The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1180–1203
https://doi.org/10.1214/17-AOAS1083
© Institute of Mathematical Statistics, 2018

A TESTING BASED APPROACH TO THE DISCOVERY OF DIFFERENTIALLY CORRELATED VARIABLE SETS

BY KELLY BODWIN[1], KAI ZHANG[2] AND ANDREW NOBEL[3]

University of North Carolina at Chapel Hill

Given data obtained under two sampling conditions, it is often of interest to identify variables that behave differently in one condition than in the other. We introduce a method for differential analysis of second-order behavior called Differential Correlation Mining (DCM). The DCM method identifies differentially correlated sets of variables, with the property that the average pairwise correlation between variables in a set is higher under one sample condition than the other. DCM is based on an iterative search procedure that adaptively updates the size and elements of a candidate variable set. Updates are performed via hypothesis testing of individual variables, based on the asymptotic distribution of their average differential correlation. We investigate the performance of DCM by applying it to simulated data as well as to recent experimental datasets in genomics and brain imaging.

1. Introduction.

In many statistical problems, one has two datasets that measure the same variables under different conditions. It is common in the analysis of such data to assume that the samples in each dataset are generated from two underlying distributions. Even when the data is high dimensional, differences between the distributions may be present for only a small number of variables, and it is often of interest to identify these key variables. In this paper, we present a new method of second-order comparative analysis, called Differential Correlation Mining (DCM), that identifies sets of variables such that the average pairwise correlation between variables in the set is higher in one sample condition than in another. The method does not make use of auxiliary information, apart from the separation of samples into predetermined groups (e.g., treatment vs. control). DCM is theoretically applicable to both low and high-dimensional settings and is computationally feasible for high-dimensional data (10^5 variables).

Most often, differential behavior between sample groups is measured by *first-order* statistics, which are functions of a single variable. Familiar first-order statistics include the sample mean and the sample variance. A well-studied example of

Received February 2016; revised March 2017.

[1] Supported in part by NSF Graduate Research Fellowship Grant DGE-1144081 and NSF Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute.

[2] Supported in part by NSF Grants DMS-1309619, DMS-1613112 and IIS-1633212.

[3] Supported in part by NSF Grants DMS-1613072 and DMS-1310002 and NIH Grants R01 HG009125-01 and R01 MH101819-01.

Key words and phrases. Differential correlation mining, association mining, biostatistics, genomics, high-dimensional data.

first-order differential analysis is the study of differential gene expression in microarrays [see Cui and Churchill (2003) for a canonical example, or Soneson and Delorenzi (2013) and the references therein for an overview of several methods]. Other applications of first-order differential analysis include text analysis for authorship identification [Stamatatos (2009)], studies of brain functionality based on regional activation [Phan et al. (2002)], and investigation of cultural bias in standardized testing [Wainer and Braun (2013)].

The use of first-order statistics allows for analysis of only a single variable at a time. To study relationships between pairs of variables, one requires functions of two variables, which specify *second-order* statistics. Examples of second-order statistics include correlation, covariance, and distance. When one wishes to understand interactions between many variables (as in clustering problems), data may be summarized in matrix form, where each entry in the matrix represents the observed value of a second-order statistic. It is common to look within a matrix of relational data for groups of variables that have high pairwise association. In applications of nondifferential second-order analysis, variable groups may represent, for example, social groups communication networks [Lewis et al. (2008)], genes in common protein pathways [Jiang, Tang and Zhang (2004)], or functionally similar brain regions [Greicius et al. (2002)].

While there is a large literature on clustering and networks, to the best of our knowledge, there is relatively little work comparing second-order behavior across two sample conditions. The many insights obtained from ordinary second-order variable set selection lead us to believe that a second-order differential approach will be of scientific interest. The methods introduced in this paper fall under the broader heading of *differential association mining*. As in ordinary association mining, we are interested in the pairwise behavior of variables; however, in the differential setting, we must consider two different relational matrices. In some cases, simply taking the difference of the matrices and applying ordinary clustering methods would suffice. However, most second-order statistics—including the focus of this paper, the linear correlation coefficient—require a more careful treatment. For instance, two sample correlation matrices will exhibit vastly different random behavior based on the sample sizes of the corresponding datasets, and will have a complex dependency structure when the corresponding population correlation matrices are not the identity.

The DCM method proposed here addresses differential correlation mining in a direct way. (Section 1.2 considers possible alternatives based on existing work.) DCM seeks variable sets that form differentially correlated (DC) cliques. In a graph, a clique is a set of nodes that is fully connected, in the sense that there is an edge between every pair of nodes in the set. Informally, a DC clique is a set of variables such that each variable in the set has a positive (usually large) average differential correlation with the other variables in the set.

More formally, let $\mathbf{R}_1$, $\mathbf{R}_2$ be the $p \times p$ population correlation matrices of the distributions underlying sampling Conditions 1 and 2, respectively. Let $A \subset [p]$, where $[p]$ is the index set $\{1, \ldots, p\}$, and define

$$\Delta(i, A) = \frac{1}{|A|} \sum_{j \in A} (\mathbf{R}_1 - \mathbf{R}_2)_{ij} \tag{1.1}$$

to be the average difference of correlations between variable i and variables in index set A. Here, the subscript ij denotes the element in the ith row and jth column of the corresponding matrix, and $|A|$ is the cardinality of the set A. We formally define DC cliques as follows.

DEFINITION 1.1. Let $\mathbf{R}_1$, $\mathbf{R}_2$ be given and let $\Delta(\cdot, \cdot)$ be defined as in (1.1). An index set $A \subseteq [p]$ with at least two elements is a *DC clique* for $\mathbf{R}_1 - \mathbf{R}_2$ if:

1. $\Delta(i, A) > 0$ if and only if $i \in A$.
2. The set A cannot be written as a disjoint union of nonempty index sets $A_1, A_2 \subset [p]$ such that A_1 and A_2 satisfy Condition 1 above.

Condition 1 ensures that no relevant variables are omitted from a DC clique (every variable that is positively differentially correlated relative to the set A is included in A) and that a DC clique does not contain any extraneous elements. Condition 1 implies that a DC clique has larger average pairwise correlation under the first distribution than under the second. Condition 2 ensures that a DC clique cannot be subdivided into two smaller DC cliques. Importantly, the definition places *no conditions* on the correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$. In particular, $\mathbf{R}_1$ and $\mathbf{R}_2$ need not be sparse, and need not satisfy any structural constraints such as bandedness. For a given pair $\mathbf{R}_1$, $\mathbf{R}_2$, it may happen that no DC cliques exist, or that the entire variable set forms a DC clique.

Note that the definition of DC cliques is not symmetric: in general, the DC cliques for $\mathbf{R}_1 - \mathbf{R}_2$ will be different from those for $\mathbf{R}_2 - \mathbf{R}_1$. The difference lies not in the relational structure itself, but rather in how we order the sample conditions (1 or 2). For example, in biological data, one sample group may involve a treatment condition, while the other is a reference or control group. A DC clique for $R_1 - R_2$ would contain genes that are more highly correlated in Condition 1 than Condition 2, for example, a protein pathway that is more active in Condition 1. This structure is illustrated in Figure 1.

The asymmetry in DC cliques could be eliminated by replacing the relevant section of (1.1) by a symmetric notion of difference such as $|\mathbf{R}_1 - \mathbf{R}_2|$. However, a variable set based on absolute difference (or similar) could contain a mixture of elements with positive correlation to A *and* elements with negative correlation to A. Such mixed groups would not exhibit the unified block structure of the type seen in Figure 1. Further, large variable sets with strong average negative correlation cannot occur. Simple algebra shows that since $\mathbf{R}_1$ is positive definite, the average pairwise correlation in Condition 1 of any set A with k elements must be at least $-\frac{1}{(k-1)}$.

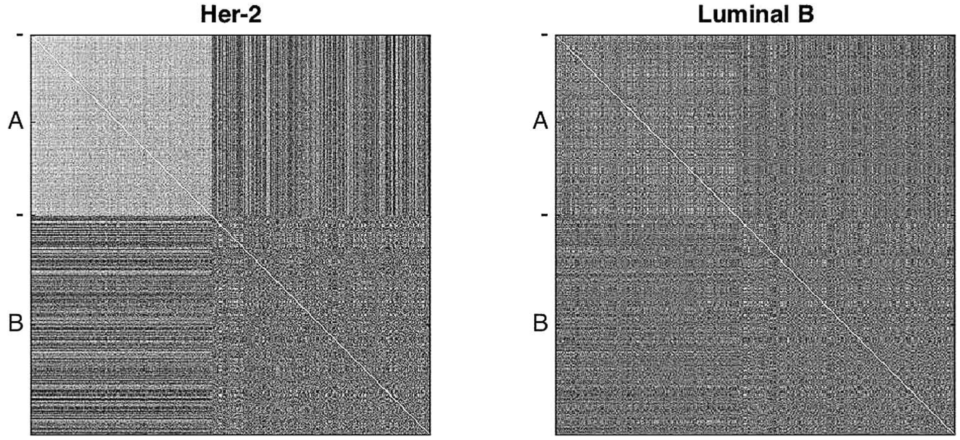

FIG. 1. *Sample correlation matrices for each of two breast cancer tumor subtypes, showing observed DC clique* (A) *and random genes* (B).

