Testing for Associations of Opposite Directionality in a Heterogeneous Population

Fangyuan Zhang[1] · Jie Ding[2] · Shili Lin[3]

Received: 21 February 2016 / Revised: 12 August 2016 / Accepted: 13 September 2016 /
Published online: 22 September 2016

Abstract In gene networks, it is possible that the patterns of gene co-expression may exist only in a subset of the sample. In studies of relationships between genotypes and expressions of genes over multiple tissues, there may be associations in some tissues but not in the others. Despite the importance of the problem in genomic applications, it is challenging to identify relationships between two variables when the correlation may only exist in a subset of the sample. The situation becomes even less tractable when there exist two subsets in which correlations are in opposite directions. By ranking subset relationships according to Kendall's tau, a tau-path can be derived to facilitate the identification of correlated subsets, if such subsets exist. However, the current tau-path methodology only considers the situation in which there is association in a subsample; the more complex scenario depicting the existence of two subsets with opposite directionality of associations was not addressed. Further, existing algorithms for finding tau-paths may be suboptimal given their greedy nature. In this paper, we extend the tau-path methodology to accommodate the situation in which the sample may be drawn from a heterogeneous population composed of subpopulations portraying positive and negative associations. We also propose the use of a cross entropy Monte Carlo procedure to obtain an optimal tau-path, CEMC$_{tp}$. The algorithm not only can provide simultaneous detection of positive and negative correlations in the same sample, but also can lead to the identification of subsamples that provide evi-

Electronic supplementary material The online version of this article (doi:10.1007/s12561-016-9167-7) contains supplementary material, which is available to authorized users.

✉ Shili Lin
shili@stat.osu.edu

[1] Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX, USA

[2] Department of Medicine and Genetics, Stanford University, Stanford, CA, USA

[3] Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

dence for the detected associations. An extensive simulation study shows the aptness of CEMC$_{tp}$ for detecting associations under various scenarios. Compared with two standard tests for detecting associations, CEMC$_{tp}$ is seen to be more powerful when there are indeed complex subset associations with well-controlled type-I error rates. We applied CEMC$_{tp}$ to the NCI-60 gene expression data to illustrate its utility for uncovering network relationships that were missed with standard methods.

Keywords Cross entropy Monte Carlo (CEMC) · Tau-path · Heterogeneous sample · Subset associations · Gene networks

1 Introduction

There is often a need to identify relationships, which may or may not be linear, between two variables. However, it is usually unknown, a priori, whether a relationship holds over an entire sample, or only in a subset of the sample. Under some circumstances, even more complex situations can arise, in which there may be associations of opposite directionality in two subsets, with the potential of yet other observations portraying no association, all in the same sample. In other words, the sample is heterogeneous in the association pattern, representing a mixture of homogeneous subpopulations of different association directions and strengths. A number of examples in which such scenarios may occur were discussed in the literature, including the study of association between high-density lipoprotein (HDL) and the risk of myocardial infarction [1,2]. In gene regulatory networks, it is possible that the patterns of gene co-expression exist only in a subset of the sample [3]. For example, the activation of a gene may have a positive, negative, or neutral effect on another gene, depending on the underlying cells, cancer types, or other conditions. As such, the expressions of two genes over a set of cells may exhibit heterogeneity, with different association directions over a different subset of cells. As another example, the relationships between single nucleotide polymorphisms and gene expressions over multiple tissue types can be mined from the GTEx database [4], but such relationships are likely to be heterogeneous, as an association may exist in some tissues, but not in the other tissues.

A number of measures are frequently used to study association between two variables. Pearson's correlation coefficient is useful for measuring linear relationships [5,6]. Spearman's rank correlation [7] and Kendall's tau [8,9], on the other hand, are both nonparametric statistics used to measure the degree of monotonic association between two rankings without the assumption of linearity. However, the correlation signals of all these measures will be weakened or canceled when uncorrelated observations or correlated observations with opposite directions exist in the same sample. This situation will occur when the observations are sampled from a heterogeneous population containing homogeneous subpopulations with monotonic relationships.

The originally proposed tau-path offers a solution for detecting correlation that exists only in a subset of the sample [10,11]. It is a procedure based on a sequential development of Kendall's tau measure of monotone association. The "optimal" sequence is achieved by reordering the observations so that the sample tau coefficients $\{\tau_k\}$ for the first k (=2, ..., n) of the n bivariate observations form a "maximum" monotonic decreasing path, ending at the usual Kendall's tau coefficient τ $(=\tau_n)$ over all the n observations. Tau-path not only can indicate whether there is measurable evidence of association in a subsample, but also can identify which subsample is involved. Nevertheless, current algorithms for finding the tau-path are suboptimal given its greedy nature [10]. Another shortcoming of the existing tau-path approach is that it was proposed for detecting association of a single direction in a subset. As such, the more complex correlation scenario as described above is not explored or discussed.

To address these issues, in this paper, we present an efficient optimization algorithm, Cross-Entropy Monte Carlo tau-path, CEMC$_{tp}$, for finding an optimal tau-path based on an objective criterion without resolving to exhaustive search. CEMC is an approach originally proposed to solve difficult combinatorial problems [12]. Although CEMC has mainly been used to solve problems in engineering and computer science, it was successfully adapted for tagging SNP selection [13] and for rank aggregation of results from multiple studies in genomics [14]. In addition to tackling the optimization problem, we also extend the tau-path methodology to address the scenario of two subsamples with opposite association directions in the same sample. Simulation results demonstrate the validity of CEMC$_{tp}$ and show that it can be more powerful compared to other methods when the underlying population is heterogeneous. Importantly, when the underlying population is homogeneous, there appears to be little loss in efficiency. Finally, to illustrate the applicability of CEMC$_{tp}$ to genomic studies, we used it to analyze the NCI-60 gene expression data, leading to the identification of several potential gene networks that were missed using traditional approaches.

2 Methods

2.1 Tau-Path

Suppose $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is a sample of n pairs of observations from a population characterized by a pair of random variables (X, Y) that may have different correlation patterns in different subpopulations. Let $p = (p(1), p(2), \ldots, p(n))$ be a permutation that reorders the original observation sequence $(1, 2, \ldots, n)$ such that $p(k) = l$ denotes that the original observation l is now the kth ordered element under permutation p. We use S to denote the collection of all possible permutations; thus the cardinality of the set S is $||S|| = n!$. For each permutation $p \in S$, define

$$\tau_k(p) = \frac{2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \text{sign}[(x_{p(j)} - x_{p(i)})(y_{p(j)} - y_{p(i)})]}{k(k-1)}, \quad k = 2, \ldots, n.$$

The sequence $\tau(p) = (\tau_2(p), \tau_3(p), \ldots, \tau_n(p))$ is called the tau-path under permutation p [10]. The goal is to find a particular permutation p^* for which the elements in the tau-path are "sequentially maximal monotone decreasing", that is, $\tau_2(p^*) \geq \tau_3(p^*) \geq \cdots \geq \tau_n(p^*)$, with each $\tau_k(p^*)$ being the maximum for the reordered elements up to that point [10]. The rationale for seeking such a permutation is to facilitate the detection of a subset correlation, with the uniform full-set association (i.e., association in a homogeneous sample) being a special case. To see this, first note that the standard Kendall's $\tau(\equiv\tau_n(p))$ is the same for all permutations $p \in S$. Therefore, if all observations come from a population portraying an association between the two variables uniformly, $\tau_n(p^*)$ will be the statistic used to assess the evidence of existence of such an association. On the other hand, if the sample comes from a heterogeneous population composed of subpopulations, then the tau-path of permutation p^* provides evidence of decreasing strength of a positive association (we will multiply the observations for one of the two variables by a negative sign to assess negative correlation). If the strength degrades substantially after a certain point, then this provides evidence for a subset association, and the observations supporting such an association can then be identified.

2.2 Tau-Score

We first note that p^* may not be unique [10]. Further, the greedy nature of the algorithms proposed thus far does not guarantee that a p^* will be obtained. We consider an alternative criterion as an approximation to the objective. Specifically, we define a tau-score for each permutation p as $T(p) = \sum_{k=2}^{n} \tau_k(p)$. Our restated goal is then to find a permutation p^* that leads to $T(p^*)$ achieving the maximum among all $p \in S$. In other words, we seek p^* =argmax$\{T(p),\ p \in S\}$. To gain a better understanding of the tau-score measure, we note that $T(p)$ is in fact a weighted average of the concordance (+1) or discordance (-1) contribution from each pair of observations to the tau-path. Specifically, the weight for sign$[(x_{p(j)} - x_{p(i)})(y_{p(j)} - y_{p(i)})]$ is $2(1/j - 1/n)$ for all $i = 1, \ldots, j-1$, which is inversely "proportional" to the order of observations under the permutation. Therefore, the tau-score gives more weight to elements that rank earlier under the permutation.

