Balanced Functional Module Detection in genomic data

David Tritchler¹², Lorin M. Towle-Miller ¹* and Jeffrey C. Miecznikowski¹

¹Department of Biostatistics, University at Buffalo, Buffalo, NY 14260, USA and ²Biostatistics Division, University of Toronto, Toronto, ON M5S 1A1, Canada

*To whom correspondence should be addressed.

Associate Editor: Marieke Lydia Kuijjer

Received on March 26, 2021; revised on July 26, 2021; editorial decision on August 12, 2021

Abstract

Motivation: High-dimensional genomic data can be analyzed to understand the effects of variables on a target variable such as a clinical outcome. For understanding the underlying biological mechanism affecting the target, it is important to discover the complete set of relevant variables. Thus variable selection is a primary goal, which differs from a prediction criterion. Of special interest are functional modules, cooperating sets of variables affecting the target which can be characterized by a graph. In applications such as social networks, the concept of balance in undirected signed graphs characterizes the consistency of associations within the network. This property requires that the module variables have a joint effect on the target outcome with no internal conflict, an efficiency that may be applied to biological networks.

Results: In this paper, we model genomic variables in signed undirected graphs for applications where the set of predictor variables influences an outcome. Consequences of the balance property are exploited to implement a new module discovery algorithm, balanced Functional Module Detection (bFMD), which selects a subset of variables from high-dimensional data that compose a balanced functional module. Our bFMD algorithm performed favorably in simulations as compared to other module detection methods. Additionally, bFMD detected interpretable results in an application using RNA-seq data obtained from subjects with Uterine Corpus Endometrial Carcinoma using the percentage of tumor invasion as the outcome of interest. The variables selected by bFMD have improved interpretability due to the logical consistency afforded by the balance property.

Contact: lorinmil@buffalo.edu

Supplementary information: Supplementary data are available at Bioinformatics Advances online.

1 Introduction

Clinical outcomes of interest, such as risk factors, disease diagnosis or treatment success may be better understood by identifying associated genes within high-dimensional genomic datasets. Additionally, these associated genes should be associated with one another, such that they may form a ‘module’ or ‘pathway’ that coordinate together to ultimately control the targeted outcome of interest. We denote these outcome-associated sets of variables as ‘functional modules’. Many variable selection methods currently exist to identify functional modules, but these existing functional module detection methods fail to consider the concept of ‘balance’ frequently used in other applications such as social networks. Balance assures no contradictions exist between variables within the functional module. For example, suppose the expression of a certain module gene promotes an increase in the outcome variable while another gene in the same module suppresses it; balance says that the two genes must be negatively correlated, otherwise they would conflict with respect to their effect on the outcome.

Since observed correlations in a balanced signed graph can be composed to reflect all path signs in the underlying graph, we can use the observed correlations to determine if a subset of the high-dimensional set of variables satisfies the balance property. This leads to our new module discovery algorithm, balanced Functional Module Detection (bFMD). The bFMD method detects which variables are contained within the module; after they are selected, various existing analytical methods may be used on the detected module variables for obtaining the graphical structure. For example, modeling predictors as an undirected graph whose nodes influence the target is related to traceable regressions (Wermuth, 2012). Traceable regression analyzes ordered causal blocks of variables and identifies direct and indirect effects among the variables. In our simple scenario, there are two blocks, the block consisting of only the target and a block of potential predictors. The traceable regression analysis
of our semidirected graph would be partitioned into an undirected graphical model analysis (e.g. Edwards, 2000; Edwards et al., 2010) of the predictors and individual regressions of the target on each predictor (Wermuth and Cox, 2013). Dimension reduction prior to a fully detailed traceable regression analysis is desirable (Bourgon et al., 2010; Towle-Miller et al., 2020), and the set of variables selected should make causal sense. Thus, initial analysis by bFMD can expand the application of existing methods to high-dimensional data and yield more interpretable models.