As defined above, DC cliques are features of the underlying population distributions of the data. In practice, we will replace $\mathbf{R}_1$, $\mathbf{R}_2$ with estimates from observations, accounting for the uncertainty in these estimators, to select empirical DC cliques. The broad objective of DCM is to use observed data to identify DC cliques, or approximations of these, without prior knowledge of the identity, number, or size of the DC cliques present in the population. It is worth noting that the DCM algorithm and supporting analysis described here are easily adapted to a nondifferential correlation mining algorithm. An implementation of a correlation mining procedure is included along with the public DCM software.

REMARK. Some bioinformatics literature uses the phrase "Differential Co-Expression," sometimes abbreviated "DC," as an umbrella term for all differential second-order gene expression behavior. In this paper, "DC" will refer specifically to differential correlation; when a distinction must be made with co-expression or covariance, this will be made explicit.

1.1. *An example.* To motivate our definition of DC cliques, we provide an illustrative real-world example. Figure 1 shows an empirical DC clique identified by DCM in real data from The Cancer Genome Atlas (TCGA) Research Network (http://cancergenome.nih.gov/). The two sample conditions under consideration are Her-2 type breast cancer tumors and Luminal B type tumors, as classified by Perou et al. (2000). (Further results for the TCGA dataset are provided in Section 5.)

Figure 1 shows the sample correlation matrices within each tumor type, restricted to a set of 202 variables consisting of an empirical DC clique of size 102 selected by DCM (A), and 100 randomly chosen variables (B). The variables B

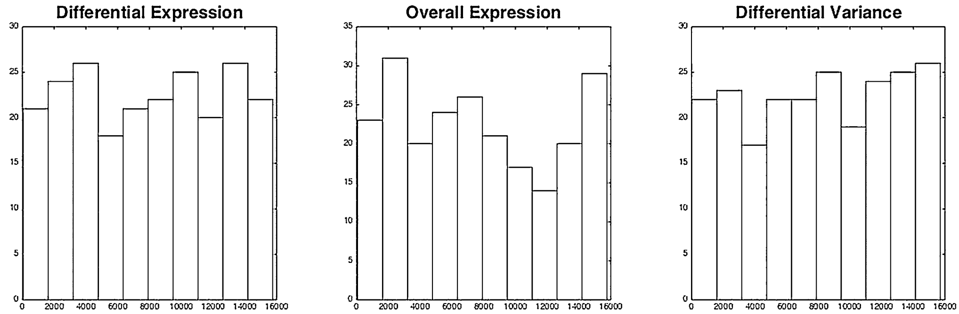

FIG. 2. *Ranks of genes in observed DC clique* (A) *out of* 15,785 *total genes*. (*Ranked by*: *differential expression*, *as measured by* p-*values of* 2-*sample* t-*tests*; *mean overall expression among Her*-2 *samples*; *and ratio of sample variances between Her*-2 *and Luminal B*.)

are included for contrast, and to show that the differential correlation observed in A is not present in the entire dataset. The figure illustrates the second-order behavior and the differential nature of the identified DC clique A. The block pattern in the upper left corner of the Her-2 matrix shows that every entry in the correlation matrix of A is large, suggesting that all the variables of A are strongly pairwise correlated. The Luminal B sample correlation shows a similar pattern, but it is much less pronounced. No such pattern is seen among the variables in B.

In general, the results of DCM are distinct from those found by first-order analysis (e.g., differential expression). For example, Figure 2 shows the relative differential expression, overall expression level, and differential variation for the above estimated DC clique A. For this plot, we ranked all genes in the study ($p = 15{,}785$) by (a) t-statistic of differential mean expression between Her-2 and Luminal B samples, (b) overall expression in Her-2 samples, and (c) ratio of sample variations (F-statistic) for Her-2 versus Luminal B samples. The histograms in Figure 2 show the ranking of the genes in A. The overall uniformity of the histograms indicates that the variables in the observed DC clique A do *not* exhibit standard first-order differential behavior. Similar results were observed for all other data studied in this paper.

By targeting DC cliques, the DCM method identifies variables whose *joint* behavior is different across sample conditions. The results are readily interpretable as sets of variables that interact strongly under one sample condition but only weakly (or not at all) under another. In this paper, we will demonstrate DCM is an effective and efficient way to identify differentially correlated variable sets from observed data.

1.2. *Related work.* Below we provide an overview of work that is either directly related to DCM or may be reasonably adapted to the DC clique paradigm. In what follows, let $\mathbf{R}_1$, $\mathbf{R}_2$ denote the population correlation matrices of two data distributions, and let $\widehat{\mathbf{R}}_1$, $\widehat{\mathbf{R}}_2$ denote the corresponding sample correlation matrices.

Mining from single correlation matrices. Nondifferential correlation mining, in which one searches for highly associated variables from a single dataset, has been well studied, typically in the context of clustering. Kriegel, Kröger and Zimek (2009) provides a survey of clustering methods for high-dimensional data based on correlation distance. Datta and Datta (2002) and Jiang, Tang and Zhang (2004) and the references therein give an overview of methods developed specifically for clustering of gene expression. In general, typical clustering or community detection methods must be adapted for application to correlation distances to correct for bias [see, e.g., MacMahon and Garlaschelli (2015) for an illustrative example].

Detection of isolated changes in correlation structure. Existing approaches to differential correlation mining are based largely on examining individual variables for changes in second-order structure across two sample conditions. For example, one may treat $\widehat{\mathbf{R}}_1$ and $\widehat{\mathbf{R}}_2$ as the adjacency matrices of two fully connected, weighted networks, and then look for variables whose connectivity pattern is very different across the two networks [Gill, Datta and Datta (2010), Xia, Cai and Cai (2015)]. Most methods approach differential correlation mining by developing a statistic to measure the change in pairwise correlations of an individual variable: Hu, Qiu and Glazko (2010) uses the covariance distance (total difference of covariances), Choi and Kendziorski (2009) use a direct difference of sample correlations, Fukushima (2013) uses the difference of Fisher transformed sample correlations, and Liu et al. (2010) use a filtration (or thresholding) step before summing square correlation differences. These methods then permute samples across the two classes to measure the significance of the original differential correlation. Significant variables may then be selected by an appropriate multiple testing procedure.

Estimation and hypothesis testing. Much theoretical work is devoted to testing equality of high-dimensional covariance and correlation matrices. When the sample size n is substantially larger than the dimension p, classical results are applicable, for example, likelihood ratio tests as discussed in Anderson (1959) and Muirhead (1982), or results like those of Steiger (1980) for testing individual sample correlation. In the high-dimensional ($p > n$) setting, Cai and Jiang (2011), Cai, Zhang and Zhou (2010), Cai, Liu and Xia (2013) have developed minimax rate optimal tests for the equality of covariance matrices under sparsity assumptions. Results for correlation (rather than covariance) are less prevalent; recent work includes tests for sets of sample correlation coefficients [Donner and Zou (2014)], tests for rank-based correlation matrices [Zhou et al. (2015)], and tests for detecting overall dependence [Bassi and Hero (2012)].

In some cases, optimal testing procedures can inform methods for estimation of high-dimensional covariance and correlation matrices. Particularly relevant is the work of Cai and Zhang (2014), which yields an estimator for the difference matrix $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$. This estimator is implemented and discussed further in Section 4. Other approaches to high-dimensional estimation include: Bickel and Levina (2008), who discuss a thresholding estimator for covariance matrices; Peng, Zhou and Zhu (2009), who estimate partial correlations in sparse regression models; and Rajaratnam, Massam and Carvalho (2008), who make use of graphical model techniques for covariance matrix estimation.

Finally, the work of Sheng, Witten and Zhou (2016) proposes an approach to differential correlation mining by testing subsections of the difference of correlation matrices $\mathbf{R}_1 - \mathbf{R}_2$. Like DCM, the proposed method seeks to identify DC clique-like structure by appealing to classical asymptotic results. However, the method relies on a sequential testing and screening procedure that is infeasible for high-dimensional settings ($\sim 10^2$ or more). As such, despite the close relationship between this method and DCM, we were not able to include it in the simulation study in Section 4.