2.3 CEMC for Finding an Optimal Permutation

Finding the tau-path with the largest tau-score is combinatorial in nature as the number of possible permutations is $n!$, and thus an exhaustive search will not be tractable even for a sample of moderate size. To determine the best order through optimizing $T(p)$, we propose to adopt the CEMC approach similar to the earlier adaptations [13,14].

The idea is to "sort" the observations through estimating the probability of each observation being in a particular position (rank). We accomplish this by organizing the probabilities into a matrix for applying the CEMC algorithm. Specifically, we consider random matrix $\boldsymbol{Z} = (Z_{jr})_{n*n}$, in which each row/column vector is composed of 0's and exactly one 1 at a random position. We use $\boldsymbol{v} = (v_{jr})_{n\times n}$ to denote the corresponding probability matrix, in which the probabilities in each column always sum to 1. This can be interpreted as each column of $\boldsymbol{Z}$ being an independent (across columns) realization of a multinomial distribution whose probability vector is the corresponding column in the $\boldsymbol{v}$ matrix. Thus, the probability mass function for $\boldsymbol{Z}$ can be specified as follows:

$$P_v\{\mathbf{Z} = z = (z_{jr})_{n\times n}\}$$
$$\propto \prod_{r=1}^{n}\prod_{j=1}^{n}(v_{jr})^{z_{jr}} \times I\left(\sum_{j=1}^{n} z_{jr} = 1, r = 1, \ldots, n;\quad \sum_{r=1}^{n} z_{jr} = 1, j = 1, \ldots, n\right).$$

A realization of $\mathbf{Z}$, z, uniquely determines the corresponding candidate order of the observations (that is, determining the underlying permutation) through a deterministic function $f(z)$ without the need to reference the probability matrix. That is, for each column r, the row (corresponding to an observation) with the "1" entry identifies the observation that takes the rth rank (position in the permuted sequence) under the permutation, $r = 1, 2, \ldots, n$. Given the $1-1$ correspondence between p and z, finding p^* is equivalent to finding z^* that maximizes $T(f(z))$ [13,14]. Detailed description of an efficient algorithm for finding the optimal order can be found in [14] and is summarized in Appendix A for the current article to be self-contained.

2.4 CEMC$_{tp}$ Algorithms for Determining Positive and Negative Association

We describe the steps of a CEMC tau-path algorithm for detecting (sub)samples that are correlated in the two variables. This algorithm is designed to detect associations in scenarios where the population may be homogeneous or heterogeneous. We use $X = \{x_1, \ldots, x_n\}$ to denote the n observations from the first component of the bivariate random variable, while $Y = \{y_1, \ldots, y_n\}$ denotes the n corresponding observations from the second component.

1. Generate m permutations of Y and denote them as $Y^{(1)}, \ldots, Y^{(m)}$. In our examples below, we set $m = 500$, as that appears to be sufficient for obtaining type-I error rates close to the nominal values.
2. (a) Generate positive tau-paths: use CEMC OEA (Appendix A) to determine the optimum tau-path for each of the $m + 1$ data sets $(X, Y), (X, Y^{(1)}), \ldots,$ and $(X, Y^{(m)})$, and denote them as $\tau_+ = \left(\tau_2^{(+)}, \ldots, \tau_k^{(+)}, \ldots, \tau_n^{(+)}\right), \tau^{(+l)} = \left(\tau_2^{(+l)}, \ldots, \tau_k^{(+l)}, \ldots, \tau_n^{(+l)}\right), l = 1, 2, \ldots, m$.
   (b) Generate negative tau-paths: determine the optimum tau-path for each of the $m + 1$ data sets $(X, (-1) \times Y), (X, (-1) \times Y^{(1)}), \ldots$, and $(X, (-1) \times Y^{(m)})$, and denote them as $\tau_- = \left(\tau_2^{(-)}, \ldots, \tau_k^{(-)}, \ldots, \tau_n^{(-)}\right), \tau^{(-l)} = \left(\tau_2^{(-l)}, \ldots, \tau_k^{(-l)}, \ldots, \tau_n^{(-l)}\right), l = 1, 2, \cdots, m$.
3. (a) Calculate the p value for positive association. First, find the smallest upper quantile q_+ for τ_+ along the path. Specifically, suppose $q_+(k)$ is the upper quantile of $\tau_k^{(+)}$ among $\tau_k^{(+1)}, \ldots, \tau_k^{(+m)}, k = 2, \ldots, n$. Then $q_+ = \min\{q_+(k), k = 2, \ldots, n\}$. The p value, p_+, is then defined as the proportion of all the tau-paths (out of the total of $m + 1$) that have an upper quantile at least as small as $q+$ in at least one of the positions ($k = 2, \ldots, n$) along the path.
   (b) Similarly, find the p value, p_-, for negative association based on the $m + 1$ negative tau-paths τ_- and $\tau^{(-l)}, l = 1, \ldots, m$.

4. (a) Assessing evidence of significance for positive association and obtaining the associated subsample. If $p_+ \leq \alpha/2$ for a predetermined pathwise significance level α, then we declare the detection of significantly positive association (in at least a subset of the sample). If positive association is detected, we find $k^* = \text{argmin}_k\{q_+(k), k = 2, \dots, n\}$, and $\{p^*(1), \dots, p^*(k^*)\}$ is the (sub)sample that support the evidence of positive association, where p^* is the optimal permutation of the observed sample from the OEA.
   (b) Assessing evidence of significance for negative association and obtaining the associated subsample. Similarly as above, if $p_- \leq \alpha/2$, then we assert detection of negative association, and we find the subsample that supports such a detection.

Note that instead of finding the p values, one may construct the upper and lower path-wise confidence bounds instead. The tau-path τ_+ breaking the upper bound at any position along the path will lead to the conclusion of positive association while tau-path τ_- breaking the lower bound at any position along the path will lead to the conclusion of negative association. These two procedures are equivalent and will lead to the same conclusion.

The above algorithm provides a procedure for detecting the existence of positive and negative associations within a sample controlling for path-wise false positives. Moreover, the observations that support the evidence of such associations are also identified. However, to compare with existing methods that only detect monotonic (but not necessarily linear) association, we also propose a slightly modified algorithm that combines evidence from both positive and negative associations. This modification takes place in Steps 3 and 4 of the CEMC$_{tp}$ algorithm; details are given in Appendix B. The R package for performing CEMC$_{tp}$ analysis can be downloaded from http://www.stat.osu.edu/~statgen/SOFTWARE/CEMCtp.

2.5 Three Methods for Comparisons

2.5.1 Tau-Score Method

The CEMC$_{tp}$ algorithm as described in the previous section can be computationally intensive as one needs to perform path-wise evaluation to control for false positives. As such, we consider a variation, the tau-score method, CEMC$_{ts}$, in which the calculation of p values is based on the tau-score, not the tau-path, and thus can be more computationally efficient. More specifically, in Step 3 of the algorithm, we first compute the tau-scores $T_+, T^{(+l)}, l = 1, \dots, m$; the upper quantile of T_+ among the $m + 1$ tau-scores is then taken as the p value for detecting positive association. The p value for detecting negative association can be found analogously. The procedure for assessing overall (combined) evidence of association using CEMC$_{ts}$ follows the same idea as described in Appendix B.

2.5.2 Two Conventional Methods

In addition to CEMC$_{ts}$, we also compare CEMC$_{tp}$ with two conventional methods for calculating correlation: Pearson's correlation coefficient and Kendall's tau. The assessment of significance for these two methods are based on the same permuted samples as in the CEMC methods to construct the underlying null distributions. Since both methods are devised for detecting evidence of monotonic association in a homogeneous population, they are compared to the versions of CEMC$_{ts}$ and CEMC$_{tp}$ that are appropriate for assessment of overall evidence of association.

