Gene Copy Number Analysis for Family Data Using Semiparametric Copula Model

Ao Yuan[1], Guanjie Chen[2], Zhong-Cheng Zhou[3], George Bonney[1] and Charles Rotimi[2]

[1]National Human Genome Center, Howard University, U.S.A. [2]Center for Research on Genomics Global Health, NHGRI, NIH, U.S.A. [3]SuiZhou Central Hospital, SuiZhou, HuBei, 441300 P.R. China.

Abstract: Gene copy number changes are common characteristics of many genetic disorders. A new technology, array comparative genomic hybridization (a-CGH), is widely used today to screen for gains and losses in cancers and other genetic diseases with high resolution at the genome level or for specific chromosomal region. Statistical methods for analyzing such a-CGH data have been developed. However, most of the existing methods are for unrelated individual data and the results from them provide explanation for horizontal variations in copy number changes. It is potentially meaningful to develop a statistical method that will allow for the analysis of family data to investigate the vertical kinship effects as well. Here we consider a semiparametric model based on clustering method in which the marginal distributions are estimated nonparametrically, and the familial dependence structure is modeled by copula. The model is illustrated and evaluated using simulated data. Our results show that the proposed method is more robust than the commonly used multivariate normal model. Finally, we demonstrated the utility of our method using a real dataset.

Keywords: cluster, copula, family data, gene copy number, semiparametric model

Introduction

The gene copy number (also called "copy number variants"—CNV) is the number of copies of a particular gene in the genome of an organism. Recent evidences show that gene copy number (GCN) amplifications and deletions are common characteristics of many genetic diseases. For example, GCN can be elevated in cancer cells as demonstrated in the epidermal growth factor receptor (EGFR) gene in patients with non-small cell lung cancer (Cappuzzo et al. 2005) and also higher copy number of CCL3L1 has been associated with susceptibility to human HIV infection (Gonzalez et al. 2005). Thus identifying these genetic gains and losses provides useful information about specific disease susceptibility or resistance. GCN analysis among normal people within the human genome is also of interest. However, these genetic characteristics are usually not directly observable. Recent technological development in array comparative genomic hybridization (a-CGH) provides scientists with an efficient tool to conduct whole genome and high-density region specific investigation of GCN (Solinas et al. 1997; Pinkel et al. 1998; Snijders et al. 2001).

Briefly, a-CGH technique involves the labeling of genomic DNA from disease tissues (e.g. cancer) and normal control tissue (reference) with different colors (fluorochrome). These samples are then co-hybridize to a metaphase spread from a normal reference cell. After hybridization, emission from each of the two fluorescent dyes is measured, and the signal intensity ratios are indicative of the relative copy number of the two samples. The ratio of the two fluorochrome intensities is then calculated and regions where the disease DNA are amplified or deleted are readily detected on the metaphase spread. The resulting data are in the form of microarrays. This technique not only gives us information about copy number gains and losses in the disease genomic DNA but also allows the identification of the specific chromosomes and the regions of the chromosomes where these changes occurred.

However, the a-CGH data does not provide direct measurements of the GCN changes. Hence, several statistical approaches for analyzing and describing results from these experiments have been developed. Differences exist in these approaches and newer approaches addressing some of the limitations of existing method are needed. For example, some of these methods do not take into account the spatial dependence within the chromosome (Hodgson et al. 2001; Pollack et al. 2002; Cheng et al. 2003; Wang et al. 2004) while others have implemented such dependence structure into their models to enhance the inference

Correspondence: Ao Yuan, National Human Genome Center, Howard University, 2216 Sixth Street, N.W., Suite 206, Washington DC, 20059 U.S.A. Tel: 202-806-4361; Email: ayuan@howard.edu

(Jong et al. 2004; Picard et al. 2004; Fridlyand et al. 2004; Eilers et al. 2004). With the exception of a few that are Bayesian (Barash and Friedman, 2002; Daruwala et al. 2004; Broet and Richardson, 2005), most of the existing methods are frequentist's. All of the existing methods are designed for analyzing data from unrelated persons and are therefore effective in explaining horizontal changes in GCN. However and when available, family data present wonderful opportunity to investigate the vertical kinship effects of GCN as well as the horizontal changes. For this type of data, the main challenge is to model the high dimensional familial dependence structure, and no such approach was found following a careful review of the literature. In this paper, we present such a method in which we used the nonparametric approach to model the marginal distributions and then linked the joint distribution by a copula structure.

Typically, GCN changes observed from a-CGH experiments are classified into three groups corresponding to the three statuses of copy number changes—amplification, deletion and normal; Thus, allowing the microarray responses to have similar features. The practical challenge in the problem that we describe here is that of high dimensionality due to familial dependence among pedigree members. As we indicated above, several of the statistical tools for microarry data clustering deal with low dimensional data (usually one dimensional) and do not take into account the familial dependence among the pedigree members. Such methods can be divided into two main groups, the model based and non-model based (semiparametric). The former assumes specific probability models for the sub-distributions of response in each cluster and is efficient when the specifications are correct but may be seriously biased if implemented specifications deviate from the true unknown models. The semiparametric does not make any assumption about the models except that of a mixing structure, in which the unknown sub-distributions are estimated nonparametrically from the data themselves, and the inference is robust. This method is adequate when the data size is large so that the sub-distributions can be estimated accurately. Yuan and He (2008) proposed such a method for low dimensional microarray clustering for independent data generated from unrelated persons. For data with high dimensionality, the commonly used multivariate normal model rarely fits the actual data, and the nonparametric method is not directly applicable in cluster analysis, so neither of the above models based or non-model based methods are suitable for analyzing the dependence and high dimensionality of family data.

In statistics, the copula is a widely used tool for modeling the dependence structure of high dimensional data (Sklar, 1959; Joe, 1997), and is particularly suitable for pedigree data modeling. Here we propose and implement a semiparametric copula model to address this problem. Specifically, the marginal distributions are estimated nonparametrically, the within pedigree dependence structure is modeled by copula with parameters specified by the kinship coefficients. A penalty term is used against non-unimodality of the sub-distributions. The optimal partition is performed using a classification-estimation (of densities)-maximization-estimation (of parameters) algorithm. The algorithm shares the property of ascending the penalized semiparametric likelihood, just like the well known EM algorithm for ascending the parametric likelihood, and thus, under fair conditions, converges to the optimal partition of the microarray.

The Method

In a-CGH data, the fluorescence ratios between two samples, case and control, are measured across a genomic region. For loci with different copy number changes, the corresponding log-ratio measurement tend to be different. Thus in a-CGH data analysis, often a three-state mixture model is used: deletion state, normal state and amplification state, and we arbitrarily lable them as state 1, 2 and 3. Genes with copy number deletion tend to have smaller log-ratio measurements, those with normal status tend to have moderate measurements, and those with amplification tend to have larger measurements.

We focus on the case of a given chromosome. When there are more than one chromosome under consideration, the method is similar by modeling the chromosomes one by one. Suppose there are n loci of interest and r pedigrees of individuals. The measurement at each locus for each individual is observed. The j-th pedigree has s_j individuals ($j = 1, ..., r$), at locus k, the l-th individual of the j-th pedigree has microarray measurement y_{jkl}. Denote $y_{jk} = (y_{jk1}, \ldots, y_{jks_j})'$ be the measurements of the j-th pedigree at locus k, for each fixed pair (j, k) the y_{jkl}'s are familiarly dependent due to kinship.