1.3. *Outline.* In the next section, we describe in detail the three main steps of the DCM procedure. Section 3 provides a closer examination of the test statistic used in the procedure, including a discussion of its asymptotic distribution. We apply DCM to simulated data in Section 4, and compare the results to possible alternative procedures based upon existing work. Finally, in Section 5 we present the results of two applications of DCM to the aforementioned TCGA dataset and to brain activity data from the multiinstitutional Human Connectome Project.

2. The DCM procedure. In this section, we present details of the three components of the proposed DCM procedure: initialization, set update, and residualization. The initialization step employs a simple greedy algorithm to select an initial variable set A. Once the initial set is determined, it is passed to an update algorithm that iteratively refines the set, making use of a hypothesis testing framework to test variables for differential correlation. When an estimated DC clique is found, the residualization step prepares the data for further search by removing the differential correlation of the discovered set.

An important advantage of this type of approach is that the number and size of output sets are chosen adaptively based on testing principles. The DCM method does not require prespecification of the number of clusters (as in kmeans), nor does it require an additional decision about cluster size (as in hierarchical clustering). Rather, the multiple testing procedure in the iterative step of DCM naturally determines the number of variables in an output set. DCM also differs from typical clustering procedures in that it does not require the calculation of a full $p \times p$ dissimilarity matrix, which can be a computational advantage in high-dimensional data.

The DCM procedure is summarized below. Detailed pseudocode is supplied as supplemental material.

THE DCM PROCEDURE

- ▹ *Initialization:* Identify a good initial variable set A using a greedy algorithm that identifies a local maximum of a simple score function.
- ▹ *Iteration:* Refine the initial set A. At each iterative step, repeat the following until termination.
    - ▹ *Test* the differential correlation of each variable i with respect to A. Let A' be the set of variables with significant differential correlation, as determined by an FDR controlling multiple testing procedure.
    - ▹ *Terminate* if $A' = A$ or a cycle is observed.
    - ▹ *Update:* Set A to be A'.
- ▹ *Return:* Output variable set A.
- ▹ *Residualization:* Remove the effect of the DC clique A.
- ▹ *Repeat* search with new initial set as many times as desired.

Iterative updating using multiple testing was first applied by Wilson et al. (2014) in the context of community detection for binary networks. DCM makes use of the same search paradigm; however, a fundamentally different treatment is required to address differential correlation. In particular, the work of Wilson et al. (2014) performs hypothesis tests based on a fully constructed null model, whereas DCM requires no structural assumptions on the null distribution of the data beyond equal correlation ($\mathbf{R}_1 = \mathbf{R}_2$) and some mild moment conditions (see Theorem 1).

We now provide a more in-depth discussion of each step of the procedure.

2.1. *Notation.* We assume that the data under condition c consists of n_c independent samples drawn from a distribution F_c with correlation matrix $\mathbf{R}_c$, for $c = 1, 2$. Let $\mathbb{X}_1 = (\mathbf{U}_1, \dots, \mathbf{U}_p) \in \mathbb{R}^{n_1 \times p}$ and $\mathbb{X}_2 = (\mathbf{V}_1, \dots, \mathbf{V}_p) \in \mathbb{R}^{n_2 \times p}$ denote the resulting data matrices in standard sample-by-variable form. Thus $\mathbf{U}_j \in \mathbb{R}^{n_1}$ denotes the measurements of variable j under Condition 1, while $\mathbf{V}_j \in \mathbb{R}^{n_2}$ denotes the measurements of variable j under Condition 2. Let $\mathbb{X}_{1,A} = (\mathbf{U}_j)_{j \in A}$ and $\mathbb{X}_{2,A} = (\mathbf{V}_j)_{j \in A}$ denote the restriction of $\mathbb{X}_1$ and $\mathbb{X}_2$, respectively, to a variable set $A \subset [p]$. Similarly, let $\mathbf{R}_{c,A}$ denote the correlation matrices under distribution F_c restricted to the variables in A.

Let $\tilde{\mathbf{U}}_j$ and $\tilde{\mathbf{V}}_j$ be the standardized versions of $\mathbf{U}_j$ and $\mathbf{V}_j$, respectively, such that $\|\tilde{\mathbf{U}}_j\| = \|\tilde{\mathbf{V}}_j\| = 1$, and define $\tilde{\mathbb{X}}_1 = (\tilde{\mathbf{U}}_1, \dots, \tilde{\mathbf{U}}_p)$ and $\tilde{\mathbb{X}}_2 = (\tilde{\mathbf{V}}_1, \dots, \tilde{\mathbf{V}}_p)$. Finally, for $c = 1, 2$, let $\widehat{\mathbf{R}}_c$ denote the usual sample correlation matrices from data of $\mathbb{X}_c$ (and $\widehat{\mathbf{R}}_{c,A}$ that of the appropriate restricted datasets). Thus $(\widehat{\mathbf{R}}_1)_{ij} = \widehat{\mathrm{cor}}(\mathbf{U}_i, \mathbf{U}_j) = (\tilde{\mathbb{X}}_1^t \tilde{\mathbb{X}}_1)_{ij}$ and a similar relation holds for $\widehat{\mathbf{R}}_2$.

2.2. *Initialization.* The set update procedure in the second step of DCM readily identifies variables that are significantly differentially correlated relative to a given variable set A, and is most effective when the initial set of variables exhibits at least low levels of differential correlation. (When applied to a randomly chosen set of variables, the set update procedure typically returns an empty set.) The core search procedure could be run exhaustively, beginning with every variable set $A \subset [p]$, but this is not computationally feasible for data sets of high or moderate dimension. As an alternative, we identify initial variable sets exhibiting a moderate degree of differential expression using a greedy search procedure. We then pass this initial skeleton clique to the set update process to be fleshed out into a final estimated DC clique.

The initialization procedure seeks a local maximum of the score function

$$S(A) = \sum_{i,j \in A} \left\{(n_1 - 3)^{1/2}\varphi(\widehat{\mathbf{R}}_1) - (n_2 - 3)^{1/2}\varphi(\widehat{\mathbf{R}}_2)\right\}_{ij}, \tag{2.1}$$

where φ is the element-wise Fisher transformation of sample correlations, namely

$$\varphi(r) = \frac{1}{2}\log\left(\frac{1-r}{1+r}\right). \tag{2.2}$$

To find a local maximizer of $S(\cdot)$, we begin with a random seed A. We consider only pairwise swaps in which we replace an element of A with one from A^c. The set A is then updated by making the swap that produced the largest increase in the score. Since exactly one element is added and removed at each stage, the size of the variable set remains constant. The cardinality of A is user-specified (with a default of 50). Due to the subsequent set update procedure, we find that many initial cardinalities result in the same final outcome. (As a rule, erring on the side of initial cardinalities that are *smaller* than the expected output set size is advisable, to avoid drowning out signal with too much noise.) Because of the random seeding, the algorithm is not purely deterministic. However, in practice the same local maximum is reached from most seeds.

We make use of the variance-stabilizing Fisher transformation in the initialization procedure as a way of roughly capturing *significance* of differential correlation instead of simply maximizing over absolute differences $\widehat{R}_1 - \widehat{R}_2$. The transformation, and subsequent weighting by degrees of freedom, ensures that the first and second terms in the sum are approximately standardized. As such, sets maximizing $S(\cdot)$ are good ballpark guesses for true DC cliques. In the core set update procedure (Section 2.3), we employ a precise testing approach to measure significance of average differential correlation, so the initial sets need not be perfect. It is simply computationally more efficient to "warm-start" the algorithm with a reasonable set than to apply the core refinement procedure from random starting points.

Pseudocode for the implementation of the initializing algorithm is provided as supplemental material. A closely related method is implemented in Section 4 for comparison with DCM.

2.3. *Core set update procedure.* The heart of the DCM procedure is the set update algorithm, which makes use of multiple testing principles to iteratively refine a variable set A. Recall that the goal of DCM is to estimate DC cliques from the data. To this end, the set update procedure is designed to identify variable sets exhibiting the properties of a true DC clique up to a level of statistical significance.