3 Simulation Study

3.1 Simulation Models and Settings

To evaluate the performance of CEMC$_{tp}$, and to compare its performance with CEMC$_{ts}$, Pearson correlation, and Kendall's tau, we carried out a simulation study based on a variety (a total of 16) of homogeneous/heterogeneous population settings. Each of the 16 population settings were characterized by two distributions, with potentially different association patterns to create subpopulations. One is a standard bivariate normal distribution with density function

$$f(t_1, t_2) = \frac{|A|^{1/2}}{2\pi} exp\left(-\frac{1}{2}\sum_{i,j=1}^{2,2} A_{i,j}t_it_j\right). \tag{1}$$

The other is a standard bivariate t distribution with the degree of freedom being 1 and the density function being

$$f(t_1, t_2) = \frac{|A|^{1/2}}{2\pi}\left(1 + \sum_{i,j=1}^{2,2} A_{i,j}t_it_j\right)^{-3/2}. \tag{2}$$

In both (1) and (2), $A = \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix}$, and $\Sigma = A^{-1} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ where ρ is the correlation parameter. When $\rho = 0$, the two normally distributed variables are independent, whereas the two t variables are still dependent although uncorrelated.

The 16 settings with various proportions of samples being positively associated, negatively associated, or uncorrelated are given in Table 1. In the first 8 settings, the samples are mixtures of positively correlated and uncorrelated samples. That is, for each of these settings, there exists only one subset of the sample with the two variables positively correlated. For example, in the first setting, there are a total of 60 pairs of observations, among which ten are sampled from a subpopulation depicted by a bivariate normal or t distribution with a correlation coefficient of 0.9, and the remaining 50 are drawn from a subpopulation in which the two random variables are uncorrelated. In the last 8 settings, negatively correlated samples are also included. In other words, there are two subsets in the sample, one positively, while the other negatively, correlated. For example, in setting 12, among the 120 pairs of observations, 40 are drawn from a subpopulation in which the two random variables are positively associated with a correlation of 0.9; 40 are sampled from another subpopulation where the two

Table 1 Simulation settings portraying various patterns of population heterogeneity

Setting	Positive		Negative		Uncorrelated
	No.	ρ	No.	ρ	No.
1	10	0.9	0	0	50
2	10	0.6	0	0	50
3	30	0.9	0	0	30
4	30	0.6	0	0	30
5	50	0.9	0	0	10
6	50	0.6	0	0	10
7[a]	60	0.9	0	0	0
8[a]	60	0.6	0	0	0
9	60	0.9	60	−0.6	0
10	80	0.9	40	−0.6	0
11	40	0.9	80	−0.6	0
12	40	0.9	40	−0.6	40
13	80	0.9	20	−0.6	20
14	20	0.9	80	−0.6	20
15	40	0.9	20	−0.6	60
16	20	0.9	40	−0.6	60

[a] Settings 7 and 8 are in fact from a homogeneous population with all 60 samples positively correlated

variables are negatively associated with a correlation of −0.6; and the remaining 40 are from yet another subpopulation in which the two variables are uncorrelated. Thus, the sample in setting 12 comes from a heterogeneous population composed of three homogeneous subpopulations. In addition, to obtain the type-I error rates, we also consider two settings (not shown in Table 1), with either 60 (matching those in settings 1–8) or 120 (matching those in settings 9–16) pairs of uncorrelated observations.

3.2 Results for Detecting Association

Figure 1 depicts the performances of Pearson correlation, Kendall's tau, CEMC$_{ts}$, and CEMC$_{tp}$ for settings 1–8 when the sample is a mixture of positively correlated and uncorrelated observations. Specifically, Fig. 1a shows the power to identify the existence of correlation among the normally distributed samples when correlation may exist only in a subset. We can see that no matter which method is used, the power increases as the proportion of correlated observations increases. The Pearson correlation has higher power when the correlation is moderate (0.6). When the correlation is strong (0.9), CEMC$_{ts}$ and CEMC$_{tp}$ outperform the other two methods for all the mixing proportions considered, although we note that the differences are all small. This result is consistent with earlier results when Kendall's tau and tau-path were compared [10]. We hypothesize that this occurs because the estimated ordering is likely to be highly variable with only moderate strength of association. On the other hand, CEMC$_{ts}$ and CEMC$_{tp}$ are better choices when there is a more highly associated subpopulation, even if the subpopulation is small. The type-I error rates are 0.039, 0.033, 0.043 and 0.05 for Pearson, Kendall, CEMC$_{ts}$, and CEMC$_{tp}$, respectively, indicating

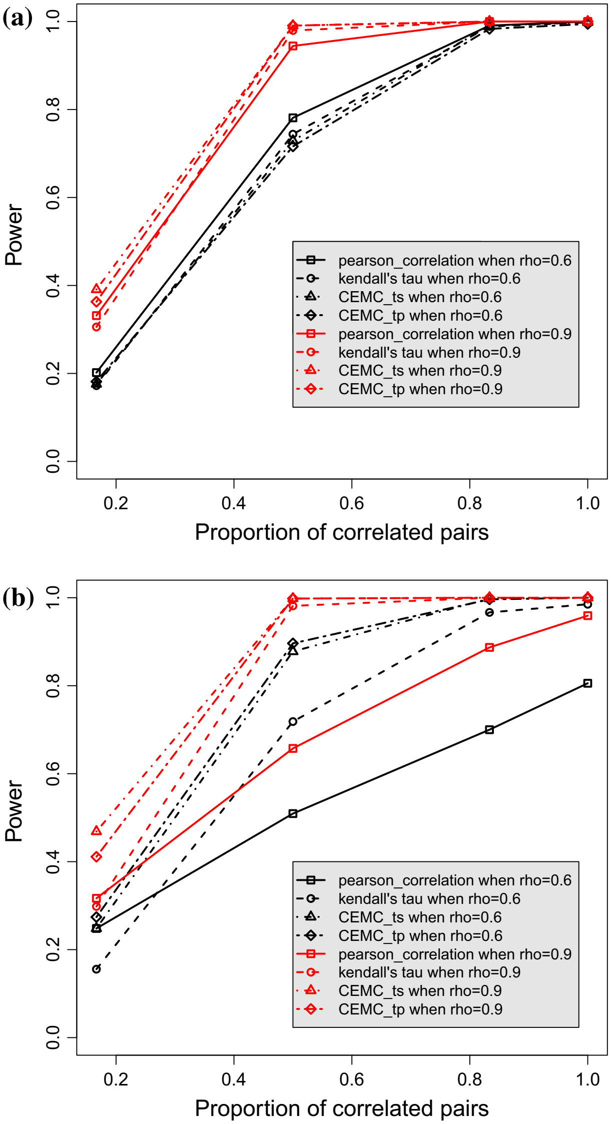

Fig. 1 Power of detecting associations based on four measures: Pearson's correlation coefficient, Kendall's tau, CEMC$_{ts}$, and CEMC$_{tp}$ for samples from mixtures of **a** normal distributions and **b** *t* distributions

that all the four methods have a well-controlled type-I error rate at the 5 % nominal level.

Figure 1b shows the power for identifying the existence of correlation among the *t* distributed samples. Similar to the results for samples from the normal distributions, the power of all four methods increases as the proportion of correlated pairs increases. However, contrary to the normal results, CEMC$_{ts}$ and CEMC$_{tp}$ give higher power than Pearson correlation or Kendall's tau for both levels of correlation strength (0.6 and 0.9) and all the mixing proportions.

The type-I error rates are 0.052, 0.059, 0.037 and 0.044 for the four methods, Pearson, Kendall, CEMC$_{ts}$, and CEMC$_{tp}$, respectively. Even though CEMC$_{ts}$ is more computationally advantageous, its performance is largely on par with CEMC$_{tp}$ for the simulation settings considered.

Figure 2 shows the results for settings 9–16 when the samples are mixtures of positively correlated, negatively correlated, and uncorrelated pairs of observations. The results for each of the eight settings are represented by the eight sets of circles, with the center of each set depicting the corresponding setting. The (perpendicular) distances from the center to the three sides ('+", "−" and "uncorrelated") of the equilateral triangle are the numbers of observations that are positively correlated ($\rho = 0.9$), negatively correlated ($\rho = -0.6$) or uncorrelated, respectively. Note that they sum to 120, as the total sample size is 120 for all the eight settings. For example, setting 9 has an equal number of positively and negatively correlated, but no uncorrelated, observations. Thus, the center of the corresponding circles sits in the middle of the "uncorrelated' side, that is, with an equal distance (60) to both the "+" and the "−" sides. The size (radius) of each circle shows the power of the corresponding method.