Generally this question is formulated as a cluster problem, in which each of the gene locus in classified into one of the clusters B_1, B_2 and B_3 represent the three states deletion, normal and amplification. Let y be a general random vector of the observation y_{jk}'s, a mixture model on y is specified as

$$f(y) = \sum_{i=1}^{3} \alpha_i f(y \mid B_i), \quad (1)$$

Where $f(\cdot|B_i)$ is the sub-density of the responses in the i-th cluster, and the α_i's are the mixing proportions satisfying $0 \leq \alpha_i \leq 1$, $\Sigma_{i=1}^{k} \alpha = 1$. In the literature usually the $f(\cdot|B_i)$'s are specified as multi dimensional normal density functions with cluster specific mean vectors and variance matrices. Typically for this type of pedigree data the dimension is around 10 to 15.

In practice, such high dimensional dependent data hardly conforms to a multivariate normal distribution. A commonly used statistical tool to deal with high dimensional dependence structure is the copula. In this method, it is only necessary to specify each of the marginal densities, and then use a link (copula) to compose all the marginal densities into a joint multivariate density with desired dependence structure. There are large number of copulas to be used, and some optimality criteria to select the best copula for a given problem and data. When the copula is selected, we can incorporate the kinship coefficients among the pedigree members into its dependence structure. Also, there is the question of how to specify the marginal densities. There are various parametric densities to choose from, but if the wrong one is used the results may be seriously biased. On the other hand, for data with large sample size, the nonparametric density can adapt to any distributional feature. Since we do not know the true sub-densities we model each of the marginal densities by nonparametric method for robustness. Finally, a modified BIC criterion is used to select the optimal number of clusters. We describe each of the above steps in different sub-sections below.

The marginal distributions

Since commonly available pedigree data usually consist of three generations and to account for the age and gender difference, the distributions of the measurements are divided into six groups in the following order: first generation male, first generation female, second generation male, second generation female, third generation male and third generation female. We use G_s to denote the s-th group. For example if an individual with observation y_{jkl} is a second generation female in any given pedigree, she is in group 4, we simply denote $y_{jkl} \in G_4$, and so on. Denote $f_s(\cdot|B_i)$ be the sub-density of array cluster i of group s.

Since the $f_s(\cdot|B_i)$s are unknown, they can be estimated by the well known nonparametric estimator (Rosenblatt, 1969)

$$\hat{f}_s(y_{jkl}|B_i) = \frac{1}{n_{is} h_{n_{is}}} \sum_{y_{uvw} \in C_i \cap G_s} K\left(\frac{y_{uvw} - y_{jkl}}{h_{n_{is}}}\right), \quad (2)$$

where n_{is} is the sample size (number of individuals) of group s in cluster i, $K(\cdot)$ is arbitrary given kernel density, and $h_{n_{is}}$ is a given bandwidth to be specified below.

In the density estimation literature, the choice of kernel is not of particular importance (Diggle, 1983; Silverman, 1986). Studies suggest that most unimodal densities perform about the same as the other when used as a kernel, and the choice between kernels can be made on other grounds such as computational efficiency. However, there are some popular options in practice for different reasons. For some general introduction for the choice of kernels,we refer to Silverman (1986) and Scott and Wand (1991). The normal kernel (i.e. $K(\cdot)$ is the density function of the standard normal distribution) is widely used in practice for convenience and other nice features.

In contrast, the choice of bandwidth is crucial in density estimation (Silverman, 1986). Interesting proposals which address this problem can be found in Fan and Gijbels (1992). There is literature on automatic methods that attempt to minimize a lack-of-fit criterion, such as integrated squared error. From Silverman (1986), we choose to use the bandwidth

$$h_{n_{is}} = 0.9 \hat{\sigma}_{is} (n_{is})^{-1/5} \quad (3)$$

where $\hat{\sigma}_{is}^2$ is the empirical variance of the y_{jkl}'s in the s-th group and the i-th cluster.

In the copula formulation we also need the corresponding marginal distribution functions. Let $Fs(\cdot|B_i)$ denote the marginal distribution functions for cluster i and group s, $\hat{F}_s(\cdot \mid B_i)$ for its empirical estimate,

$$\hat{F}_s(y_{jkl} \mid B_i) = \frac{1}{n_{is}} \sum_{y_{uvw} \in C_i \cap G_s} \chi\left(y_{uvw} \leq y_{jkl}\right), \quad (4)$$

where $\chi(\cdot)$ is the indicator function.

The joint distribution

The copula is a commonly used statistical tool to model multivariate joint distribution, it appeared in the early work of Hoeffding, Fréchet and others and formally introduced by Sklar (1959). We first give a very brief account of it and we refer to Hunchinson and Lai (1990); Joe (1997) and Nelson (1999) for detailed review.

A function C defined on $[0, 1]^d$ is a d-variate copula if $C(F_1(x_1), ..., F_d(x_d))$ is a joint distribution function for any marginal distribution functions $F_1(x_1)$, ..., $F_d(x_d)$. The marginal distributions of $C(F_1(x_1), ..., F_d(x_d))$ itself are just $F_1(x_1)$, ..., $F_d(x_d)$. This property provides a convenient way to construct a joint distribution with given marginal ones. On the other hand, given a set of marginal distribution functions $F_1(x_1)$, ..., $F_d(x_d)$, there is a unique copula C such that $C(F_1(x_1), ..., F_d(x_d))$ is the true joint distribution of them (Sklar, 1959). Also, for any joint d-dimensional distribution function $F(\ldots)$, let $F_i^{-1}(\cdot)$ be the quantile functions of the i-th margin, then the function $C(x_1, \ldots, x_d) = F(F_1^{-1}(x_1), \ldots, F_d^{-1}(x_d))$ is a d-variate copula. Let $c(\ldots)$ be the density function (the total derivative) of $C(\ldots)$ when exists. Let $f_i(\cdot)$ be the density function of $F_i(\cdot)$, the density function $f(x_1, ..., x_d)$ of the copula distribution function $C(F_1(x_1), ..., F_d(x_d))$ is given by

$$f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \prod_{i=1}^{d} f_i(x_i). \quad (5)$$

Given a copula, the dependence structure can be characterized in several ways, including Pearson's correlation, Kendall's tau, Spearman's rho, tail dependence, etc. Kendall's tau is generally easier to compute for copulas, so we use this dependence measure. For a two-dimensional copula, Kendall's tau is given by

$$\begin{aligned} \tau &= 4\int_0^1 \int_0^1 C(u, v)c(u, v)dudv - 1 \\ &= 2P((X_1 - X)(Y_1 - Y) > 0) - 1, \end{aligned}$$

where (X_1, Y_1) and (X, Y) are independent with the same distribution. $-1 \le \tau \le 1$, $\tau = 0$ for independence, -1 and 1 for perfect negative and positive dependence. Genest et al. (1995) suggested a pseudo-likelihood approach to estimate the dependence parameters, in which the observed data is transformed via the empirical marginal distributions to obtain pseudo-data that are used in the estimation. Using the special relationships among relative pairs, we can implement the dependence parameters in the copula via the relationships among kinship coefficients, Kendall's tau and the copula dependence parameters without estimation.