Consider a single iterative step, at which we update a given variable set A. We wish to determine whether each variable i (including those in A itself) ought to be included in the updated set A'. Since our eventual goal is to discover a DC clique, we perform hypothesis tests based upon the principles of Definition 1.1. For a given variable set A, the tests for variable i may be written as

$$H_0(i, A) : \Delta(i, A) = 0 \quad \text{vs.} \quad H_1(i, A) : \Delta(i, A) > 0. \tag{2.3}$$

Recall that $\Delta(i, A)$, as defined in (1.1), is a difference of average pairwise correlations between i and elements of A, so (2.3) is a test of differential correlation *relative* to the fixed set A. We then update the set A to $A' = \{i : H_0(i, A) \text{ was rejected}\}$ by simultaneous multiple hypothesis testing. This process continues until a fixed point $A = A'$ is reached.

To test the hypotheses in (2.3), we require a test statistic. A natural choice is the corresponding sample quantity

$$\hat{\Delta}(i, A) = \frac{1}{|A|} \sum_{j \in A} (\widehat{\mathbf{R}}_1 - \widehat{\mathbf{R}}_2)_{ij}. \tag{2.4}$$

In addition to being a straightforward choice, this test statistic exhibits several desirable properties discussed in Section 3.

Let $\delta(i, A)$ denote the realized value of the test statistic $\hat{\Delta}(i, A)$ for a particular dataset. It is clear that large positive values of $\delta(i, A)$ provide support for the alternate hypothesis in (2.3), while values that are negative or close to zero provide evidence in favor of the null. Thus, to test the hypotheses, for each $i = 1, \dots, p$ we calculate a p-value of the form

$$\mathrm{p}(i : A) = \mathbb{P}_0\big(\hat{\Delta}(i, A) > \delta(i, A)\big), \tag{2.5}$$

where the probability $\mathbb{P}_0$ is the (unknown) distribution of $\hat{\Delta}(i, A)$ under the null hypothesis $\Delta(i, A) = 0$. Since we make no assumptions about the distributions of data under Conditions 1 and 2, we make use of asymptotic results to approximate the above probability. We show in Section 3.2 that, under appropriate regularity assumptions, and for large enough sample sizes n_1 and n_2,

$$\mathrm{p}(i : A) \approx 1 - \Phi\left(\frac{\delta(i, A)}{\hat{\sigma}_0}\right), \tag{2.6}$$

where $\hat{\sigma}_0^2$ is an estimate of the variance of $\hat{\Delta}(i, A)$ that can be computed from the available data. (The exact form of $\hat{\sigma}_0^2$ is given in Appendix 2.)

The p-values $\{\mathrm{p}(i : A)\}_{i=1}^{p}$ quantify the significance of the differential correlation of each variable relative to A. To select a set of significant variables A', we apply the modified FDR procedure of Benjamini and Yekutieli to the p-values. Specifically, we carry out the following steps:

1. Order the p-values $\{\mathrm{p}(i : A)\}_{i=1}^{p}$ as $\{\mathrm{p}_{(1)}, \dots, \mathrm{p}_{(p)}\}$.
2. Define the cutoff index k^* by

$$k^* = \max\left\{k : \mathrm{p}_{(k)} < \left(\sum_{i=1}^{p} 1/i\right)^{-1} \left(\frac{k\alpha}{p}\right)\right\}. \tag{2.7}$$

3. Let $A' = \{i : \mathrm{p}(i : A) \le \mathrm{p}_{(k^*)}\}$.

Recall that we impose no assumptions on the structure of correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$. In particular, it is possible that p-values $\mathrm{p}(i : A)$ and $\mathrm{p}(j : A)$ may be negatively correlated. For example, it is common in genetics for individual pairs of genes to exhibit negative correlation; in this case, a low p-value for one gene will imply a high p-value for the other. Most common multiple testing methods assume independence or positive dependency between p-values. The possibility of negative dependency of p-values necessitates a more conservative multiple testing method such as that of of Benjamini and Yekutieli (2001), which controls the expected false discovery rate at level α under negative dependence.

The main search procedure terminates when it degenerates ($A = \varnothing$) or converges ($A = A' \neq \varnothing$). For the degenerate case, the interpretation is simple: the initial set (chosen in the first step of the DCM procedure) was not significantly differentially correlated. In the second case, we have identified an empirical DC clique, in the sense that by design, a nonempty fixed point A meets the first requirement of a DC clique in Definition 1.1 up to a level of statistical significance. The only other possible outcome of the iterative search procedure is a multiset cycle, which is discussed in Section 2.5.

REMARK. The DCM algorithm does not require the use of Benjamini and Yekutieli (2001) specifically; any multiple testing method controlling FDR would suffice in principle. In our experience, changes to the underlying multiple testing procedure had only minor effects on the results.

2.4. *Residualization.* In general, we expect multiple DC cliques in a dataset. The residualization step allows the DCM procedure to search the same dataset many times, avoiding repeated results. Suppose an empirical DC clique A has been selected. Our approach is to estimate a rank one approximation of correlation matrices $\widehat{\mathbf{R}}_{1,A}$ and $\widehat{\mathbf{R}}_{2,A}$ via factor analysis [Harman (1960)]. We then substitute the relevant submatrices, $\mathbb{X}_{1,A}$ and $\mathbb{X}_{2,A}$, with residualized data for which the estimated rank one correlation has been removed. Methods of estimation and removal of low-rank correlation have been well established in the literature. In the DCM software, we use the implementation of Friguet, Kloareg and Causeur (2009) for the R Statistical Software version and the method of Bishop (2006) for the Matlab version.

By opting for rank-one approximation, we are taking a conservative approach to residualization. It is conceivable that the correlation structure of A is of higher rank. If so, A may be selected more than once by DCM; however, since each time the data is being further residualized, we are guaranteed to eventually remove all structure of A. In practice, we have yet to encounter a duplicate result from real data.

2.5. *Special cases.* *Minimality:* A nonempty fixed point A of the set update procedure has the property that, analogously to Definition 1.1, $H_0(i, A)$ is rejected if and only if $i \in A$. The second condition of Definition 1.1, however, is not guaranteed in general. It is possible that DCM may select a large set that in truth consists of two (or more) disjoint population DC cliques. These cases are well addressed by the residualization step. When a conglomerate estimated DC clique is residualized, the *joint* structure is removed, leaving behind the individual structure of the disjoint DC cliques. Further runs of the DCM algorithm are then able to identify the separate DC cliques.

In extreme cases, the sampled data may be such that the disjoint DC cliques are, by chance, correlated enough to have negligible remaining individual structure after residualization. This correlation may render the multiple DC cliques indistinguishable in the data from a combined DC clique.

Cycles: Under certain conditions, the main search procedure terminates in a cycle of two or more sets. When the set update procedure oscillates between two sets A_1 and A_2, we restart the search on the intersection $A = A_1 \cap A_2$. In this case, the algorithm usually converges to a fixed point in the vicinity of the intersection. If the oscillation persists, we output the intersection $A = A_1 \cap A_2$. This overlap set has the property that $H_0(i, A)$ will be rejected for all $i \in A_1, A_2$, so it is worth attention as an empirical DC clique.

Cycles of length greater than two are rarely observed in real or simulated data. However, to protect against longer cycles leading to infinite loops, the algorithm terminates at a maximum iteration limit.

Repetition to exhaustion: In principle, the DCM procedure can be run from many initial sets. In practice, we consider the procedure to have been "run to exhaustion" if every variable has been included in at least one initial set and/or output set. Our implementation of the method is thus designed to randomly choose initial sets at each run from the set of unused variables. Note that this approach does *not* prevent variables from appearing in multiple output sets.

3. Properties of the test statistic. We now discuss some properties of the test statistic $\hat{\Delta}(i, A)$ used in the calculation of p-values for the set update procedure.