Figure 2a shows the power for identifying the existence of correlation among the normally distributed samples. The circles on the right side of the triangle are generally larger than those at symmetric positions on the left side, and the circles at the bottom are generally larger than those on the top, implying that more samples with stronger correlations increases the detection power for all the methods. The two circles each with a radius close to 1 on the right bottom corner show that all methods have the highest power when most observations are correlated strongly in one direction. In almost all the settings, CEMC$_{ts}$ and CEMC$_{tp}$ obtain the same or higher power than Pearson's correlation and Kendall's tau. This is because subsets with opposite directions may cancel each other's signals when Pearson correlation and Kendall's tau are used to summarize the evidence of association over the entire sample as a whole. CEMC$_{ts}$ and CEMC$_{tp}$, on the other hand, are robust to mixtures of associations of different directions. The type-I error rates are 0.037, 0.039, 0.028 and 0.037 for the four methods in the same order as before, once again demonstrating the ability of all methods to achieve the correct size.

Figure 2b shows the power for identifying the existence of correlation among the *t* distributed samples. Similar to the results from the normally distributed settings, all methods achieve higher power with more observations strongly correlated in one direction. In particular, Pearson's correlation, CEMC$_{ts}$ and CEMC$_{tp}$ exhibit high power consistently in several settings. In contrast, the power of Kendall's tau varies substantially from setting to setting. The type-I error rates are 0.052, 0.051, 0.052 and 0.037 for Pearson, Kendall, CEMC$_{ts}$, and CEMC$_{tp}$, respectively. Contrary to the results from a single correlated subsample, when there are two correlated subsamples of different directions, the results from CEMC$_{ts}$ and CEMC$_{tp}$ can be a bit more different, especially for the normal samples. In the five instances where there are appreciable differences, CEMC$_{tp}$ has better power in four of them (settings 11, 12, and 16 in the normal samples and setting 16 in the *t* sample).

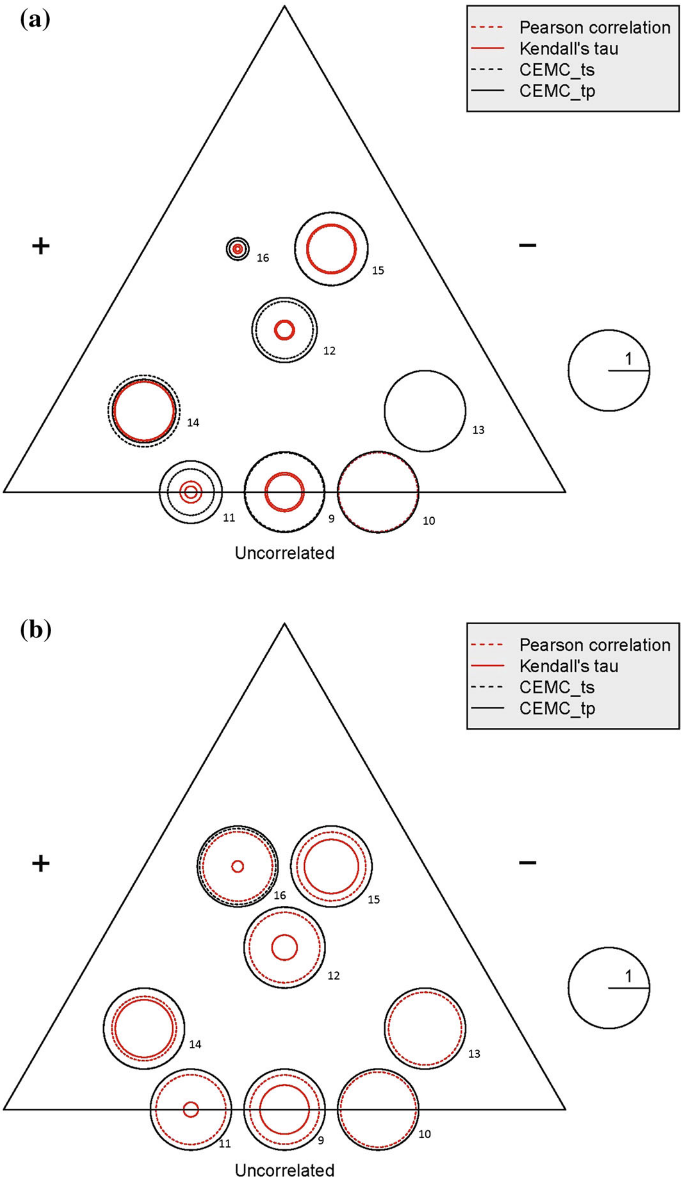

Fig. 2 Power for detecting associations of different directionalities based on four measures: Pearson's correlation coefficient, Kendall's tau, CEMC$_{ts}$, and CEMC$_{tp}$ for samples from mixtures of **a** normal distributions and **b** t distributions.

Finally, to show that our permutation procedure for obtaining the p values is efficient, we also used samples simulated under the null settings to construct the reference distribution for the test statistics. That is, we drew samples from uncorrelated normal or t distributions instead of permuted samples to calculate the p values. The results corresponding to Figs. 1 and 2, are shown in Supplementary Figures S1–S4, from which one can see that the results are very similar between those based on the null samples and those based on the permuted samples.

3.3 CEMC$_{tp}$ Inference on Association with Directional Information

As we have already pointed out earlier, a feature of CEMC$_{tp}$ that separates it from the rest is that it can detect existence of associations of opposite directionality in a sample from a heterogeneous population. Recall that whether a sample is from a homogeneous or a heterogeneous population is unknown, a priori, and thus detection power is an issue that deserves further consideration. To better understand the detection power of CEMC$_{tp}$, we considered an expanded collection of settings (the first 6 columns of Table 2), which includes the eight settings (9–16) in Table 1 that portray a mixed sample of positively and negatively correlated observations.

The power for simultaneous detection of positive and negative correlations in the same sample are given in the last four columns of Table 2. From the table, we can see that when there are the same number of observations supporting associations in both the positive and negative directions (settings 9a and 9c), there is a fairly even power for detecting both types of associations for samples from the normal as well as the *t* bivariate distributions. When there is strong association in one direction but only weak association in the other direction, the power for detecting association can be quite different. For example, in the situations where the number of associated observations are the same or the stronger association is in fact supported by more observations (settings 9, 9b, 10, 12, 12a, 13, and 15), the stronger association is detected with much higher power, especially when the samples are from a mixture of normal distributions. On the other hand, when the stronger association is supported by fewer observations,

Table 2 Simulation settings and the power for detecting associations using CEMC$_{tp}$

Setting	Positive		Negative		Uncor.	Normal		*t*	
	No.	ρ	No.	ρ	No.	Posi.	Nega.	Posi.	Nega.
9	60	0.9	60	−0.6	0	0.99	0.05	1.00	0.57
9a	60	0.9	60	−0.9	0	0.96	0.98	1.00	1.00
9b	60	0.6	60	−0.9	0	0.04	1.00	0.67	1.00
9c	60	0.6	60	−0.6	0	0.15	0.20	0.92	0.84
10	80	0.9	40	−0.6	0	1.00	0.00	1.00	0.08
11	40	0.9	80	−0.6	0	0.50	0.41	0.91	0.99
12	40	0.9	40	−0.6	40	0.78	0.05	1.00	0.55
12a	40	0.6	40	−0.9	40	0.05	0.85	0.70	0.99
13	80	0.9	20	−0.6	20	1.00	0.00	1.00	0.01
14	20	0.9	80	−0.6	20	0.02	0.77	0.36	1.00
15	40	0.9	20	−0.6	60	0.90	0.00	1.00	0.26
16	20	0.9	40	−0.6	60	0.17	0.11	0.79	0.92