For pedigree data, the dependence relationships among familial members (i, j) are best described by the kinship coeffcients, $\gamma_{ij} = \Delta_{7ij}/2 + \Delta_{8ij}/4$, where Δ_{7ij}, Δ_{8ij}, Δ_{9ij} are the condensed kinship coefficient (Jacquard, 1974) between relative pair i and j. The $\Delta_{kij}s$ $(k = 1, ..., 9)$ are the probabilities for the nine possible condensed identical by descent (IBD) status as in Jacquard (1974), in which Δ_{7ij}, Δ_{8ij} and Δ_{9ij} are commonly used in practice. They are the population probabilities of sharing 2, 1 and 0 genes IBD for relative pair (i, j), without regard to their particular genotypes, but only (i, j)'s kinship relationships, under the Mendelian inheritance. Also, $2\gamma_{ij}$ is the expected proportion of gene IBD for relative pair (i, j) at this locus. For convenience we list the values of these coefficients for some common relative pairs (Lange, 1997), and we compute corresponding Kendall's tau in the last column after the computations below.

For trait underlined by single locus or multiple loci, Pearson's correlation for relative pair (i, j) is $2\gamma_{ij}$ (Lange, 1997). Assume that gene copy number change statuses are determined only by the underlying genetic sources, and that the amounts of dependence among them are additive with respect to their shared genetic sources. Then at any fixed locus, Kendall's tau between a fixed type of relative pair (i, j) is (Appendix)

$$\tau_{ij} = 2\Delta_{7ij} + (3/2)\Delta_{8ij} + \Delta_{9ij} - 1. \quad (6)$$

As is true for Pearson's correlation, we postulate that Kendall's tau remain the same, or approximately so, when the trait is influenced by multiple loci. As the kinship coefficients are already known as in Table 1, we get Kendall's tau by the above relationships and in turn, the dependence parameters in the copula model is obtained from the relationship among the dependence parameters and kendal's tau for specified copula. Thus we can easily implement the dependence kinship coefficients in the copula in terms of Kendall's tau without estimating them. For this, we first need to review several commonly used copulas. Note, for family data the dependence are not constant among different relative types, hence copulas with constant dependence parameters, such as Clayton's copula or Frank's copula can not be used here.

Table 1. Kinship coeffcient for selected relative pairs.

Relationship	Δ_7	Δ_8	Δ_9	γ	τ
Grand parent-offspring	0	1/2	1/2	1/8	1/4
Parent-Offspring	0	1	0	1/4	1/2
Half Siblings	0	1/2	1/2	1/8	1/4
Full Siblings	1/4	1/2	1/4	1/4	1/2
First Cousins	0	1/4	3/4	1/16	1/8
Double First Cousins	1/16	6/16	9/16	1/8	1/4
Second Cousins	0	1/16	15/16	1/64	1/32
Uncle-Nephew	0	1/2	1/2	1/8	1/4

Multivariate normal copula

Let $\Phi_d(\cdot, \Theta)$ be the d-variate normal distribution function with mean vector $\mathbf{0}$ and correlation matrix $\Theta = (\theta_{ij})$, $\phi_d(\cdot, \Theta)$ be its density function. $\Phi(\cdot)$ be the one-dimensional standard normal distribution function, and $\Phi^{-1}(\cdot)$ be its quantile (inverse) function. The multivariate normal copula is defined as

$$C(u_1, \ldots, u_d; \Theta) = \Phi_d(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d); \Theta), \quad (u_1, \ldots, u_d) \in [0,1]^d$$

with density

$$c(u_1, \ldots, u_d; \Theta) = \phi_d(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d); \Theta) \prod_{j=1}^{d} (1/\phi(\Phi^{-1}(u_j)).$$

Thus for given marginal distribution functions $F_1(\cdot)$, ..., $F_d(\cdot)$ and their densities $f_1(\cdot)$, ..., $f_d(\cdot)$, the joint distribution function for the multivariate normal copula with these given margins is

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d); \Theta) = \Phi_d(\Phi^{-1}(F_1(x_1)), \ldots, \Phi^{-1}(F_d(x_d)); \Theta)$$

with density

$$f(x_1, \ldots, x_d) = c(x_1, \ldots, x_d) = \phi_d(\Phi^{-1}(F_1(x_1)), \ldots, \Phi^{-1}(F_d(x_d))) \prod_{j=1}^{d} \frac{f_j(x_j)}{\phi(\Phi^{-1}(F_j(x_j)))}.$$

For the distribution in (6), any lower dimensional joint distributions have the same form. For example the (i, j)-th joint distribution function is $F(x_i, x_j) = \Phi_2(\Phi^{-1}(F_i(x_i)), \Phi^{-1}(F_i(x_j)); \Theta_{ij})$, where Θ_{ij} is the (i, j) sub-block of the matrix Θ. From Table 1 and in this case, Spearman's and Kendall's tau are the same. For this copula, Spearman's rho (Kendall's tau) and the dependence parameters θ_{ij}'s in normal copula are related by (Lindskog, 2000)

$$\tau_{ij} = \frac{6}{\pi} \arcsin \frac{\theta_{ij}}{2}, \quad or \quad \theta_{ij} = 2 \sin(\tau_{ij} \pi / 6). \qquad (7)$$

By relationships (6) and (7), the dependence parameters θ_{ij}'s in the multivariate normal copula are easily obtained given the τ_{ij}'s, which are computed via the known kinship coefficients Δ_{kij}'s, as long as we know the kin type of relative pair (i, j).

Multivariate T-copula

Let Θ be the correlation matrix given in the multivariate normal distribution, $x = (x_1, \ldots, x_d)'$ The density function d-dimensional T-distribution with r degrees of freedom is

$$q_r(x_1, \ldots, x_d) = \frac{\Gamma((r+d)/2)}{(r\pi)^{d/2} \Gamma(r/2) |\Theta|^{1/2}} \left(1 + \frac{1}{r} x' \Theta^{-1} x\right)^{-(r+d)/2}.$$

The corresponding copula density is

$$c(u_1, \ldots, u_d) = q_r(Q_r^{-1}(u_1), \ldots, Q_r^{-1}(u_d)) \prod_{j=1}^{d} (1/q_r(Q_r^{-1}(u_j))),$$

where $Q_r(\cdot)$ is the distribution function of the T-distribution with r degrees of freedom, and $q_r(\cdot)$

is its density function. Given marginal distribution functions $F_1(\cdot)$, ..., $F_d(\cdot)$ and their densities $f_1(\cdot)$, ..., $f_d(\cdot)$, the joint density with the copula defined by this multivariate T-distribution is

$$f(x_1, \ldots, x_d) = q_r(Q_r^{-1}(F_1(x_1)), \ldots, Q_r^{-1}(F_d(x_d))) \prod_{j=1}^{d} \frac{f_j(x_i)}{q_r(Q_r^{-1}(F_j(x_j)))}.$$

For this copula, the relationships between the θ_{ij}'s and the τ_{ij}'s are the same as for the multivariate normal copula.

Selection of copula

Given several candidate copulas C_1, ..., C_h with densities c_1, ..., c_h, a natural question is how to select the optimal copula for the data. Let $\hat{F}_{jl}(\cdot)$ be the estimated marginal distribution for individual l in pedigree j (although there are only six different versions of them). For example, if individual (j, l) is in group s, then $\hat{F}_{jl}(\cdot) = \hat{F}_s(\cdot)$. The $\hat{F}_s(\cdot)$'s are defined as

$$\hat{F}_s(y_{jkl}) = \frac{1}{n_s} \sum_{yuvw \in G_s} \chi(y_{uvw} \le y_{jkl}),$$

where n_s is the number of observations in group s.