3.1. *Geometric interpretation.* The equation for $\hat{\Delta}(i, A)$ given in (2.4) expresses the test statistic directly in terms of average differential correlation. However, we may also write $\hat{\Delta}(i, A)$ in an alternate form that yields an informative geometric interpretation. Let $\tilde{\mathbf{U}}_i \in \mathbb{R}^{n_1}$ and $\tilde{\mathbf{V}}_i \in \mathbb{R}^{n_2}$ be the standardized measurements of variable i under sample Conditions 1 and 2, respectively; and let

$$\mathbf{W}_1 := \frac{1}{|A|} \sum_{j \in A} \tilde{\mathbf{U}}_j \quad \text{and} \quad \mathbf{W}_2 := \frac{1}{|A|} \sum_{j \in A} \tilde{\mathbf{V}}_j \tag{3.1}$$

be the vector means of the standardized measurements of the variables in A under each condition. It is easily shown that

$$\frac{1}{|A|} \sum_{j \in A} \widehat{\mathrm{cor}}(\mathbf{U}_i, \mathbf{U}_j) = \mathbf{W}_1^t \tilde{\mathbf{U}}_i = \|\mathbf{W}_1\| \widehat{\mathrm{cor}}(\tilde{\mathbf{U}}_i, \mathbf{W}_1)$$

and, therefore,

$$\hat{\Delta}(i, A) = \|\mathbf{W}_1\| \widehat{\mathrm{cor}}(\mathbf{W}_1, \tilde{\mathbf{U}}_i) - \|\mathbf{W}_2\| \widehat{\mathrm{cor}}(\mathbf{W}_2, \tilde{\mathbf{V}}_i).$$

Note that the vector $\tilde{\mathbf{U}}_i$ and the vectors $\{\tilde{\mathbf{U}}_j : j \in A\}$ lie on the surface of an $(n_1 - 2)$-dimensional sphere in $\mathbb{R}^{n_1}$, and that $\mathbf{W}_1$ is the geometric center (centroid) of the latter collection. The norm $\|\mathbf{W}_1\|$ is between 0 and 1; large values of $\|\mathbf{W}_1\|$ correspond to the centroid being closer to the surface of the sphere, indicating that the vectors $\{\tilde{\mathbf{U}}_j : j \in A\}$ are tightly clustered, or equivalently, highly intercorrelated. Thus the quantity $\|\mathbf{W}_1\| \widehat{\mathrm{cor}}(\mathbf{W}_1, \tilde{\mathbf{U}}_i)$ weights the similarity of $\mathbf{U}_i$ and the centroid $\mathbf{W}_1$ according to the overall similarity of the vectors $\{\tilde{\mathbf{U}}_j : j \in A\}$. Similar remarks apply to $\{\tilde{\mathbf{V}}_j : j \in A\}$ and $\mathbf{W}_2$. One may therefore interpret the average correlation between a variable i and a set A as a balance between the intracorrelation of A and the individual contribution of i. The statistic $\hat{\Delta}(i, A)$ is the difference between this measure in Conditions 1 and 2.

Figure 3 gives a simple two-dimensional representation of the geometric picture discussed above. In Condition 1, $\mathbf{U}_i$ is not strongly correlated with $\mathbf{W}_1$, but $\|\mathbf{W}_1\|$ is large because the vectors indexed by A are tightly clustered. In Condition 2, $\mathbf{V}_i$ is strongly correlated with $\mathbf{W}_2$, but $\|\mathbf{W}_2\|$ is small because the vectors indexed by A are not tightly clustered. In this example, $\hat{\Delta}(i, A)$ is close to zero, and we would likely conclude no differential correlation is present.

3.2. *Asymptotic distribution of* $\hat{\Delta}(i, A)$. We now discuss the asymptotic distribution of $\hat{\Delta}(i, A)$, from which the p-values used in Section 2.3 are derived. First, we make note of a classical result concerning sample correlations.

THEOREM 1 [Steiger and Hakstian (1982)]. *Let* $\mathbf{R}$ *be a* $p \times p$ *correlation matrix*, *and* $\widehat{\mathbf{R}}$ *the corresponding sample correlation matrix based on* n *i.i.d. samples*

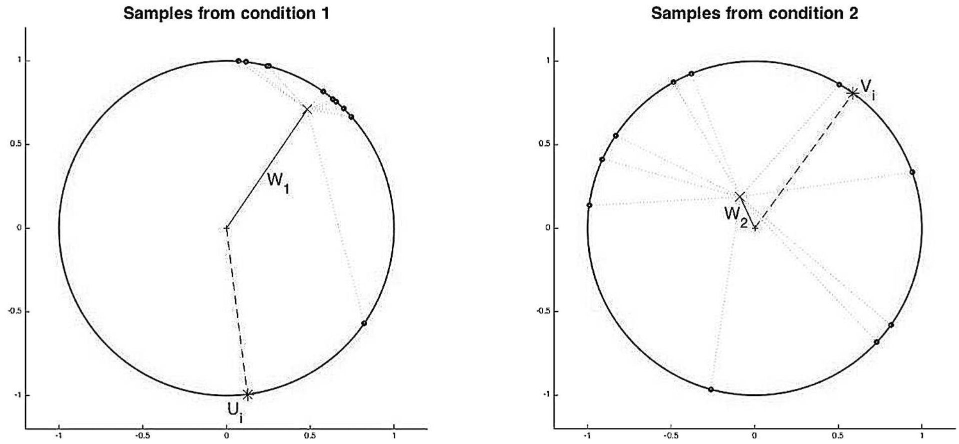

FIG. 3. *Geometric representation of data in two dimensions. Unlabelled points represent the standardized data in group A.*

of p-variate data with finite 4th moment. Let $\mathbf{P}$ and $\widehat{\mathbf{P}}$ be the vectorized versions of the matrices, of dimension $p^2 \times 1$. Then, as n tends to infinity

$$\sqrt{n}(\widehat{\mathbf{P}} - \mathbf{P}) \Rightarrow \mathcal{N}_{p^2}(\mathbf{0}, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ is a $p^2 \times p^2$ covariance matrix for which a general form is given equations (3.1–3.5) *in* Browne and Shapiro (1986).

Using Theorem 1, one may evaluate the asymptotic distribution of $\hat{\Delta}(i, A)$, which is a function of $\mathbf{P}$ and $\widehat{\mathbf{P}}$. (A proof of Corollary 1.1 is supplied in Appendix 1.)

COROLLARY 1.1. *Let A be a fixed index set and let $\hat{\Delta}(i, A)$ be defined as in* (2.4), *with sample correlation matrices $\widehat{\mathbf{R}}_1$ and $\widehat{\mathbf{R}}_2$ based on n_1 and n_2 independent samples from distributions F_1 and F_2, respectively. Let $\sigma_0^2(i, A) := \operatorname{var}(\hat{\Delta}(i, A) \mid H_0)$, where H_0 is the null hypothesis in* (2.3). *Then, under H_0 and as $\min(n_1, n_2) \to \infty$,*

$$\frac{\hat{\Delta}(i, A)}{\sigma_0(i, A)} \Rightarrow \mathcal{N}(0, 1). \tag{3.2}$$

In practice, the variance $\sigma_0^2(i, A)$ is not known. We can use the results in Steiger and Hakstian (1982) for the asymptotic variance of $\hat{\Delta}(i, A)$, which leads to a consistent estimator $\hat{\sigma}_0(i, A)$, the derivation of which is detailed in the supplementary material to this paper [Bodwin, Zhang and Nobel (2018)]. We note that regardless of the size of A, the calculation of $\hat{\sigma}_0(i, A)$ requires basic operations on only three n_1 vectors and three n_2 vectors. Such algebraic simplification is important, since in practice the variance estimate must be calculated separately for *every* variable $i \in [p]$ and for multiple iterative steps of the DCM algorithm.

REMARK. The results of Corollary 1.1 and the variance estimator in the supplementary material [Bodwin, Zhang and Nobel (2018)] apply to variable sets of fixed cardinality ($|A| = k$) as n_1 and n_2 tend to infinity. In practice, one may encounter variables sets for which $k > n_1, n_2$. Simulations suggest that the DCM algorithm still identifies DC cliques with high success and controls false discovery in such settings even when the cardinality of $|A|$ greatly exceeds the sample size.

4. Simulation study. To test the DCM method against possible alternatives, we implemented a simple study of performance on simulated data. We created artificial datasets containing a single DC clique and compared the results of several methods to the known truth. Although the simulated setting is not a perfect representation of real data situations, it readily illustrates the advantages of DCM as opposed to existing methods.

4.1. *Simulated data.* We generated data with a single embedded DC clique, consistent with Definition 1.1. Our study varied the following parameters: size of the DC clique (k), total number of variables (p), strength of the true correlations in each sample condition (ρ_1 and ρ_2), and samples sizes of the two conditions (n_1 and n_2). In both sample conditions, the DC clique signal was layered on top of either (a) uncorrelated Gaussian noise or (b) a randomly real data sample from The Cancer Genome Atlas gene expression data.

4.2. *Methods implemented.* To compare DCM to alternate approaches, we implemented or adapted representative methods from those in Section 1.2 to search for DC cliques.

Detection of isolated changes *(DCP)*. Although the goal of DCM is to identify *sets* of variables, certain existing methods are designed to find *individual* (or isolated) variables whose correlations structure changes across conditions. The Differential Correlation Profile (DCP) method of Liu et al. (2010) is one such approach, using permutation of samples to determine the significance of correlation difference for each individual variable. Importantly, this approach identifies a list of individual differentially correlated variables, rather than a united set. For the purposes of this study, we treated the collection of selected variables as an estimated DC clique.