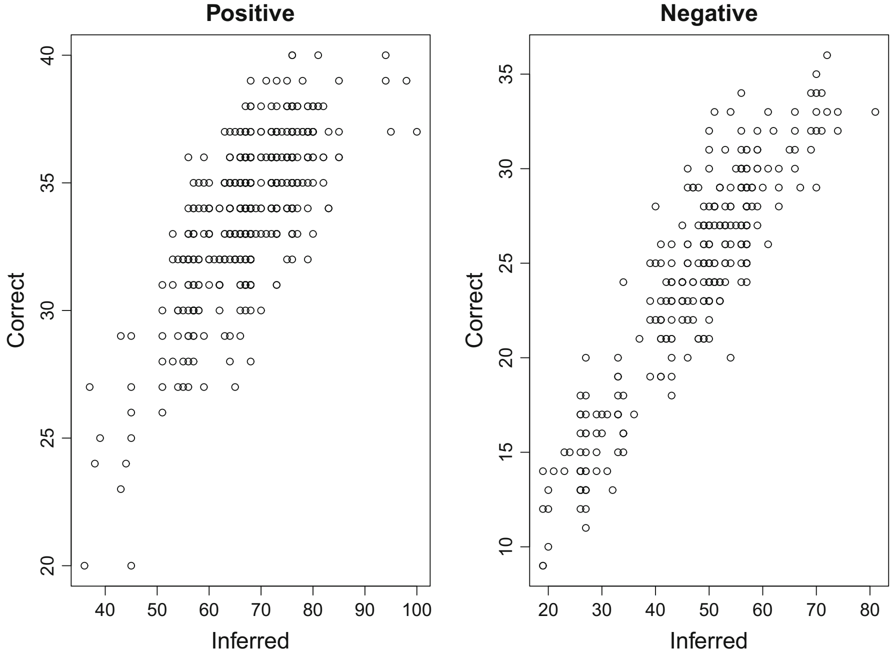

Fig. 3 Scatter plots of number of observations inferred to be associated versus number of correctly inferred observations for samples from a mixture of t distributions. *Left* positive association; *Right* negative association

the detection power may even out (as in settings 11 and 16) or can actually be swung to the other direction with extremely lopsided sample sizes (as in setting 14). In contrast to the ability of CEMC$_{tp}$ for detecting associations in both directions in a heterogeneous sample, the Pearson's correlation and the standard Kendall's tau measures may fail to detect the overall association (e.g., see settings 11 and 12 in Fig. 2a and settings 11 and 16 in Fig. 2b). Overall, from Table 2, we can see that for each of the settings, the sample from a mixture of t distributions tends to have higher power than that from a mixture of normal distributions for detecting associations of both directions; in some cases the power can be much larger. Once again, the results are practically the same when p values were computed based on null samples rather than permuted samples (Supplementary Table S1).

Another unique feature of CEMC$_{tp}$ is that, when positive and/or negative associations are detected, the observations that provide such evidence can be inferred as well. To demonstrate this capability, we show the results for setting 12 when the samples were generated from a mixture of t distributions. The scatter plots for the number of inferred observations versus the number of correct observations (i.e., the inferred observation is indeed generated from the corresponding distribution) when the association is detected are provided in Fig. 3. As seen from the figure, the observations that were inferred to support the detected association, either positive or negative, may contain some that were not actually generated from the corresponding distribution. In fact, the ratio of inferred versus correct observations is about two to one, which is not surprising given the results presented in Figs. 1 and 2, as a guideline. Given that the Pearson's correlation has reasonably good power of detecting the overall association, many of the unassociated observations may be inferred to be associated in either direction. Results from the other settings painted similar pictures.

To further understand and visualize these results, in Fig. 4b we plotted a sample dataset from setting 12 with the t-mixture. We can see that, excluding the point on the bottom-left corner, the rest of the data points show a random cloud, which explains the low power for standard Kendall's tau (Fig. 2b). The higher power for Pearson's correlation appears to be driven by the influential point on the bottom-left corner, creating an apparent positive linear association. On the other hand, CEMC$_{tp}$ was able to identify a subset of observations that are positively associated and another subset that are negatively associated, recovering the underlying setting. As seen from the figure, a majority of the true positively correlated data points were correctly identified, while there are a few uncorrelated or negatively correlated data points that were incorrectly inferred to be in the positively correlated subsample, corroborating the results seen in Fig. 3. On the other hand, although many of the negatively correlated data points were also correctly identified, there are about one third of them that were not correctly inferred, which is also not surprising given that the strength for negative association (-0.6) is weaker than that for positive association (0.9). A sample dataset from setting 12 with the normal-mixture is also visualized to facilitate better understanding of the result in Figs 2 and 3, and Table 2. One can see from Fig. 4a that there is little information on association based on the whole set, leading to low power for the standard Kendall's tau and the Pearson correlation. Further, a majority of the positively and negatively correlated data points were correctly inferred, although many data points from the uncorrelated subsample were incorrectly inferred to be negatively associated.

4 A Real Data Analysis—Gene Network Inference

We considered NCI-60, which is a panel of 60 diverse human cancer cell lines used by the National Cancer Institute (NCI) to study a variety of issues. The NCI-60 data include the expression measures of 12,625 genes for each of the 60 cell lines, and can be downloaded from an NCI-hosted FTP site [15]. Since biochemical functions are determined largely by specific enzymes, genes in the same network may be turned on or off differently in different cell lines. Thus, for each pair of genes, the positively correlated expression (co-expression) or the negatively correlated expression patterns may exist only in a subset of the cell lines. To investigate whether CEMC$_{tp}$ can recover such heterogeneous network relationships that may have been missed by other methods, we apply Pearson's correlation, Kendall's tau, CEMC$_{ts}$, and CEMC$_{tp}$ to the data to identify potential gene pairs and to compare their performances. To be more focused in our illustration and comparison, we first computed the Kendall's tau for every pair of genes and binned them into 20 groups according to the overall strength of association: $[-1, -0.9]$, $(-0.9, -0.8]$, ..., $(-0.1, 0]$, $(0, 0.1]$,..., and $(0.9, 1]$. We then randomly selected 10 pairs from each of the bins for investigation and comparisons

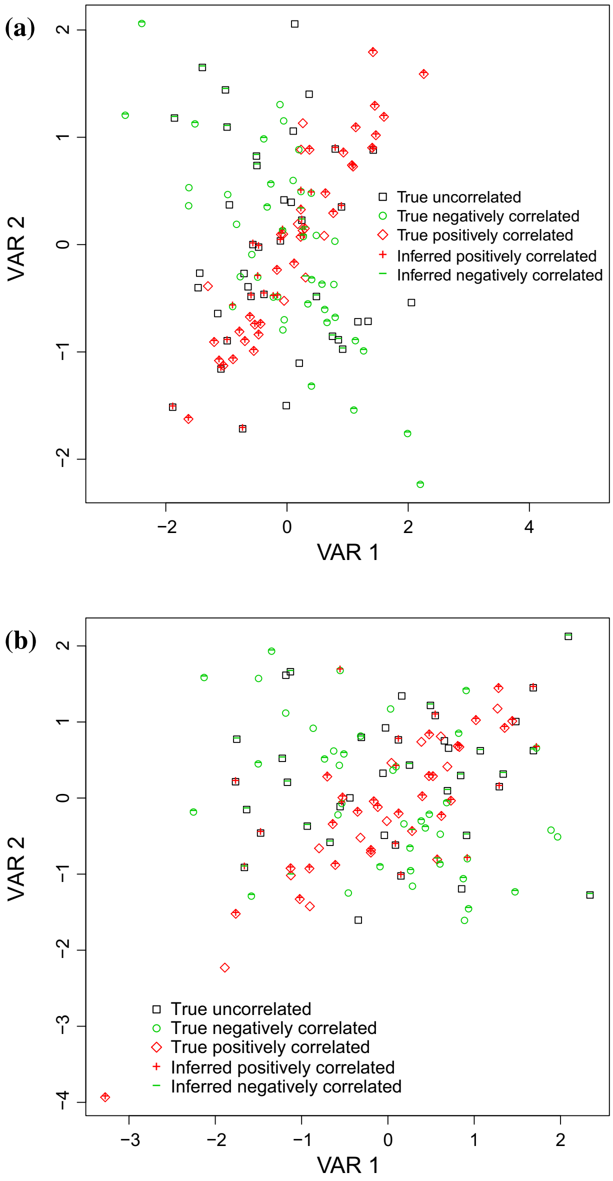

Fig. 4 A sample dataset with three underlying subsamples (*square* uncorrelated; *diamond* positively correlated; *circle* negatively correlated) and inferred positively correlated (+) and negatively correlated (−) subsets for samples from mixtures of **a** normal distributions and **b** *t* distributions

of the methods. It turns out that only one pair has Kendall's tau correlation lower than −0.6, and thus we have a total of 161 pairs.