When there are parameters to be estimated in the copula, the optimal copula can be selected by AIC criteria (Oakes, 1989; Dias and Embrechts, 2003). Here, our copula has no parameters to be estimated, by the likelihood principle and (5), an intuitive way is to select the C with

$$\boldsymbol{C} = \arg\max_{\boldsymbol{i}} \prod_{\boldsymbol{j}=1}^{\boldsymbol{r}} \prod_{\boldsymbol{k}=1}^{\boldsymbol{n}} \boldsymbol{c_i}(\hat{\boldsymbol{F}}_{\boldsymbol{j}1}(\boldsymbol{y}_{\boldsymbol{jk}1}), \ldots, \hat{\boldsymbol{F}}_{\boldsymbol{jsj}}(\boldsymbol{y}_{\boldsymbol{jksj}}))$$

or equivalently, to avoid computation overflow or underflow,

$$\boldsymbol{C} = \arg\max_{\boldsymbol{i}} \frac{1}{\boldsymbol{r}} \sum_{\boldsymbol{j}=1}^{\boldsymbol{r}} \sum_{\boldsymbol{k}=1}^{\boldsymbol{n}} \log \boldsymbol{c_i}(\hat{\boldsymbol{F}}_{\boldsymbol{j}1}(\boldsymbol{y}_{\boldsymbol{jk}1}), \ldots \hat{\boldsymbol{F}}_{\boldsymbol{jsj}}(\boldsymbol{y}_{\boldsymbol{kks}_j})). \quad (8)$$

This is equivalent to choosing the copula with the largest likelihood.

Now for given copula, the joint density for the data $y = \{y_{jkl}\}$ is modeled by

$$\hat{f}(y) = \prod_{j=1}^{r} \prod_{k=1}^{n} \sum_{i=1}^{3} \alpha_i \hat{f}(y_{jk}|B_i) \quad (9)$$

where

$$\hat{f}(y_{jk}|B_i) = c(\hat{F}_{j1}(y_{jk1}|B_i), \ldots, \hat{F}_{js_j}(y_{jks_j}|B_i)) \prod_{l=1}^{s_j} \hat{f}_{jl}(y_{jkl}|B_i).$$

We point out that although we used the same notation c, for different families, the number of individuals may differ and so are the dimensionalities of the c's.

However, under the semi-parametric mixture model assumption, the sub-distributions can take any shape, even the shape of the entire distribution, and as a result any cluster partition will give about the same likelihood value via (9). So optimizing (9) over all possible cluster partitions will not be able to identify the desired clusters. Thus we put some constraints on the selection of clusters such that the sub-distribution is approximately unimodal and optimizing model (9) will give the desired clusters, as in Yuan and He (2008). The reference Yuan and He will be refered to as YH in subsequent citations. However there are two major differences between the method we are proposing and that of YH. Our method can handle high-dimensional data and the link among the marginal densities in copula.

Specifically, let $g(\cdot|B_i)$ be the multivariate normal density with mean given by the sample mean for data in B_i, and covariance matrix Θ, for observations in B_i (i = 1, ..., 3), and denote $g = (g_1, g_2, g_3)$, where $g_i = g(\cdot|B_i)$. g is used as shape constraints for the $\hat{f}(\cdot|B_i)$s. Intuitively, for each fixed i, when the 'correct' partitions are specified, the differences between the $\hat{f}(\cdot|B_i)$s and $g(\cdot|B_i)$s will be relatively small. The Kullback-Leibler divergence $\boldsymbol{D}(\hat{\boldsymbol{f}}(\boldsymbol{B_i}), \boldsymbol{g}(\boldsymbol{B_i}))$ is be used to quantify this difference between the two densities $\hat{f}(\cdot|B_i)$ and $g(\cdot|B_i)$ with $\boldsymbol{D}(\hat{\boldsymbol{f}}(\boldsymbol{B_i}), \boldsymbol{g}(\boldsymbol{B_i})) = \int_{\boldsymbol{B_i}} \hat{\boldsymbol{f}}(\boldsymbol{y}|\boldsymbol{B_i}) \log[\hat{\boldsymbol{f}}(\boldsymbol{y}|\boldsymbol{B_i}) / g(y|B_i)]dy$. Note that $\boldsymbol{D}(\hat{\boldsymbol{f}}(\boldsymbol{B_i}), \boldsymbol{g}(\boldsymbol{B_i}))$ is non-negative and is zero only if $\hat{f}(\cdot|B_i) \equiv g(\cdot|B_i)$. An empirical version of it is given by

$$\boldsymbol{D}(\hat{\boldsymbol{f}}(\boldsymbol{B_i}), \boldsymbol{g}(\boldsymbol{B_i})) = \sum_{\boldsymbol{y_{jk}} \in \boldsymbol{B_i}} \log[\hat{\boldsymbol{f}}(\boldsymbol{y_{jk}}) / \boldsymbol{g}(\boldsymbol{y_{jk}})]$$

and we set

$$\boldsymbol{D}(\hat{\boldsymbol{f}}, \boldsymbol{g}|\boldsymbol{B}) := \sum_{\boldsymbol{i}=1}^{3} \boldsymbol{D}(\hat{\boldsymbol{f}}(\boldsymbol{B_i}), \boldsymbol{g}(\boldsymbol{B_i})).$$

Let $\boldsymbol{L}_0(\alpha \mid \boldsymbol{y}, \hat{\boldsymbol{f}}, \boldsymbol{B})$ be the log-likelihood of (9). Now, instead of optimizing (9), we optimize over all possible partitions of clusters, the penalized log-likelihood,

$$\boldsymbol{L}(\alpha \mid \boldsymbol{y}, \hat{\boldsymbol{f}}, \boldsymbol{g}, \boldsymbol{B}) = \boldsymbol{L}_0(\alpha \mid \boldsymbol{y}, \hat{\boldsymbol{f}}, \boldsymbol{B}) - \lambda \boldsymbol{D}(\hat{\boldsymbol{f}}, \boldsymbol{g} \mid \boldsymbol{B})$$
$$= \sum_{j=1}^{r} \sum_{k=1}^{n} \log \left(\sum_{i=1}^{3} \alpha_i \hat{\boldsymbol{f}}^{1-\lambda}(\boldsymbol{y}_{jk} \mid \boldsymbol{B}_i) \boldsymbol{g}^{\lambda}(\boldsymbol{y}_{jk} \mid \boldsymbol{B}_i) \right), \quad (10)$$

for some $0 \leq \lambda \leq 1$ to be specified. This model can be viewed as an extension of the traditional mixture model. When $\lambda = 0$, it corresponds to a nonparametric specification of sub-distributions, when $\lambda = 1$ it is a full parametric model given by the $g(\cdot|B_i)$s, and when $0 < \lambda < 1$ it corresponds to an intermediate model. By doing this, we are forcing the distributions to be close to normal, more than what is needed for unimodal. The tunning parameter λ is chosen according to simulation for the given type of data. The choice of a multi-variate normal here is for convenience as other choices could be made but may result in additional complication.

The CEME algorithm

However, directly optimizing the mixture model (10) is usually not easy. A common practice of estimating the cluster membership of each observation in the data while evaluating the maximum likelihood estimate $\hat{\alpha}$ of α in (10) is the EM algorithm (Dempster et al. 1977). The EM algorithm is a much easier (though much slower) endeavor computationally than the direct optimization.