We emphasize that the goal of our method is variable selection, not prediction. The outcome variable is modeled to constrain the variable selection for the logical consistency of signs. Optimizing prediction and optimizing variable selection are not the same thing. For selection and scientific understanding, we want all the relevant variables, while prediction can stress parsimony. While a good prediction equation will favor functional variables, it does not require that all network variables are included to provide the basis for a biological explanation. A good example is the lasso (Tibshirani, 1996) which eliminates correlated variables, which might cloud understanding the biological mechanism. The object of our analysis to provide the scientist with a parcel of variables that allow a complete picture of the underlying biology, a set with no obvious self- contradictions in causing the response.

Our development starts with a weighted signed graph model for predictor variables where the edge weights can be positive or negative. To alleviate cumbersome notation, we will define a semidirected graph and associated directed signed edges from the predictors and individual regressions of the target on each predictor. For a given node \( i \) and \( j \), the weight \( w_{ij} \) can be positive or negative, to describe direct or inverse covariation. A univariate outcome predictor variables where the edge weights can be positive or negative.

1.1.2 Functional modules

We informally define a ‘functional module’ to be a set of coordinated variables \( X = \{x_i: i = 1, 2, \ldots, p\} \) that influence an outcome variable \( Y \). Each variable in \( X \) influences \( Y \) either directly or indirectly by operating through other variables in the set \( X \). Such modules that help describe the response \( Y \) can be profitably described by an underlying weighted undirected graph \( G = (X, E, w) \) under special properties.

\[ G \text{ and } Y \text{ create a semidirected graph } G \cup Y \text{ where each node in } G \text{ is a potential predictor of } Y \text{. The directed portion of } G \cup Y \text{ corresponds to the edges connecting from } G \text{ to } Y \text{, and the undirected portion of } G \cup Y \text{ corresponds to the edges that connect variables within } X \text{ but not } Y \text{. We represent a semidirected path } x: i \rightarrow Y \text{ in } G \cup Y \text{ as two segments: an undirected path } y: i \rightarrow j \text{ from } i \text{ in } G \text{ terminating in some } j \text{ in } G \text{ followed by a directed edge from } j \text{ to } Y \text{. We define the weight of the path } z \text{ from } i \text{ to } Y \text{ along } y \text{ to be } w^{i\rightarrow Y} = w^y \beta_{YX_{1:j}} \text{ where } w^y \text{ represents the path weight from } i \text{ to } j \text{ in } G \text{ and } \beta_{YX_{1:j}} \text{ represents the partial regression coefficient of } j \text{ in the regression model for response } Y \text{ including all variables in } X \text{ except } x \text{. Note that sign}(x_i) = \text{sign}(w^i) \times \text{sign}(\beta_{YX_{1:j}}) \text{ since } \beta_{YX_{1:j}} \text{ characterizes the directed edge, and the path weight equals } \beta_{YX_{1:j}} \text{ when a path consists of only a directed edge from } x_i \text{ to } Y \text{.}

1.1.3 Balance in functional modules

Building on the definitions described in previous sections, we further require the signed graph \( G \) to be balanced (Harary, 1953). A characterization of balance that best results in desired properties of a functional module is described in Definition 1.1. For example, if \( i \) is related to \( j \) by a path of module elements, no other \( i \rightarrow j \) paths in the network should exist with opposing signs. This balanced mechanism ensures no conflicts, and it greatly improves the interpretability in biological systems by identifying sets of variables that operate in this cooperative fashion.

Definition 1.1 Balanced Signed Graph. A signed graph \( G \) is ‘balanced’ if and only if for every two nodes of \( G \), all paths joining them have the same sign (Chartrand, 1977, Ch. 8; Harary, 1953).

To model our functional network, we extend the concept of balance to semidirected graphs by requiring all semidirected paths \( x: i \rightarrow Y \text{ to have the same sign, which results in a similar interpretation} \text{ to balance within } G \). More precisely, we require a functional network \( G \cup Y \) to have the properties described in Definition 1.2. In summary, \( G \) is a module or network (by the second condition) with no internal conflict (by the first and fourth conditions), and the module affects \( Y \) (by the third condition).

Definition 1.2 Balanced Functional Network \((G \cup Y)\). \( G \cup Y \) is a ‘balanced functional module’ when the following are satisfied.

1. \( G \) is balanced.
2. \( G \) is connected: at least one path connecting each pair of elements within \( G \) exists.
3. Every node of \( G \) is connected to \( Y \) by a semidirected path with at least one directed edge.
4. For a given node \( i \in G \), all \( i \rightarrow Y \) paths have the same sign.