We visualize the results for the performance comparisons between CEMC$_{tp}$ and the other methods using the scatter plots of their $-\log_{10}(p$ value)'s. In Fig. 5, the left

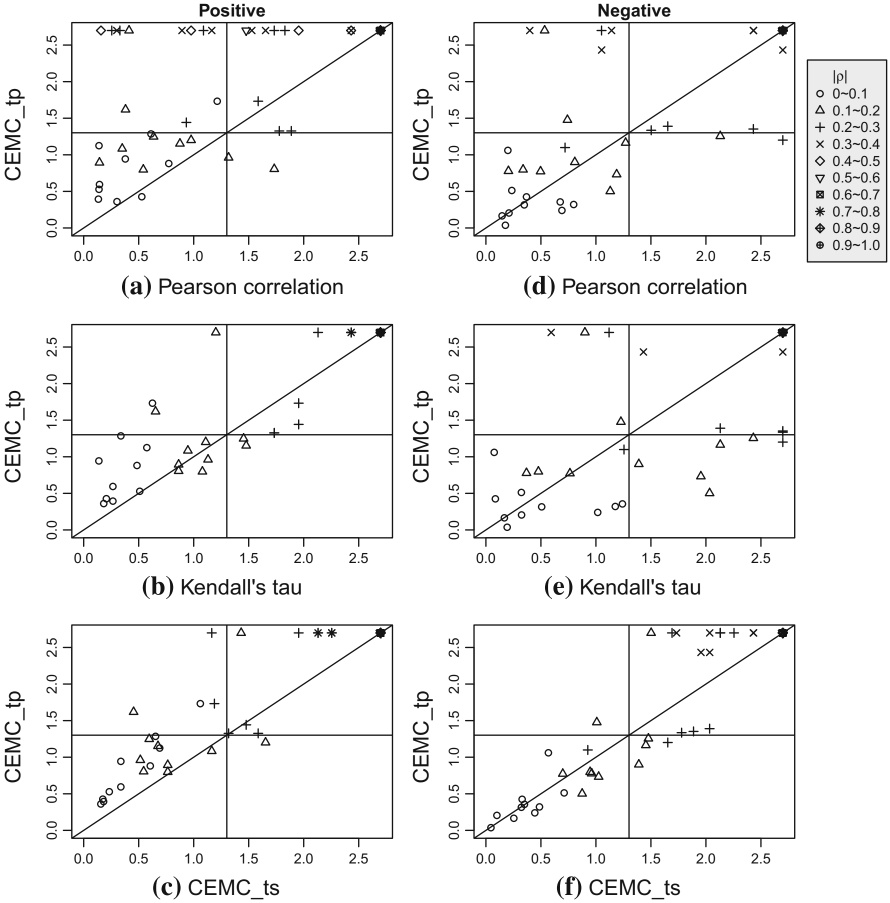

Fig. 5 Scatter plots of $-\log_{10}(p \text{ value})$ for CEMC$_{tp}$ versus the same for each of the three comparison methods: Pearson correlation, Kendall's tau, and CEMC$_{ts}$. *Left* positive association; *Right* negative association

column shows the results for detecting positive associations while the right column is for the negatively associated ones. The horizontal and the vertical lines mark the value $-\log_{10}(0.05)$, thus a point located above the horizontal line or to the right of the vertical line indicates that the pair of genes is identified to be correlated by the corresponding method at the 5 % significance level. As seen from the plots, when Kendall's tau has a very small absolute value $|\rho|$ (say from −0.1 to 0.1), all methods generally regard them as noise. On the other hand, when Kendall's tau correlation has a moderately large absolute value (say greater than 0.5), all the methods identify them as correlated genes. Therefore, in the two extremes, all methods yield similar results. The more interesting cases are when the absolute value of the overall Kendall's tau correlation is moderately small. In those situations, CEMC$_{tp}$ generally has a higher chance of identifying them to be correlated. This is expected, as CEMC$_{tp}$ takes the underlying substructure of the sample, i.e., subset correlations, into consideration, leading to increased power when there are subsets with strong correlations. In contrast, such signals may be washed out

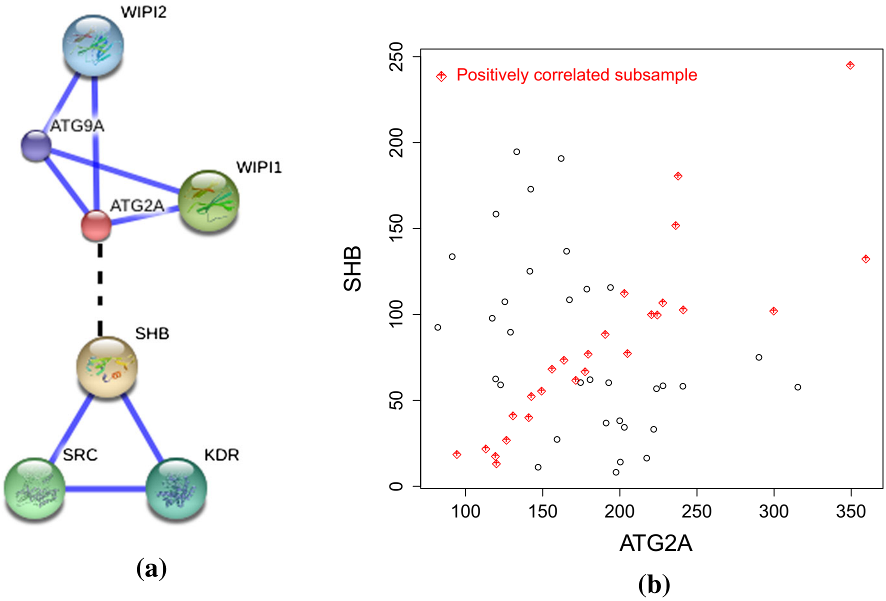

Fig. 6 A network featuring a pair of genes: ATG2A and SHB. **a** Two sub-networks, each for ATG2A and SHB, are shown as those connected by *solid lines*. The significantly positive association between ATG2A and SHB detected by CEMC$_{tp}$ is depicted by the *dashed line*. **b** The detected subsample showing positive correlation is highlighted in the scatterplot of all data points

when a measure is forced to consider the whole set, such as the Pearson's correlation or the standard Kendall's tau, as illustrated in two simulated datasets in Fig. 4.

We illustrate the added values of the results from CEMC$_{tp}$ for understanding gene networks using STRING [16]. The STRING Database (http://string-db.org/) is comprised of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources: genomic context, high-throughput experiments, (conserved) co-expression, and previous knowledge. As it integrates interaction data from these sources for a large number of organisms, connection between some parts of genes that are present only on a subset of observations might have been missed.

We illustrate the results using three examples, pairs ATG2A & SHB, APBA1 & CRK, and PPP6C & TOP3B. Each of the two genes in the first pair belongs to a different network (connected by solid lines in Fig. 6a). Applications of the four methods to this pair of genes yielded a positive association at the 5 % significance level by CEMC$_{tp}$ (p value $= 0.019$), but not by any of the other methods. The result from CEMC$_{tp}$ indicates that the expression levels between these two genes are correlated only in 25 of the 60 cell lines. By taking the possibility of subset correlations into consideration, CEMC$_{tp}$ appears to be able to uncover the relationship (connected by the dotted line in Fig. 6a) and identify the cell lines that contribute to the detected correlation. The connection of these two subunits into a single network is made possible by considering subset relationships, which appears to have been lost when the full set is considered. To see this more clearly, we plotted the expression levels of ATG2A (x-axis) versus that of SHB (y-axis) for each of the 60 cell lines (Fig. 6b). It is easily seen that there is