For fixed k, let $u_{ij} = 1$ if the i-th locus belongs to the j-th cluster, $u_i = (u_{i1}, u_{i2}, u_{i3})$ be its membership vector, and $u = \{u_{ij}\}$. Treating u as missing data, (y, u) is referred to as the "complete" data. Then the likelihood for the "complete" data is

$$\prod_{j=1}^{r} \prod_{k=1}^{n} \prod_{s=1}^{3} \left(\alpha_s \hat{f}(y_{jk} \mid B_s) \right)^{u_{ks}}.$$

Although we used the same notation $\hat{f}(y_{jk}|B_s)$ for each fixed B_s, the dimension of the data y_{jk} may vary for different pedigree j, as well as the density $\hat{f}(y_{jk}|B_s)$. The corresponding log-likelihood is

$$\boldsymbol{L}_0(\alpha \mid \boldsymbol{y}, \hat{\boldsymbol{f}}, \boldsymbol{u}, \boldsymbol{B}) = \sum_{j=1}^{r} \sum_{k=1}^{n} \sum_{s=1}^{3} \boldsymbol{u}_{ks} \left(\log \alpha_s + \log \hat{\boldsymbol{f}}(\boldsymbol{y}_{jk} \mid \boldsymbol{B}_s) \right).$$

By the same reason as (10), we optimize the penalized "complete data" log-likelihood

$$\boldsymbol{L}(\alpha \mid \boldsymbol{y}, \hat{\boldsymbol{f}}, \boldsymbol{q}, \boldsymbol{u}, \boldsymbol{B}) = \sum_{j=1}^{r} \sum_{k=1}^{n} \sum_{s=1}^{3} \boldsymbol{u}_{ks} \left(\log \alpha_s + \log [\hat{\boldsymbol{f}}^{1-\lambda}(\boldsymbol{y}_{jk} \mid \boldsymbol{B}_s) \boldsymbol{g}^{\lambda}(\boldsymbol{y}_{jk} \mid \boldsymbol{B}_s)] \right), \quad (11)$$

where $g(y_{jk}|B_s)$ is the analogue of $\hat{f}(y_{jk}|B_s)$. The above log-likelihood is optimized iteratively, with the clusters B_s's are classified along each iteration. We specify the starting values at iteration zero as below.

Starting values

Set $\alpha_s^{(0)} = 1/3, \ (s = 1,2,3); u_{ks}^{(0)} = 1/3, \ (s = 1,2,3); k = 1, \ldots, n)$. Divide the n loci into 3 region of roughly equal sizes, and lable them as the $B_s^{(0)}$'s. Let $\hat{f}^{(0)}(\cdot|B_s^{(0)})$ be the nonparametric estimate of $f_j(\cdot)$ using only the measure responses in $B_s^{(0)}$. Denote $(B^{(t)} = B_1^{(t)}, B_2^{(t)}, B_3^{(t)})$ be the estimate of $B=(B_1,B_2,B_3)$ at the t-th iteration of the algorithm.

Given the current t-th iteration estimates $\alpha^{(t)} = \alpha_1^{(t)}, \alpha_2^{(t)}, \alpha_3^{(t)}), u_{ij}^{(t)}$'s, $B(t), f(t)(\cdot B_s^{(t)})$'s and $q^{(t)}(\cdot|B_s^{(t)})$'s from the t-th iteration, we update them in the $(t + 1)$-th iteration according to the following CEME steps.

1. *Classification-step*: Each response locus k, is classified into a candidate cluster $\tilde{B}_s$, if

$$\prod_{j=1}^{r} \left((\alpha_s^{(t)} f^{(t)}(\mathrm{y}_{jk}|B_s^{(t)}) \right) = \max_{1 \leq l \leq 3} \prod_{j=1}^{r} \left((\alpha_l^{(t)} f^{(t)}(\mathrm{y}_{jk}|B_s^{(t)}) \right).$$

   This is the optimal classification rule in the sense of minimizing the expected loss (Anderson, 1984), and it is also the so-called Bayesian assignment. In the cases of ties, we use uniform random assignment among the tied clusters. Let $\tilde{B} = (\tilde{B}_1, \tilde{B}_2, \tilde{B}_3)$ be a candidate classification of the clusters after this step.
2. *Expectation-step*: Let U_{ks}'s be the associated random variables of the u_{ks}'s, and $g^{(t)}(\cdot|B_s^{(t)})$ be the multi-dimensional normal density with mean and covariance matrix empirically estimated from the data in $B_s^{(t)}(s = 1,2,3)$.

As in YH, for $k = 1, ..., n$; $s = 1, 2, 3$, we have

$$\hat{u}_{ks}^{(t+1)} := E\left(U_{ks}|y, \alpha^{(t)}, f^{(t)}, g^{(t)}\right) = \frac{\prod_{j=1}^{r}\left(a_s^{(t)} f^{(t)^{1-\lambda}}(y_{jk}|B_s^{(t)}) g^{(r)^{\lambda}}(y_{jk}|B_s^{(t)})\right)}{\Sigma_{l=1}^{3}\prod_{j=1}^{r}\left(a_l^{(r)} f^{(t)^{1-\lambda}}(y_{jk}|B_l^{(t)}) g^{(t)\lambda}(y_{jk}|B_l^{(t)})\right)},$$

where the expectation is taken with respect to the constrained log-likelihood L. Denote $u^{(t+1)} = \{u_{ks}^{(t+1)}\}$.

3. *Maximization-step*: Compute the MLE $\alpha^{(t+1)}$ of a given $u^{(t+1)}$ as in YH

$$\alpha_s^{(t+1)} = \frac{\Sigma_{k=1}^{n} u_{ks}^{(t)}}{\Sigma_{k=1}^{n}\Sigma_{k=1}^{3} u_{kl}^{(t)}} = \frac{1}{n}\sum_{k=1}^{n} u_{ks}^{(t)}.$$

4. *Estimation-step*: To update the estimation of the density $f^{(t)}(\cdot)$ of current iteration t to $f^{(t+1)}(\cdot)$ for the next iteration $t + 1$, we first compute candidate sub-marginal density $\tilde{f}_{jl}(\cdot|\tilde{B}_s)$ for individual l in pedigree j at locus k and cluster s. If this individual is in group v, then

$$\tilde{f}_{jl}(y_{jkl}|\tilde{B}_s) = \frac{1}{\tilde{n}_{sv}\tilde{h}_{sv}}\sum_{y_{abc}\in\tilde{B}_s\cap G_v} K\left(\frac{y_{abc} - y_{jkl}}{\tilde{h}_{sv}}\right)$$

with

$$\tilde{h}_{sv} = 0.9\tilde{\sigma}_{sv}(\tilde{n}_{sv})^{-1/5}, (j = 1, \ldots, r; l = 1, \ldots, s_j; k = 1, \ldots, n; s = 1, 2, 3; v = 1, \ldots, 6)$$

where $\tilde{n}_{sv}$ is the number of responses for group v in cluster $\tilde{B}_s$, and $\tilde{\sigma}_{sv}^2$ is sample variance of this group. Similarly, the candidate sub-marginal distribution functions are

$$\tilde{F}_{jl}(y_{jkl}|\tilde{B}_s) = \frac{1}{\tilde{n}_{sv}}\sum_{y_{abc}\in\tilde{B}_s\cap G_v}\chi(y_{abc}\le y_{jkl}), (j = 1, \ldots, r; l = 1, \ldots, s_j; k = 1, \ldots, n; s = 1, 2, 3; v = 1, \ldots, 6).$$

Then use (5) to get the candidate sub-joint density for the j-th pedigree at locus k, as follows

$$\tilde{f}(y_{jk}|\tilde{B}_s) = c(\tilde{F}_{j1}(y_{j1k}), \ldots, \tilde{F}_{js_j}(y_{js_jk}))\prod_{l=1}^{r}\tilde{f}_{jl}(y_{jkl}), (j = 1, \ldots, r; k = 1, \ldots, n; s = 1, 2, 3)$$

and that for all the pedigree at locus k is

$$\prod_{j=1}^{r}\tilde{f}(y_{jk}|\tilde{B}_s), (k = 1, \ldots, n; s = 1, 2, 3).$$

Let $\tilde{g}$ be the reference densities corresponding to $\tilde{B}$.