Mining a single correlation matrix *(WGCNA, NetTop)*. One approach to mining differential correlation is to analyze each sample condition separately, then compare results. The Network Topology (NetTop) method of Bockmayr et al. (2013) creates network representations for each of the two sample conditions by thresholding the corresponding Fisher-transformed sample correlation matrices. Connected components that appear in one network and not the other are considered to be differentially correlated variable sets.

The Weighted Gene Co-Expression Network Analysis (WGCNA) method of Langfelder and Horvath (2008) is a hybrid approach which mines for clusters (or "modules") in a single correlation matrix, then tests each module for differential *expression* across conditions. Thus, although the WGCNA method involves both differential and second-order elements, it is not designed to search for DC cliques or similar structures. For the purposes of this simulation study, we applied WGCNA to samples from Condition 1 only. We then tested the output module for differential correlation via a standard t-test over sample correlations in Conditions 1 and 2. In this way, we attempted to only select variable sets exhibiting differential correlation, even though WGCNA does not naturally identify modules with this property.

Mining dissimilarity matrices *(hclust, D-Est, DiffCoEx).* Another possible approach is to summarize differential correlation in a single dissimilarity matrix, then select variable sets via ordinary clustering methods. We implemented a straightforward version of this approach, applying hierarchical clustering to the difference of sample correlation matrices, $\widehat{\mathbf{R}}_1 - \widehat{\mathbf{R}}_2$. To circumvent the challenge of selecting a cutoff in the dendrogram, we instead chose the first cluster of size less than or equal to the true DC clique. (In practice, the true size would not be known, so we would be less sure of the "best" cutoff point.) We also applied this idealized hierarchical clustering to $\widehat{\mathbf{D}}$, the estimator suggested in Cai and Zhang (2014) for directly estimating $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$. Finally, the DiffCoEx method of Tesson, Breitling and Jansen (2010) is a modification of WGCNA; a dissimilarity matrix is created based on adjusted sample correlations, then the clustering approach of WGCNA is applied.

4.3. *Results.* We applied the seven proposed methods (DCM, DCP, NetTop, WGCNA, hclust, D-EST, and DiffCoEx) to several simulated datasets at each of many parameter combinations. We found that all methods behaved similarly with regard to changes in sample sizes n_1, n_2, and clique size k (relative to p). Here, we present only the results regarding the correlation signal size, ρ_1 versus ρ_2, and the different background types, to illustrate key differences in performance between methods. By default, the other parameters were set to be $n_1 = n_2 = 100$, $k = 100$, and $p = 1000$.

To control false discovery, we disregarded output variable sets with more than 5% false positive elements. Figure 4 shows the percent of variables in the seeded DC clique that were successfully identified by each method after false discovery screening, for various strengths of true differential correlation ($\rho_1 - \rho_2$ grows). Figure 5 examines the scenario where $\rho_1 = \rho_2 \neq 0$; that is, when correlation was present in both sample conditions but not differential. Figure 5 shows the size of selected variable sets—ideally, DC mining methods would return no results in these cases. All result reflect an average of 10 simulations at each data point, with all other parameters set to default values.

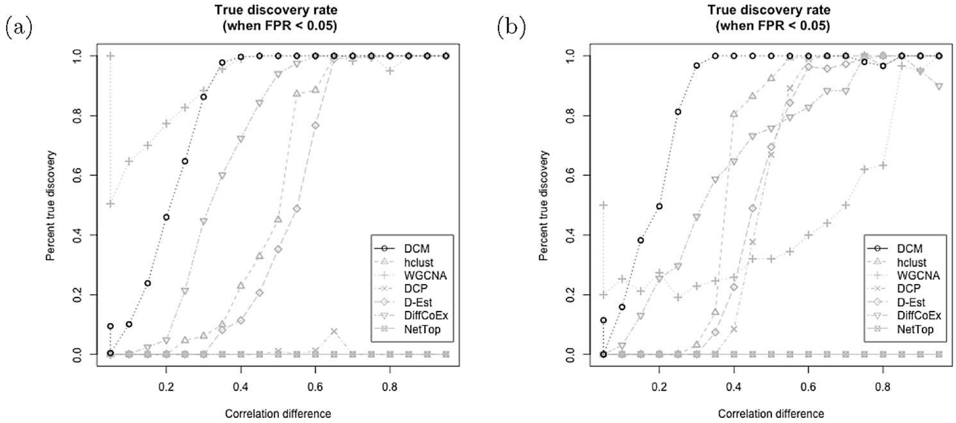

FIG. 4. *True discovery rates when false positive controlled at* 0.05 *level*, *for Gaussian noise background* (a) *and real data background* (b).

DCM was able to control false positives in all cases except for some error when there was very low signal in the real data background, which may be due to actual signal being present in the randomized real data. DCM also began to reliably detect DC cliques at a lower signal (around a correlation difference of 0.2 at the default parameters) than every method except WGCNA with Gaussian background.

In randomized real data [Figure 4(b)], *WGCNA* did not control the false positive rate. WGCNA is a method for nondifferential analysis, so, when applied to Condition 1 data, it correctly identifies many correlated variables, even though they are often equally correlated in Condition 2. Although we have adapted the method to

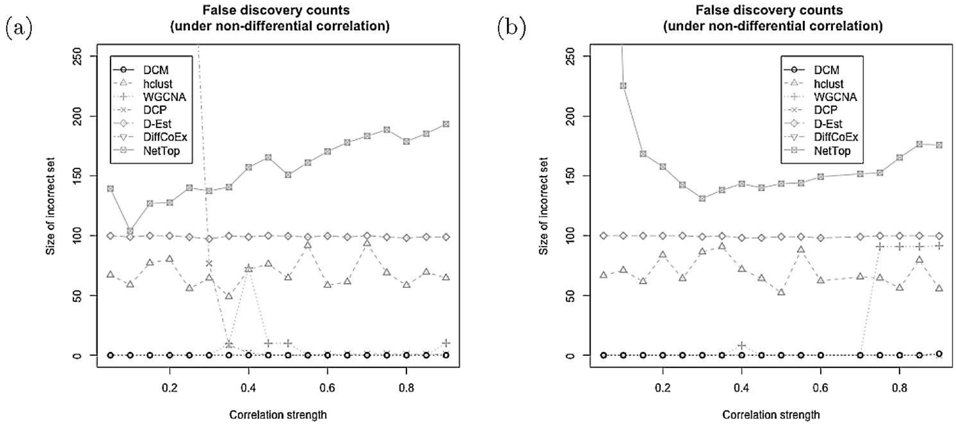

FIG. 5. *Sizes of incorrect variables sets when no differential correlation is present*, *for Gaussian noise background* (a) *and real data background* (b).

test selected modules for differential correlation, true DC cliques are obscured by existing nondifferential structure.

The *hclust* and *D-EST* approaches behave as expected: because we chose a cutoff of the hierarchical clustering dendogram by size, our approach necessarily returns a nonempty variable set. The false positive rate was consequently high for small or no signal. Similarly, *NetTop* relies on a thresholding procedure to maximize differences between conditions, so it is likely to find signal even when none is present. However, even if the false positives were perfectly controlled in some way, these methods show a lower detection point than DCM.

DiffCoEx performed the strongest in our simulations, as it was able to control false discovery in most cases while still detecting DC cliques at a reasonable rate. DCM, however, proved more sensitive without sacrificing error control.

Finally, *DCP*, and any approach that seeks isolated structure rather than unified sets, is likely to greatly overselect variables in the uncorrelated background case because the mutual behavior of the variables in a DC clique will induce some correlation structure in the extraneous variables. Figure 1 illustrates this phenomenon, as there is some pattern in the cross correlation between variables in B and A. This result emphasizes the danger of the common approach of looking for isolated changes in correlation structure of individual variables, rather than searching for DC cliques: vestigial correlation patterns may be misleading.

REMARK. We also implemented versions of the iterative testing update procedure using different hypothesis testing approaches, including a Normal approximation to Fisher-transformed data and a classic likelihood ratio test as derived in Muirhead (1982). We found that neither approach yielded a higher discovery rate (with controlled FDR) than DCM.

4.4. *Computation.* Figure 6 shows the computation times for all tested methods on a log scale and an absolute scale. Since modern datasets tend to have dimension in the tens or hundreds of thousands of variables, the exponential differences between method runtimes are crucial to the practicality of analysis. All methods except the basic hclust required exponentially more runtime than DCM.

