100705

[33] Schwarz G. Estimating the dimension of a model. Ann Stat 1978; 6: 461-464. https://doi.org/10.1214/aos/1176344136

[34] Kim JM, Jung YS, Soderberg T. Directional Dependence of Genes Using Survival Truncated FGM Type Modification Copulas. Communications in Statistics - Simulation and Computation 2009; 38: 1470-1484. https://doi.org/10.1080/03610910903009336

[35] Golub GH, Van Loan CF. Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press; 1996.

[36] Kaufmann WK, Nevis KR, Qu P, Ibrahim JG, Zhou T, Zhou Y, *et al.* Defective cell cycle checkpoint functions in melanoma are associated with altered patterns of gene expression. J Invest Dermatol 2008; 128: 175-187. https://doi.org/10.1038/sj.jid.5700935

[37] Kaufmann WK, Carson CC, Omolo B, Filgo AJ, Sambade MJ, Simpson DA, *et al.* Mechanisms of chromosomal instability in melanoma: Chromosomal Instability in Melanoma. Environ Mol Mutagen 2014; 55: 457-471. https://doi.org/10.1002/em.21859

Received on 21-04-2017 Accepted on 07-05-2017 Published on 08-12-2017

https://doi.org/10.6000/1929-6029.2017.06.04.1