We update the quadruple $(B^{(t+1)}, f^{(t+1)}, F^{(t+1)}, g^{(t+1)})$ as

$$(B^{(t+1)}, f^{(t+1)}, F^{(t+1)}, g^{(t+1)}) = \begin{cases}(\tilde{B}, \tilde{f}, \tilde{F}, \tilde{g}), \\ (B^{(t)}, f^{(t)}, F^{(t)}, g^{(t)})\end{cases}$$

$$\begin{cases}\text{if } L(\alpha^{(t+1)}|y, \tilde{f}, \tilde{g}, \tilde{B}) \ge L(\alpha^{(t)}|y, f^{(t)}, g^{(t)}, B^{(t)}), \\ \text{otherwise.}\end{cases}$$

The estimate of $f(\cdot)$ at the $(t + 1)$-th iteration is then

$$\hat{f}^{(t+1)}(\cdot) = \sum_{s=1}^{3}\hat{\alpha}_s^{(t+1)}\hat{f}^{(t+1)}(\cdot|B_s^{(t+1)}).$$

Note that at each iteration t, $\alpha^{(t)}$, $B^{(t)}$, and the $u_{ij}^{(t)}$s are updated, but not necessarily so for $f^{(t)}$ and $g^{(t)}$.

The above four steps are iterated until convergence of $(\alpha^{(t)}, f^{(t)}, B^{(t)})$. (Note by the following Proposition, we may check the stability of the $\alpha^{(t)}$ as a simple criterion for the convergence of the triple). We may use the relative error criterion for the convergence of the ($\alpha^{(t)}$'s, that is, for some prespecified $\delta > 0$, we stop the iteration when $\Sigma_{s=1}^{3}\left|(\alpha_s^{(t+1)} - \alpha_s^{(t)})/\alpha_s^{(t)}\right| \le \delta$. Typically, $\delta \le 0.01$.

As in YH, we have

Proposition. *For each fixed k, the sequence* $\{L(\alpha^{(t)}|y, f^{(t)}, g^{(t)}, B^{(t)})\}$ *is increasing in t*, and *there is a stationary point* (α^*, f^*, B^*) *satisfying*

$$\prod_{j=1}^{r}\left(\alpha_s^* f^*(y_{jk}|B_s^*)\right) = \max_l\prod_{j=1}^{r}\left(\alpha_l^* f^*(y_{jk}|B_l^*)\right), \quad \forall k \in B_s^* (s = 1, 2, 3).$$

When (α^*, f^*, B^*) *is the unique stationary point, we have, as* $t \to \infty$,

$$(\alpha^{(t)}, f^{(t)}, B^{(t)}) \to (a^*, f^*, B^*).$$

Application

Simulation study

We simulate $r = 10$ pedigrees, each has four individuals, father, mother and two sibs, and we assume there are $n = 200$ loci of interest, which are divided into 3 clusters as $B_1 = (1, 80)$, $B_2 = (81, 150)$ and $B_3 = (151, 200)$, with cluster means $\mu_1 = (4.9, 4.2)$, $\mu_2 = (9.9, 9.2)$ and $\mu_3 = (14.9, 14.2)$ for male and female individuals. We generated two datasets by simulation using the normal copula and multi-normal models. Each of the datasets were analyzed using both normal copula and multi-normal models.

To simulate data from the Multivariate normal copula model, let A be the Cholesky decomposition of Θ. To sample from this copula distribution: for $k = 1, ..., n$ and $i = 1, ..., 5$

1. generate r independent samples $\boldsymbol{Z}_{1k}, ..., \boldsymbol{Z}_{rk}$ from $N(\mathbf{0}, I_4)$.
2. Let $\boldsymbol{u}_{lk} = A\boldsymbol{Z}_{lk}$ $(l = 1, ..., r)$.
3. For $k \in B_i$, if j is for male, set $x_{ljk} = \Phi(u_{l1k}) + \mu_{k1}$; if j is for female, set $x_{ljk} = \Phi(u_{l1k}) + \mu_{k2}$, and $x_{lk} = (x_{l1k}, ..., x_{l4k})$, where $\Phi(\cdot)$ is the distribution function of the standard normal. Then $x_{1k}, ..., x_{rk}$ is a sample from the 4-variate normal copula model with correlation matrix Θ. The results are displayed in Table 2 below, with tuning parameter λ of values 0.25, 0.5, 0.75 and 1.

In the above, $\lambda = 1$ corresponds to a normal model, and $0 < \lambda < 1$ correspond to a mixed model. For this type of data, the model has difficulty in parameter convergence for small values of λ, reflecting the fact that the multivariate data distribution is too noisy for nonparametric part of the model to work alone; thus a parametric unimodal component is needed to help cluster the data. The normal copula model has larger likelihood value in all these cases. This means the normal copula model is more robust than the multivariate normal model. For the data from multi-normal model, when $\lambda = 0.5$, 38 loci from cluster three are classified to cluster two. Over all, $\lambda = 0.75$ performs well for all the data set, and so we recommend this value of λ in this analysis.

To assess the robustness of the method, we simulated larger data sets with family sizes of 4 and 5, each with 100 families and 200 candidate loci. The simulated clusters and means for male and female are: cluster one, 1–70, (4.9, 4.2); cluster two, 81–150, (8.9,8.2) and cluster three, 171–200, (12.9,12.2) respectively. To reflect some complexity we added minor clusters to some of the clusters. The means for male and female for the minor clusters are 71–80, (5.4, 4.7) and 151–170, (12.4,11.7). The results are summarized in Table 3.

Overall, the results are consistent: the smaller the value of λ, the better the model fitness, as indicated by larger likelihood value. This means that the nonparametric model component capture the data distribution in fine details. But in many cases, the computation breaks down for $\lambda = 0$ as pointed out earlier. It is seen that for either the data generated from multi-normal or normal copula distributions, the overall performances of the semiparametric model is robust for a range of the tuning parameter λ.

Real data analysis

We use the proposed method to analyze the Genetics Analysis Workshop 15 (GAW15) data

Table 2. Cluster results for normal copula and multi-normal models for 10 pedigrees and 200 loci.

Data	λ	Cluster 1	Cluster 2	Cluster 3	Log-likelihood
Normal Copula	0.25	1–80	81–150	151–200	–11438825.76
	0.50	1–80	81–150	151–200	–22088970.51
	0.75	1–80	81–150	151–200	–32749199.84
	1.00	1–80	81–150	151–200	–43412062.82
Multi-normal Model	0.50	1–80	81–150(+38)	151–200(–38)	–3117275.81
	0.75	1–80	81–150	151–200	–3644388.69
	1.00	1–80	81–150	151–200	–4652713.94

Table 3. Summary cluster results from normal copula and multi-normal data sets for 100 pedigrees with 4 or 5 Family Members.