In Theorem 1.1, we establish a consequent constraint on triplets of paths in a balanced functional network. Note the Supplementary Appendix contains proofs for all subsequent theorems.

Theorem 1.1 For balanced functional network \( G \cup Y \), let \( x: i \rightarrow Y \text{ equal a semidirected } i \rightarrow Y \text{ path, } y \text{ equal an } i \rightarrow j \text{ path in } G \text{, and } z: j \rightarrow Y \text{ equal a semidirected } j \rightarrow Y \text{ path. Then } w^{x \rightarrow Y} \cdot w^{y \rightarrow X} \cdot w^{z \rightarrow Y} > 0 \text{.}

1.2 Statistical model

We represent \( G \) statistically as a Gaussian undirected graphical model with a nonsingular concentration matrix \( \Theta = \Sigma^{-1} \). The properties of \( \Theta \) will mimic those from previous studies (Anufriev and Panvhenko, 2015; Sulaimanov and Koeppl, 2016; Wermuth, 1980; Whittaker, 2009). Specifically, we denote the diagonal of \( \Theta \) as \( D = \text{diag}(\Theta_{11}, \Theta_{22}, \ldots, \Theta_{pp}) \) where \( d_i = \Theta_{ii} = \text{Var}(x_i|X^{1:i})^{-1} \) denotes the reciprocal of the partial variance of \( x_i \). Additionally, letting \( I \) denote an identity matrix of appropriate size, the adjacency matrix \( A = I - D^{-1/2} \Theta D^{-1/2} \) contains partial correlations with zeros on its diagonal, making \( W = I - A \) a matrix of partial correlations. Note that the off-diagonal elements of \( A \) correspond to the edge weights from graph \( G \). From this, we see that the partial
coefficient matrix is \( C = D^{-1/2}W D^{-1/2} \) and the matrix of least
square partial regression coefficients is \( B = CD^{-1/2}WD^{1/2} \).
The covariance matrix \( \Theta^{-1} = \Sigma = (\sigma_{ij}) \) relates to the adjacency ma-
trix \( A \) of the graph \( G \) through the expression

\[
\Sigma = D^{-1/2} \left( (I-A)^{-1} - I \right) D^{-1/2} = D^{-1/2} \left( D^{-1/2} \Theta D^{-1/2} \right)^{-1} D^{-1/2} = D^{-1/2} \left( D^{-1/2} \Sigma D^{-1/2} \right) D^{-1/2}. \tag{1}
\]

The covariance structure of \( G \) can be described in terms of a closure matrix \( G' \) (Meyer, 2000, ch. 7; Sulaimanov and Koeppl, 2016). Powers of the adjacency matrix sum walk in \( G \) of a specific length, e.g. \( A^k \) equals the sum of all \( i-j \) walk weights in \( G \) of length \( k \). Then the Neumann series \( \sum_{k=1}^{-1} A^k \) quantifies the connectivity of \( G \) and converges to \( (I-A)^{-1} = D^{-1/2} \Sigma D^{-1/2} \) (Sulaimanov and Koeppl, 2016). The off-diagonal elements of \( \Sigma \) are then given by

\[
\sum_{k=1}^{-1} A^k w_{ij} = \sum_{k=1}^{-1} A^k w_{ij} D^{-1/2}, \quad \text{where each } w_{ij} \text{ corresponds to a weight in the closure } G' \text{ of } A \text{, making } \Sigma \text{ the scaled sum of weights of all } i-j \text{ walks in } G. \text{ When the spectral radius of } A, \text{ denoted } \sigma(A), \text{ is less than 1 then the Neumann series converges and } \text{sign}(\Sigma) = \text{sign}(w_{ij}). \text{ It may be shown then when all row sums of the absolute values in } A \text{ are less than 1, then } \sigma(A) < 1. \text{ This is interpreted as moderate con-
ditional dependence among the variables since } A \text{ consists of partial correlations (Anufriev and Panvhenko, 2015).}

Balance is defined in terms of paths in the literature, while the preceding development describes \( \Sigma \) using walks. To bridge this dis-
connect, we need Lemma 1.2.