One important limitation of common approaches to correlation mining (including DCP, D-Est, hclust, and NetTop) is that memory demands scale on the order of at least p^2, as they necessitate estimation of full p by p dissimilarity matrices. Permutation- or repetition-based methods such as DCP and NetTop are even more infeasible in high dimensions, since they require the computation of a p by p correlation matrix for each of many permutations (this is why the simulations were truncated in Figure 6). An advantage of DCM is that only a the $|A| \times p$ portion of sample correlation matrices corresponding to proposed set A must be computed at any given time.

Further simulation study results are available in the supplement to this paper [Bodwin, Zhang and Nobel (2018)].

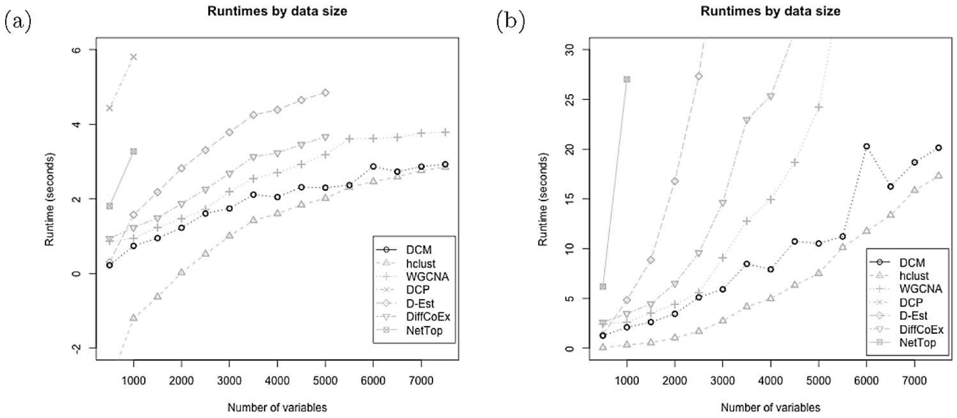

FIG. 6. *Computation time to find a single variable set at log scale* (a) *and exact scale* (b).

5. Data analyses. The efficiency of DCM allowed us to study differential correlation in two very high-dimensional settings: gene expresssion data ($\sim 10^4$ variables) and fMRI brain scan data ($\sim 10^5$ variables). Both of these datasets are beyond the computation limits of the alternate methods discussed in Section 4 without access to extraordinary computing resources.[4] In both applications, DCM was able to identify empirical DC cliques of interpretable scientific merit.

5.1. *TCGA.* As introduced in Figure 1, we applied the DCM procedure to data from The Cancer Genome Atlas, with samples from two predetermined breast cancer subtypes: Her-2 and Luminal B. A total of 18 empirical DC cliques (more correlated in Her-2 than in Luminal B) were discovered, ranging in size from 30 to 108 genes.

To illustrate how this information may be useful to genomic research, we briefly discuss one of the discovered gene sets. The set of interest contained 46 genes is listed alphabetically in Table 1. These genes are found to be highly associated with immune response, particularly the HLA (Human Leukocyte Antigen) gene class, represented by six of the genes in the set. Researchers are interested in understanding how and why some cancer subtypes trigger immune response while others do not. For example, Iglesia et al. (2014) showed that prognosis was improved for patients with Her-2 and Basal-like subtypes showing higher immunoreactive response. Further exploration of DC cliques such as the one in Table 1 may further understanding of the gene interactions that drive immune response.

5.2. *The human connectome project.* The Human Connectome Project is a multi-institutional venture aimed at mapping functional connections between parts

[4] A brief comparison of the methods using a truncated portion of the TCGA dataset is provided as supplemental material.

TABLE 1
Genes selected in empirical DC Clique for Her2 versus Luminal B samples

AGER	amt	APOL1	ARPC4	B2M	BATF2	BTN3A2
BTN3A3	C19orf38	calml4	CCDC146	CHKB-CPT1B	echdc1	ETV7
EXOSC10	FBXO6	GBP1	GBP4	GJD3	gnb3	*HLA-A*
HLA-B	*HLA-C*	*HLA-E*	*HLA-F*	*HLA-H*	HSH2D	IDO1
IL15	Irf1	LOC115110	LOC400759	LOC91316	micB	Myo15b
OASL	PILRB	Rec8	Rufy4	SAMD9L	SEC31B	STAT1
tap1	Tapbp	TTLL3	TXNDC6	Ube2l6	Zbp1	

of the human brain. The project has collected vast amounts of brain scan data, all of which is publicly available to researchers online at www.humanconnectome.org.[5] In this analysis, we made use of a dataset from the "500 Subjects MR" data release, which consists of functional magnetic resonance imaging (fMRI) brain scans for 542 healthy adult subjects. Participants performed a variety of tasks during the MR scan, designed to isolate certain types of brain functionality. Activation levels were recorded over time for ~30,000 voxels (3D coordinate locations in the brain's white matter interior) and ~60,000 greyordinates (indexed locations over the grey matter brain surface).

In this paper, we applied DCM to data from a single subject.[6] We compared two task categories:

Language-based tasks: During the scan, subjects were told brief stories and asked to answer questions after each one about what they were told.

Motor-based tasks: Subjects were attached to motion sensors at the hands, feet, and tongue. They were then asked to move one appendage at a time, in blocks of repetitions.

DCM was applied to 91,282 brain locations (or nodes) to find DC cliques that exhibit more correlation over time during language tasks than during motor tasks. On a home computer, this process took under a minute to find the first DC clique, running in Matlab. Continuing to exhaustion took approximately an hour. We discovered 10 total empirical DC cliques, in sizes ranging from 1688 (displayed) to 20.

The first empirical DC clique selected by DCM contained 1688 nodes located on the cortical surface. These nodes, or "greyordinates," are visualized as points on the smoothed exterior of the brain in Figure 7. The clear locational pattern in the nodes—despite the fact that the analysis did not take location into account—is striking. Additionally, the empirical DC clique in Figure 7 includes a concentrated group in the rear of the left cortex. This general brain region is known to be specif-

[5] Data was available in preprocessed form; see http://www.humanconnectome.org/about/project/MR-preprocessing.html for further detail.

[6] Subject #101006, a 35-year-old female.

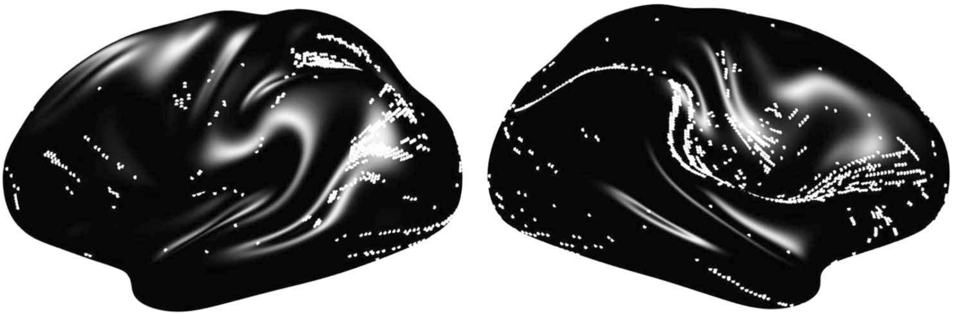

FIG. 7. *Brain locations of DC clique for languages tasks versus motor tasks.*

ically associated with language processing and auditory input [Wernicke's Area, see Wang et al. (2015)].

We also studied two other artifacts of the data for comparison. First, we identified the 1000 nodes exhibiting the strongest differential first-order behavior. These show higher mean activation during the language tasks than during the motor tasks, as measured by standard two-sample t-tests. We saw a clear grouping of nodes in the right frontal lobe. This pattern is unsurprising and appears in many studies of brain functionality that examine differential activation for language processing [Voets et al. (2006)]. This basic first-order analysis suggests that differential correlation is not redundant. None of the empirical DC cliques selected by DCM show high frontal lobe concentration; instead, they exhibit "trail-like" patterns such as the ones shown in Figure 7.

Second, we identified 1000 nodes found to be highly correlated over time for the language task data, irrespective of their behavior in the motor task data. These nodes were observed to be very tightly grouped in the interior left hemisphere. This is likely due to the nature of data measurement: fMRI brain scans measure oxygen flow in the brain, so measurements for adjacent regions tend to "blur" and show high artificial correlation [Derado, Bowman and Kilts (2010)]. In this case, the same node set is also highly correlated during motor tasks, suggesting that it is likely a byproduct of data collection. Even if this node set does represent a meaningful result—regions, perhaps, that are universally correlated regardless of task—it is not differential.