Pedigree size	Data	λ	Cluster 1	Cluster 2	Cluster 3	Log-likelihood
4	Normal Copula	0.25	1–80	81–150	151–200	–5503651.81
		0.50	1–80	81–150	151–200	–10207757.93
		0.75	1–80	81–150	151–200	–14924762.04
		1.00	1–80	81–150	151–200	–19645406.52
4	Multi-normal Model	0.50	1–80	81–150(+9)	151–200(–9)	–1716690.87
		0.75	1–80	81–150	151–200	–2213287.79
		1.00	1–80	81–150	151–200	–2726810.52
5	Normal Copula	0.25	1–80(–7)	81–150(+7)	151–200	–7841636.12
		0.50	1–80(–7)	81–150(+7)	151–200	–15364258.65
		0.75	1–80(–7)	81–150(+7)	151–200	–22940970.23
		1.00	1–80(–6)	81–150(+6)	151–200	–28643031.41
5	Multi-normal Model	0.25	1–80	81–150	151–200	–1639830.86
		0.50	1–80	81–150	151–200	–2250676.24
		0.75	1–80	81–150	151–200	–2874104.71
		1.00	1–80	81–150	151–200	–3503763.43

set with 14 pedigrees of CEPH Utah families, each with three generations and about a dozen normal individuals. Expression level of genes in lymphoblastoid cells of the above subjects were obtained using the Affymetrix Human Focus Arrays that contain probes for 8,500 transcripts. Gene copy number variations in normal people within human genome has been the subject of study (Freeman et al. 2006; Pugh et al. 2008). For 3,554 of the 8,500 SNPs tested, Morley et al. (2004) found greater variation among individuals than between replicate determinations on the same individual. These 3,554 expression phenotypes (expressed genes) were chosen for copy number change analysis. The first step is to find out the best copula model for the data. We considered three different models, the multi-normal model, the semi-parametric multivariate normal-copula model, and the semi-parametric multivariate T-copula model. Then the criterion in (8) is used to select the optimal model. The average copula likelihood values for the three models are –3217389.15, –2094272.97, –2296408.96 respectively. Thus the semi-parametric multivariate normal-copula model is the best of the three and was used for clustering. The outcome of the analysis of the GAW15 data is displayed in the figure below. The horizontal axis represents the sequential numbering of genes from 1 to 3550, and the vertical axis indicates the classified states of the genes with 1, 2 and 3 representing deletion, normal and amplification.

As shown in the figure, most of the SNPs are in clusters 1 and 3, this observation is consistent with the large variation of the expression levels. The SNPs with deletion status are more likely to be contained in cluster 1, and those with amplification status are more likely to be in cluster 3.

Concluding Remarks

We proposed, studied and demonstrated a semiparametric copula method for microarray-SNP genomewide association analysis using pedigree data. We successfully implemented the kinship relationship into the model for more robust analysis of family data than the commonly used multivariate normal model.

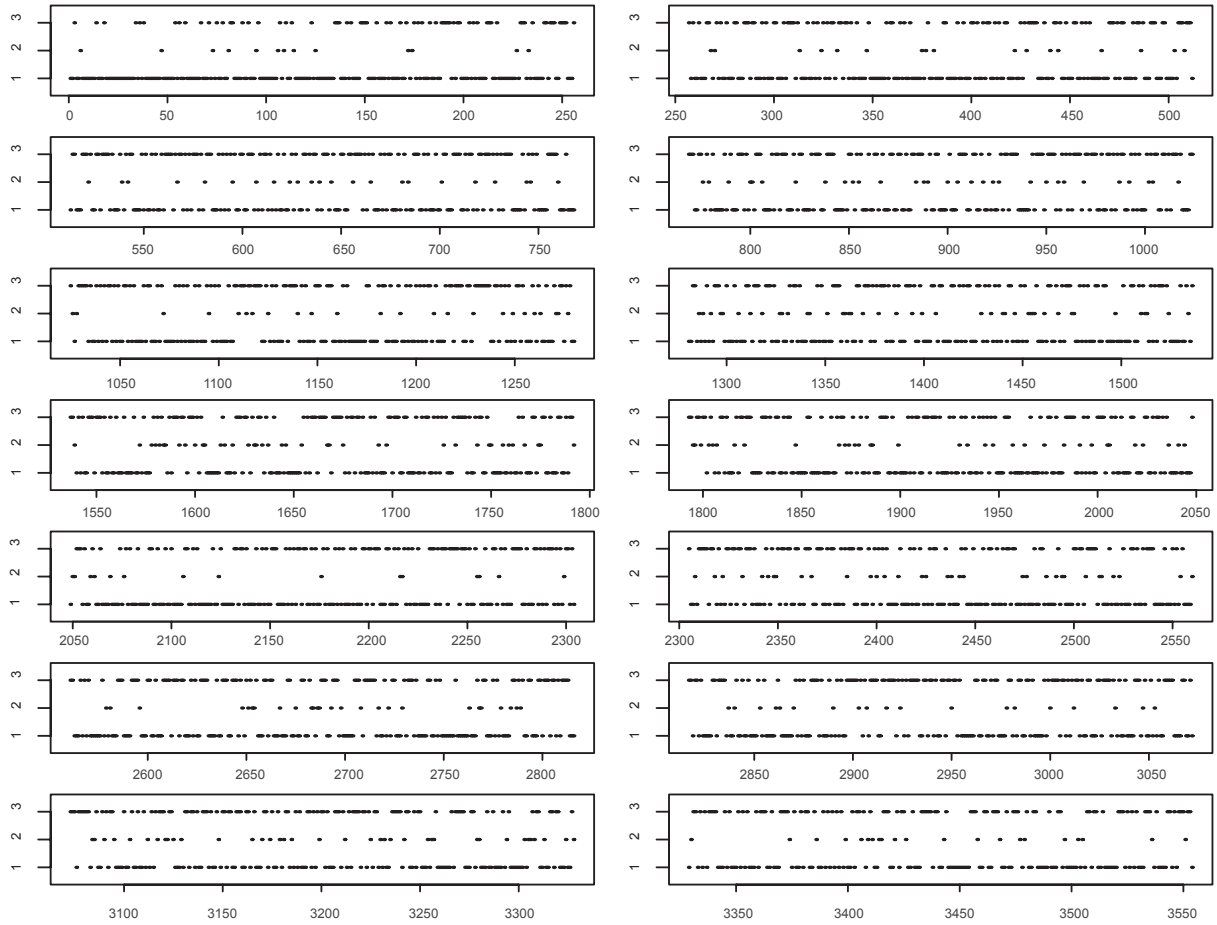

Figure 1.

Acknowledgement

This work is supported in part by the National Center for Research Resources at NIH grant 2G12RR003048, and by the Center for Research on Genomics and Global Health (CRGGH) at NHGRI/NIH.

Disclosure

The authors report no conflicts of interest.

References

Anderson, T.W. 1984. An Introduction to Multivariate Statistical Analysis. Wiley.

Barash, Y. and Friedman, N. 2002. Context-specific Bayesian clustering for gene expression data. *Journal of Computational Biology*, 9:169–91.

Broet, P. and Richardson, S. 2005. CGHmix: detection of gene copy number changes in CGH microarrys using a spatially correlated mixture model, manuscript.

Cappuzzo, F., Varella-Garcia, M., Shigematsu, H., Domenichini, I., Bartolini, S., Ceresoli, G., Rossi, E., Ludovini, V., Gregorc, V., Toschi, L., Franklin, W., Crino, L., Gazdar, A., Buun, P. and Hirsch, F. 2005. Increased HER.2 gene copy number is associated with response to gefitinib therapy in epidermal growth factor receptor-positive non-small-cell lung cancer patients. 23(22):5007–18.

Carr, D.B., Somogyi, R. and Micheals, G. 1997. Templates for looking at gene expression clustering. *Statistical Computing and Statistical Graphics Newslatter*, 8:20–9.