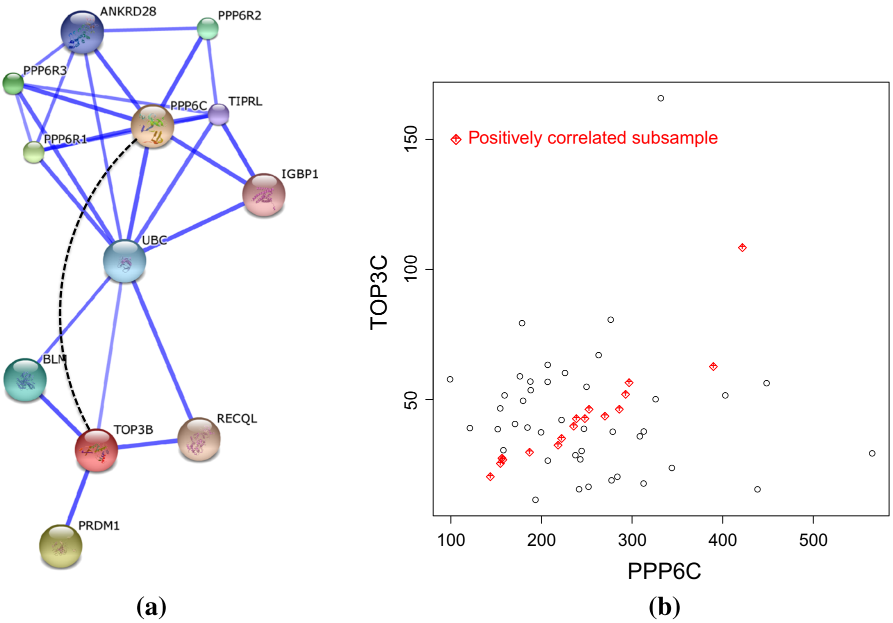

Fig. 7 A network featuring a pair of genes: PPP6C and TOP3B. **a** The significant association between PPP6C and TOP3B detected by CEMC$_{tp}$ is depicted by the *dashed arc*. **b** The detected subsample showing positive correlation is highlighted in the scatterplot of all data points.

little linear or nonlinear correlation between the expression levels of these two genes. Nevertheless, CEMC$_{tp}$ was able to detect a subset of positively correlated cell lines, showcasing its ability for mining hidden evidence.

Similarly, for gene pair APBA1 and CRK, CEMC$_{tp}$ is the only test that leads to the detection of a significant association at the 5 % level (p value = 0.033). Further, the subset correlation appears to be negative, with 17 of the cell lines providing the supporting evidence. Visualization of the network relationship is provided in Supplementary Figure S5(a), together with the scatterplot of the expression data and the subset identified to be negatively correlated as given in Supplementary Figure S5(b). For the pair PPP6C and TOP3B, although it is not strictly significant at the 5 % level (p value = 0.0518) for CEMC$_{tp}$, the evidence of significance is nonetheless much greater than the rest of the methods (all p values $>$ 0.20). Figure 7a (solid lines) shows that both of these two genes are in the same network, i.e., both of them have protein interactions with gene UBC. This added association (dashed arc in Fig. 7a now establishes their direct relationship with one another. The expression data over the 60 cell lines and the correlated subset are depicted in Fig. 7b.

5 Discussion

In real-world problems of studying relationships between two variables, such as in the study of gene networks, the scenario can be much more complicated than a monotone relationship prevailing in the entire sample of observed data, as the underlying population from which the data are drawn can be heterogeneous. However, traditional measures of correlation typically assume that the underlying population is homogeneous (even if such an assumption is rarely stated explicitly), and as such, the evidence of correlation is assessed using all observations with equal weights. A recently proposed measure, tau-path, was able to detect association that exists only in a subset of the sample, but the more complex scenario depicting the existence of two subsets with opposite directions of association (i.e., both positive and negative) was not addressed. Further, the algorithm proposed for detecting the optimal tau-path was greedy in nature, and as such, optimality in any sense is not guaranteed. To fill this void, in this paper, we extend the tau-path methodology to accommodate the more complex scenario, and we propose the adaption of a cross entropy Monte Carlo algorithm to obtain an "optimal" tau-path. The optimality is in the sense that the tau-path will achieve the maximum of the objective criterion when the number of iterations of the CEMC algorithm goes to infinity [17]. The objective criterion, the tau-score, is the weighted average of the pairwise concordance/discordance values, with the weight for each value being "proportional" to its information for providing supporting evidence of association in concert with the other observations.

In evaluating the convergence of the CEMC algorithm, we require that the average of the absolute differences in the entries of the probability matrix ($\boldsymbol{v}$) be smaller than 0.001. Based on 600 randomly selected datasets, this criterion was reached between 134–162 iterations generated from the mixtures of normal distributions, and just a bit longer for samples from the mixtures of t distributions, at 142–167 iterations. At convergence, the $\boldsymbol{v}$ matrix typically has entries that are close to 0 or 1 (recall that the probabilities in each column sum to 1) (Figure S6). A column with one 1 (or very close to 1) and the rest being 0 (or very close to 0) indicates that the particular placement of the observation at that position is definitive, as we see in the first few columns of Figure S6. This may help with visualizing the informativeness of the data and detecting where information on ordering starts to degrade.

Our extensive simulation study substantiates our expectation of the aptness of CEMC$_{tp}$ for detecting associations under various scenarios. When there is a subset with strong correlation, when there are subsets with opposite correlation directions, or when the distributions of the two variables have heavier tails, CEMC$_{tp}$ achieves higher power than Kendall's tau or Pearson's correlation coefficient, which only measure overall correlation. More specifically, when there is a monotonic relationship in the entire sample, there is little loss of power for CEMC$_{tp}$ (Figs. 1 and 2). In fact, CEMC$_{tp}$ can even have higher power when the distributions of the variables have heavier tails. On the other hand, when there are subset correlations of opposite directions, CEMC$_{tp}$ is seen to always have higher power (Fig. 2). The increase in power does not come at the expense of increased type I errors. More specifically, in our CEMC$_{tp}$ and CEMC$_{ts}$ algorithms, we control for path-wise type-I errors, leading to results with actual type-I error rates similar to the nominal values, as we demonstrated in our simulation study. In addition, CEMC$_{tp}$ has the capability of inferring the observations that support a detected correlation, although our simulation study indicates that the size of the subsample can be inflated as it may include observations from an uncorrelated subsample. We illustrate the utility of CEMC$_{tp}$ in a real data analysis by showing its ability to uncover potential relationships that appear to only exist in a subset. Examples show that filling in the "missing links" may provide a more comprehensive view of the gene networks. However, the results need experimental validations.

Despite the general advantage of CEMC$_{tp}$, it is much more computationally intensive than tests based on Kendall's tau or Pearson's correlation coefficient. Its variant, CEMC$_{ts}$, is seen to be more computationally efficient. For example, for a setting where the samples were generated from a mixture of three normal distributions, it took 28.5 and 7.9 s to complete the analysis and obtain the p value with 500 permuted samples for CEMC$_{tp}$ and CEMC$_{ts}$, respectively. When the data were generated from a mixture of t distributions, the times increase to, respectively, 52.6 and 32.0 s. Recall that CEMC$_{ts}$ differs from CEMC$_{tp}$ only in steps 3 and 4, which took a fraction of a second to complete for CEMC$_{ts}$ while it took over 20 s for CEMC$_{tp}$, essentially accounting for the differences in their computational times. For the computation of Kendall's tau and Pearson's correlation coefficients, each took less than 1 s. These computations were performed on a supercomputer cluster with 540 nodes, 12 cores/node, 48 GB of memory/node, and Intel Xeon x5650 CPUs. However, although CEMC$_{ts}$ has similar power as CEMC$_{tp}$ in some situations, it does not provide information on the subset of observations underlying a detected association. As such, it is warranted to search for a more computationally efficient procedure that has similar properties as CEMC$_{tp}$ (e.g., asymptotic optimality, identification of subsamples). Such a computational improvement is crucial for analyzing large samples upward of thousands of observations, such as those contained in The Cancer Genome Atlas (TCGA) [18] and in the GTEx database [4]. A potential strategy is to borrow the top-K CEMC idea and software [19] for finding a subsample with the strongest signal for detecting association in each direction. For a large sample, there can be substantial noise, and therefore the top-K strategy, with a reasonable choice of K to facilitate computation, may prove to be efficient without much loss of information. We have explored this idea in a preliminary analysis of a breast invasive carcinoma dataset with over 1000 individuals from TCGA [20]. Although the results are promising, much more work is needed to formalize the top-K CEMC procedure and to assess its performance.

Acknowledgements The authors would like to thank the two anonymous reviewers for their constructive comments and suggestions. This work was supported in part by the National Science Foundation grants DMS-1220772. Th authors would also like to acknowledge the allocation of computing times from the Ohio Supercomputer Center.