This example illustrates the advantage of taking a differential approach like DCM. Effects due to fMRI-driven spatial correlation or strong universal correlation can drown out signal that is truly specific to a particular sample condition. By comparing language tasks to the similar but distinct condition of motor tasks, we are able to isolate signals that are unique to language processing. The fact that the identified DC cliques show emergent locational patterns suggests that DCM is capturing a true facet of the data rather than arbitrary correlation. Since this output is unique in form, while maintaining some consistency with known brain functionality, we believe it merits further scientific investigation.

6. Conclusion and future work.

In this paper, we have introduced a new statistical method, DCM, to identify differentially correlated variable sets from observed data. The DCM algorithm is a statistically principled approach to data mining which incorporates hypothesis testing into its search procedure. It is applicable in many areas of scientific research, including statistical genetics and neuroscience. The DCM software can be run on extremely high-dimensional data (at least $\sim 10^5$ samples and/or variables) without large memory demands or long runtimes.

Future directions. Many similar association mining methods may be extended from the DCM framework. It may be of interest to study differential mining from other measures of association, such as rank-based correlation, which would require results analogous to Theorem 1. One may also consider datasets with more than two sample conditions or even a continuous response. Further theoretical work may also be able to establish the results of Corollary 1.1 in cases where the cardinality of the proposed variable set A is increasing with the sample size.

Code for public use of DCM is freely available at http://github.com/kbodwin/.

Acknowledgments. The authors wish to thank Yin Xia, Andrey Shabalin, Katherine Hoadley, and Kimberly D. T. Stachenfeld for their contributions; as well as the Editor, Associate Editor, and reviewers whose thoughtful comments and suggestions greatly improved this paper.

SUPPLEMENTARY MATERIAL

Differential correlation mining: Supplementary material (DOI: 10.1214/17-AOAS1083SUPP; .pdf). We provide the proof of Corollary 1.1, the derivation of the variance estimator, additional simulation results, extended real data results, and pseudocode for the algorithmic procedures.

REFERENCES

ANDERSON, T. W. (1959). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

BASSI, F. and HERO, A. (2012). Large scale correlation detection. In *Proc. of the IEEE International Symposium on Information Theory* 2591–2595.

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York. MR2247587

BOCKMAYR, M., KLAUSCHEN, F., GYÖRFFY, B., DENKERT, C. and BUDCZIES, J. (2013). New network topology approaches reveal differential correlation patterns in breast cancer. *BMC Syst. Biol.* **7** 78.

BODWIN, K., ZHANG, K. and NOBEL, A. (2018). Supplement to "A testing based approach to the discovery of differentially correlated variable sets." DOI:10.1214/17-AOAS1083SUPP.

BROWNE, M. W. and SHAPIRO, A. (1986). The asymptotic covariance matrix of sample correlation coefficients under general conditions. *Linear Algebra Appl.* **82** 169–176. MR0858970

CAI, T. T. and JIANG, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist.* **39** 1496–1525. MR2850210

CAI, T., LIU, W. and XIA, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108** 265–277. MR3174618

CAI, T. T. and ZHANG, A. (2014). Inference on high-dimensional differential correlation matrix. Technical report.

CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

CHOI, Y. and KENDZIORSKI, C. (2009). Statistical methods for gene set co-expression analysis. *Bioinformatics* **25** 2780–2786.

CUI, X. and CHURCHILL, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* **4** 210.

DATTA, S. and DATTA, S. (2002). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* **19**.

DERADO, G., BOWMAN, F. D. and KILTS, C. D. (2010). Modeling the spatial and temporal dependence in fMRI data. *Biometrics* **66** 949–957. MR2758231

DONNER, A. and ZOU, G. (2014). Testing the equality of dependent intraclass correlation coefficients. *J. R. Stat. Soc., D* **51** 367–379.

FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571

FUKUSHIMA, A. (2013). DiffCorr: An R package to analyze and visualize differential correlations in biological networks. *Gene* **518** 209–214.

GILL, R., DATTA, S. and DATTA, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinform.* **11** 95.

GREICIUS, M. D., KRASNOW, B., REISS, A. L. and MENON, V. (2002). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. USA* **100** 253–258.

HARMAN, H. H. (1960). *Modern Factor Analysis*. Univ. Chicago Press, Chicago, Ill. MR0159393

HU, R., QIU, X. and GLAZKO, G. (2010). A new gene selection procedure based on the covariance distance. *Bioinformatics* **26** 348–354.

IGLESIA, M. D., VINCENT, B. G., PARKER, J. S., HOADLEY, K. A., CAREY, L. A., PEROU, C. M. and SERODY, J. S. (2014). Prognostic B-cell signatures using mRNA-Seq in patients with subtype-specific breast and ovarian cancer. *Clin. Cancer Res.* **20** 3818–3829.

JIANG, D., TANG, C. and ZHANG, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Trans. Knowl. Data Eng.* **16** 1370–1386.

KRIEGEL, H.-P., KRÖGER, P. and ZIMEK, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data* **3**.

LANGFELDER, P. and HORVATH, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **9** 559.

LEWIS, K., KAUFMAN, J., GONZALEZ, M., WIMMER, A. and CHRISTAKIS, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Soc. Netw.* **30** 330–342.

LIU, B.-H., YU, H., TU, K., LI, C., LI, Y.-X. and LI, Y.-Y. (2010). DCGL: An R package for identifying differentially coexpressed genes and links from gene expression microarray data. *Bioinformatics* **26** 2637–2638.

MACMAHON, M. and GARLASCHELLI, D. (2015). Community detection for correlation matrices. *Phys. Rev. X* **5** 021006.

MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York. MR0652932

PENG, J., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. MR2541591

PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H., AKSLEN, L. A., FLUGE, O., PERGAMENSCHIKOV, A., WILLIAMS, C., ZHU, S. X., LONNING, P. E., BORRESEN-DALE, A.-L., BROWN, P. O. and BOTSTEIN, D. (2000). Molecular portraits of human breast tumours. *Nature* **406** 747–752.

PHAN, K. L., WAGER, T., TAYLOR, S. F. and LIBERZON, I. (2002). Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage* **16** 331–348.

RAJARATNAM, B., MASSAM, H. and CARVALHO, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.* **36** 2818–2849. MR2485014

SHENG, E., WITTEN, D. and ZHOU, X.-H. (2016). Hypothesis testing for differentially correlated features. *Biostatistics* **17** 677–691. MR3604273

SONESON, C. and DELORENZI, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform.* **14** 91.

STAMATATOS, E. (2009). A comparison of methods for differential expression analysis of RNA-seq data. *J. Am. Soc. Inf. Sci. Technol.* **60** 538–556.

STEIGER, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychol. Bull.* **87** 245–251.

STEIGER, J. H. and HAKSTIAN, A. R. (1982). The asymptotic distribution of elements of a correlation matrix: Theory and application. *Br. J. Math. Stat. Psychol.* **35** 208–215. MR0683508

TESSON, B. M., BREITLING, R. and JANSEN, R. C. (2010). DiffCoEx: A simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinform.* **11** 497.

VOETS, N. L., ADCOCK, J. E., FLITNEY, D. E., BEHRENS, T. E., HART, Y., STACEY, R., CARPENTER, K. and MATTHEWS, P. M. (2006). Distinct right frontal lobe activation in language processing following left hemisphere injury. *Brain* **129** 754–766.

WAINER, H. and BRAUN, H. I. (2013). *Test Validity*. Routledge, London.

WANG, J., FAN, L., WANG, Y., XU, W., JIANG, T., FOX, P. T., EICKHOFF, S. B., YU, C. and JIANG, T. (2015). Determination of the posterior boundary of Wernicke's area based on multimodal connectivity profiles. *Hum. Brain Mapp.* **36** 1908–1924.

WILSON, J. D., WANG, S., MUCHA, P. J., BHAMIDI, S. and NOBEL, A. B. (2014). A testing based extraction algorithm for identifying significant communities in networks. *Ann. Appl. Stat.* **8** 1853–1891. MR3271356

XIA, Y., CAI, T. and CAI, T. T. (2015). Testing differential networks with applications to the detection of gene–gene interactions. *Biometrika* **102** 247–266. MR3371002

ZHOU, C., HAN, F., ZHANG, X. and LIU, H. (2015). An extreme-value approach for testing the equality of large U-statistic based correlation matrices. Available at arXiv:1502.03211.

K. BODWIN
K. ZHANG
A. NOBEL
DEPARTMENT OF STATISTICS AND
OPERATIONS RESEARCH
UNIVERSITY OF NORTH CAROLINA
AT CHAPEL HILL
CHAPEL HILL, NORTH CAROLINA 27599
USA
E-MAIL: kbodwin@calpoly.edu
zhangk@email.unc.edu
nobel@email.unc.edu