Cheng, C., Kimmel, R., Neiman, P. and Zhao, L.P. 2003. Array rank order regression analysis for the detection of gene copy-number changes in human cancer. *Genomics*, 82:122–9.

Daruwala, R.S., Rudar, A., Ostrer, H., Lucito, R., Wigler, M. and Mishra, B. 2004. A versatile statistical analysis algorithm to detect genome copy number variation. *Proc. Natl. Acad. Sci. U.S.A.*, 101:1629216297.

Dempster, A.P., Laird, N.M. and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Ser. B.*, 39:1–38.

Dias, A. and Embrechts, P. 2003. Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance*, 3:1–14.

Diggle, P.J. 1983. Statistical Analysis of Spatial Point Patterns. Academic Press, London. Eilers PH, De Menezes RX. 2004. Quantile smoothing of array CGH data. Bioinformatics, Nov 30 Epub.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 99:14863–8.

Fan, J. and Gijbels, I. 1996. Local Polynomial Modeling and Its Applications. London: Chapman and Hall.

Fellenberg, M. and Mewes, H.W. 1999. Intercepting clusters of gene expression profiles in terms of metabolic pathways. In Proceedings of the German Conference on Bioinformatics' 99. Poster.

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., Carter, N.P., Scherer, S.W. and Lee, C. 2006. Copy number variation: New insights in genome diversity. *Genome Research*, doi:10.1101/gr.3677206.

Fridlyand, J., Snijders, A., Pinkel, D., Albertson, D.G. and Jain, A. 2004. Hidden Markov models to the analysis of the array CGH data. *Journal of Multivariate Analysis*, 90:132–53.

Genest, C., Ghoudi, K. and Rivest, L.P. 1995. A semiparametric estimation procedure of dependence parameters in a multivariate families of distributions. *Biometrika.*, 82:543–52.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R., Freedman, B., Quinones, M., Bamshad, M., Murthy, K., Rovin, B., Bradley, W., Clark, R., Anderson, S., O'Connell, R., Agan, B., Ahuja, S.S., Bologna, R., Sen, L., Dolan, M. and Ahuja, S.K. 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, 307(5714):1422.

Hodgson, G., Hager, J.H., Volik, S., Hariono, S., Wernick, M., Moore, D., Nowak, N., Albertson, D.G., Pinkel, D., Collins, C., Hanahan, D. and Gray, J.W. 2001. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics*, 29:459–64.

Hunchinson, T.P. and Lai, C.D. 1990. Continuous Bivariate Distributions, Emphasizing Applications. Rumsby Scientific Publishing, Adelaide.

Jacquard, A. 1974. The genetic structure of populations. New York: Springer-Verlag.

Joe, H. 1997. Multivariate Models and Dependence Concepts, Chapman and Hall, London.

Jong, K., Marchiori, E., Meijer, G., Vaart, A.V. and Ylstra, B. 2004. Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, 20:3636–7.

Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F. and Pinkel, D. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083):818–21.

Lange, K. 1997. Mathematical and statistical methods for genetic analysis, Springer-Verlag.

Lindskog, F. 2000. Modeling dependence with copulas and applications to risk management. Master thesis, Swiss Federal Institute of Technology Zurich.

Morley, M., Molony, C.M., Weber, T., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. 2004. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430:743–7.

Oakes, D. 1994. Multivariate survival distributions. *Journal of Nonparametric Statistics*, 3:343–54.

Picard, F., Robins, S., Lavielle, M., Vaisse, C. and Daudin, J.J. 2004. A statistical approach for CGH microarray data analysis. http://web.inapg.fr/ens-rech/mathinfo/recherche/mathematique/outil-A.html.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B.M., Gray, J.W. and Albertson, D.G. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20(2):207–11.

Pollack, J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Borrensen-Dale, A.L. and Brown, P. 2002. Microarray analysis reveals a major direct role of DAN. copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. U.S.A.*, 99:12963–8.

Pugh, T.J., Delaney, A.D., Farnoud, N., Flibotte, S., Griffith, M., Li, H.I., Qian, H., Farinha, P., Gascoyne, R.D. and Marra, M.A. 2008. Impact of whole genome amplification on analysis of copy number variants. *Nucleic Acids Research*, 36:No.13e80.

Rosenblatt, M. 1969. Conditional probability density and regression estimators, in P.R. Krishnaiah, ed., Multivariate Analysis II, Academic Press, New York, 25–31.

Scott, D.W. and Wand, M.P. 1991. Feasibility of multivariate density estimates. *Biometrika*, 78:197–205.

Silverman, B. 1986. Density Estimation for Statistics and Data Analysis, Chapman and Hall. Sklar A. 1959. Fonctions de répartition á n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–31.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D. and Albertson, D.G. 2001. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, 29(3):263–4.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T. and Lichter, P. 1997. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, 20(4):399–407.

Wang, J., Meza-Zepeda, L.A., Kresse, S.H. and Myklebost, O. 2004. M-CGH: analyzing microarray-based CGH experiments. *BMC Bioinformatics*, 74:1–4.

Yuan, A. and He, W. 2008. Semi-parametric clustering methods with applications to microarray data analysis. *Journal of Bioinformatics and Computational Biology*, 6:261–82.

Appendix

Proof of (6): Let (X_1, Y_1) be the traits of relative pair (i,j); (X, Y) be an independent copy of (X_1, Y_1); and A_l be the event that a relative pair share l alleles IBD (l = 0, 1, 2). Given a relative pair (i, j), by definition $P(A_2) = \Delta_{7ij}$, $P(A_1) = \Delta_{8ij}$, and $P(A_0) = \Delta_{9ij}$. By the assumption that GCN change is determined by the underlying genetic source and given A_2, the pair (i, j) share the same genetic source at the locus and the same copy number change status; thus $X_1 = Y_1$, $X = Y$. Note that the random variables X_1 and X are of continuous type and $P((X_1 - X)\ (Y_1 - Y).\ 0|A_2) = P((X_1 - X)^2.\ 0|A_2) = 1$. Also, $X_1 - X$ and $Y_1 - Y$ are random variables symmetric around 0, thus given A_0, $X_1 - X$ and $Y_1 - Y$ are independent, so the events $(X_1 - X)(Y_1 - Y).\ 0$ and $(X_1 - X)(Y_1 - Y),\ 0$ are completely random, each with probability 1/2, i.e. $P((X_1 - X)(Y_1 - Y)|A_0) = 1/2$. By the additivity assumption, given A_1, the probability of the event $(X_1 - X)(Y_1 - Y).\ 0$ is the average of those for cases of given A_2 and A_0, i.e. $P((X_1 - X)(Y_1 - Y).\ 0|A_1) = 3/4$. Then at any fixed locus, Kendall's tau between a fixed type of relative pair (i, j) is

$$
\begin{aligned}
\tau_{ij} &= 2P((X_1 - X)(Y_1 - Y) > 0 \,|\, A_2)\ P(A_2) \\
&\quad + 2P((X_1 - X)(Y_1 - Y) > 0 \,|\, A_1) P(A_1) \\
&\quad + 2P((X_1 - X)(Y_1 - Y) > 0 \,|\, A_0) P(A_0) P(A_0) - 1 \\
&= 2 \times 1 \times \Delta_{7ij} + 2 \times (3/4) \times \Delta_{8ij} + 2 \times (1/2) \times \Delta_{9ij} - 1 \\
&= 2\Delta_{7ij} + (3/2)\Delta_{8ij} + \Delta_{9ij} - 1.
\end{aligned}
$$