Appendices

Appendix A: CEMC Algorithm

Different from an exhaustive search that places a discrete uniform distribution on all the possible candidate orders, the idea of finding z^* using CEMC is to place more and more of its probability mass on the z's in a neighborhood of the z^*. This is accomplished by iteratively updating the parameter matrix $\boldsymbol{v}$. Note that a z being in a neighborhood of z^* means that the corresponding value of the objective function, the tau-score $T = T(f(z))$ in this case, is close to the maximum $T^* = T(f(z^*))$. Suppose $\boldsymbol{v}$ is the current estimate of the parameter matrix. To find the next $\boldsymbol{v}'$ so that its tau-score is getting even closer to the optimal T^*, we cast the problem as finding a good importance sampling distribution, $P_{\boldsymbol{v}'}(z)$, for estimating probability $A = P_{\boldsymbol{v}}[T(f(z) \geq a]$, which can be rare if constant a is set to be close to T^*. The choice of a will be discussed in more detail in the algorithm. The ideal importance sampling distribution is

$$Q^*(z) = \frac{I[T(f(z)) \geq a]P_{\boldsymbol{v}}(z)}{A},$$

but this is not obtainable since it involves the unknown probability A. However, we can obtain a good sampling distribution $P_{\boldsymbol{v}'}(z)$ by minimizing the cross entropy (i.e., the Kullback–Leibler distance) between it and the ideal but unobtainable distribution $Q^*(z)$, CE$(P_{\boldsymbol{v}'}(z), Q^*(z))$. This minimization can be achieved since it is equivalent to maximizing

$$E_{\boldsymbol{v}}\{I[T(f(\boldsymbol{x})) \geq a] log P_{\boldsymbol{v}'}(z)\}, \tag{3}$$

which is now free of the unknown probability A.

We find $\boldsymbol{v}_{new}$ that maximizes the expectation in (3) by a Monte Carlo approximation. Suppose $z_i, i = 1, \ldots, N$, is a sample drawn from $P_{\boldsymbol{v}}(z)$ with the current parameter specification $\boldsymbol{v}$, with their corresponding permutations denoted as $p_i = f(z_i), i = 1, \ldots, N$. Then the formula to get the update for the next parameter matrix $\boldsymbol{v}'$ is

$$\boldsymbol{v}_{new} = \arg\max_{\boldsymbol{v}'} \left\{ \frac{1}{N} \sum_{i=1}^{N} I[T(f(z_i) \geq a] log P_{\boldsymbol{v}'}(z_i) \right\}. \tag{4}$$

It has been shown that when the threshold value a is also updated iteratively, it will converge to a value (a_∞) that is close to T^* [17]. At the same time, $P_{\boldsymbol{v}^0}(z)$, $P_{\boldsymbol{v}^1}(z), \ldots,$ will converge to a distribution that places most of its probability mass on the z's that satisfy $T\big(f(z)\big) \geq a_\infty$. As suggested [14], in practice, the weighted averages of $\boldsymbol{v}$ and $\boldsymbol{v}_{new}$ as $\boldsymbol{v}'$ can better balance the rate of convergence and the chance of not being trapped in a local minimum, which is adopted in the following Order Explicit Algorithm (OEA).

The OEA Algorithm

1. Set $\boldsymbol{v}^0$ with each $v^0_{jr} \in (0, 1)$ such that $\sum_j p^0_{jr} = 1, r = 1, \ldots, n$. For instance, $p^0_{jr} = 1/n$ for all the j and r indicates so that each observation in the sample may be arranged into each position equally likely. Other choices that make use of prior information, if such is available, may also be constructed. We then draw N realizations from this initial distribution. Set $t = 0$.
2. Keep the $N_1(<N)$ realizations with the largest values of the tau-scores from the previous iteration and draw z_i, $i = N_1 + 1, \ldots, N$ from $P_{\boldsymbol{v}^t}(z)$ to form a new sample of N realizations. From this combined sample, we find the corresponding permutations and their tau-scores, p_i and $T(p_i)$, $i = 1, \ldots, N$. Let a^t be the sample upper q-quantile of the tau-scores.

3. Using the sample we update the parameter vector $\boldsymbol{v}^{t+1}$ as follows: $\boldsymbol{v}^{t+1} = (1 - \gamma)\boldsymbol{v}^t + \gamma \boldsymbol{v}_{new}$, where v_{new} is as defined in (4) and $0 < \gamma \leq 1$ is a weight parameter that is typically set to be 1/2.
4. If $\|\boldsymbol{v}^{t+1} - \boldsymbol{v}^t\| < \epsilon$, then we output the largest tau-score value and the corresponding tau-path. Otherwise we set $t = t + 1$ and go back to step 2.

Appendix B: Detection of Overall Subset Association

The CEMC$_{tp}$ algorithm presented in Sect. 2.4 is for simultaneous, yet separate, detection of positive and negative associations, and importantly, the identifications of subsamples that lead to such detections. However, for comparison with traditional association measures, we also propose the following overall association detection algorithm to facilitate such a task. The first two steps are the same as in the CEMC$_{tp}$ algorithm. However, Step 3 for calculating the p value (there will be only one overall p value) is different, which is given below. Step 4 is also different as we do not need to split the nominal significance level between the positive and negative associations.

3. Find the smallest upper quantile q_+ for τ_+ along the path. We also find the smallest upper quantile $q^{(+l)}$ for each of the tau-paths, $\tau^{(+l)}, l = 1, \ldots, m$, from the permuted data. We find q_- and $q^{(-l)}, l = 1, \ldots, m$, similarly. The overall p value for assessing subset association is the lower quantile of $\min\{q_+, q_-\}$ among the set of values $\{\min\{q_+, q_-\}, \min\{q^{(+l)}, q^{(-l)}\}, l = 1, \ldots, m\}$.
4. If the overall p value is less than α, the predetermined significance level, then the sample is said to exhibit significant evidence of association in either the full, or a subsample.

As seen from the above steps, the assessment of overall evidence of association basically considers both positive and negative associations and synthesizes the information to provide the stronger evidence.

The revised Steps 3 and 4 can also be similarly worked out for CEMC$_{ts}$ for an overall assessment of subset association.

References

1. Katz G (2014) How much do we know about HDL cholesterol? Clin Correl http://www.clinicalcorrelations.org/?p=7298
2. Voight BF et al (2012) Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet 380:572–580
3. Wang YX, Waterman MS, Huang H (2014) Gene coexpression measures in large heterogeneous samples using count statistics. Proc Natl Acad Sci USA 111:16371–16376
4. Lonsdale J, Thomas J, Salvatore M, Phillips R et al (2013) The genotype-tissue expression (GTEx) project. Nat Genet 45:580–585
5. Pearson K (1895) Notes on regression and inheritance in the case of two parents. Proc R Soc Lond 58:240242
6. Stigler SM (1989) Francis Galton's account of the invention of correlation. Stat Sci 4:7379
7. Diaconis P, Graham RL (1977) Spearman's footrule as a measure of disarray. J R Stat Soc Ser B 39:262268
8. Kendall M (1938) A new measure of rank correlation. Biometrica 30:81–89
9. Kendall M (1970) Rank correlation methods, 4th edn. Griffin, London

10. Yu L (2009) Tau-path test a nonparametric test for testing unspecified subpopulation monotone association. Ph.D. thesis. The Ohio State University, 2009
11. Yu L, Verducci JS, Blower PE (2011) The tau-path test for monotone association in an unspecified population: application to chemogenomic data mining. Stat Methodol 8:97–111
12. Rubinstein RY, Kroese DP (2004) The cross-entropy method: a unified approach to combinatorial optimization, Monte Carlo simulation, and machine learning. Springer, New York
13. Liu Z, Lin S, Tan M (2006) Genome-wide tagging SNPs with entropy-based Monte Carlo methods. J Comput Biol 13:1606–1614
14. Lin S, Ding J (2009) Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA studies. Biometrics 65:9–18
15. Shoemaker RH (2006) The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer 6:813–823
16. Szklarczyk D, Franceschini A, Wyder S et al (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015:D447–452
17. Margolin L (2005) On the convergence of the cross-entropy method. Ann Oper Res 134:201–214
18. McLendon R, Friedman A, Bigner D (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455:1061–1068
19. Schimek MG, Budinska E, Ding J, Kugler KG, Svendova V, Lin S (2015) TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. Stat Appl Genet Mol Biol 14:311–316
20. Cancer Genome Atlas Netwok (2012) Comprehensive molecular portraits of human breast tumors. Nature 490:61–70