**Lemma 1.2** For a balanced graph \( G \), the sign of an \( i-j \) walk is the sign of every \( i-j \) path in \( G \).

Note that Lemma 1.2 implies that every \( i-j \) walk contains the same sign because according to Definition 1.1, every \( i-j \) path must have the same sign to be balanced. An immediate consequence is shown in Theorem 1.3.

**Theorem 1.3** For a balanced graph \( G \) and corresponding covariance ma-
trix \( \Sigma = (\sigma_{ij}) \), if we assume moderate conditional dependence then the sign of \( \sigma_{ij} \) is equivalent to the sign of every \( i-j \) path in \( G \).

We extend the above concept to the semidirected case in Theorem 1.4.

**Theorem 1.4** For a balanced functional network \( G \cup Y \) with \( G \) satisfying the conditions of Theorem 1.3:

1. \( \text{Cor}(i, Y) \) is a positively weighted sum of all \( i-Y \) walk weights
2. The sign of \( \text{Cor}(i, Y) \) equals the sign of every \( i-Y \) path.
3. \( \text{Cor}(i, Y) \neq 0 \).

## 2 Methods

### 2.1 Identification of module elements

We use Theorems 1.1, 1.3 and 1.4 to guide the discovery of subsets within \( X \) which influence \( Y \) and behave as a balanced graph. To this end, we define the matrix \( M \) in Theorem 2.1, and balanced functional modules will appear as positive submatrices of this matrix.

**Theorem 2.1** If the graph \( \text{Cor}(i, Y) \neq 0 \) \( G \cup Y \) within \( X \) is a balanced functional module, then \( M = \text{diag}(\text{Cor}(X, Y)) \text{Var}(X) \text{diag}(\text{Cor}(X, Y)) \) (2)
is a positive matrix (i.e. all elements of \( M \) are \( > 0 \)).

Note that the \((i,j)th\) element of \( M \) equals \( \text{Cor}(x_i, Y)\text{Cor}(x_j, Y) \), making it large when \( i \) and \( j \) are highly correlated with each other (i.e. in the same module) and with \( Y \) (i.e. both functional). Also, if \( i \) and \( j \) are uncorrelated, or either is unrelated to the outcome, the corresponding matrix element will be zero. Thus, \( M \) will highlight sets of interrelated variables and also incorporate a functional constraint. Theorem 2.1 justifies a mechanism for identifying the subset of module elements by analyzing a matrix similar to (2) by using the full data collected. The data collected will contain a high-dimensional set of variables, and the module \( M \) will be a subset of the larger matrix. If we compute the matrix \( M' = \text{diag}(\text{Cor}(X', Y)) \text{Var}(X') \text{diag}(\text{Cor}(X', Y)) \) for the entire set of predictors \( X' \), then the balanced functional module \( M \) will appear as a positive submatrix of \( M' \). The nonzero elements correspond to correlated variable pairs that are additionally associated with \( Y \). When elements equal zero, they fail to satisfy all three conditions. Since the ordering of the variables is arbitrary, we can write \( M' \) as a matrix composed of zeros with a positive upper diagonal block \( M \) corresponding to the functional module. We then model the observed data matrix as \( W = M' + E \), where \( E \) is a matrix of mean zero errors. In other words, \( W \) is a noisy, observable version of the unobserved \( M^* \), and \( W \) inherits the interpretation given above for \( M \). Thus, the identification of a positive submatrix within \( W \) can approximate the functional module \( M \).

To depict this, note that \( W \) can be written as the Hadamard product

\[
E(W) = \text{Var}(X') \cdot \text{Cor}(X', Y) \text{Cor}(X', Y)^T. \tag{3}
\]

where the operator \( \cdot \) denotes element-wise multiplication. The left
matrix, \( \text{Var}(X') \), contains the information on the modular structure, and the right matrix, \( \text{Cor}(X', Y)^T \), is positive when the variables corresponding to that element are correlated with each other and with \( Y \). Taken from Miecznikowski et al. (2016), Figure 1 visually depicts the formation of the Hadamard product (3) for three simulated modules, one of which is functional.

### 2.2 Balanced Functional Module Detection

Assume a high-dimensional set of variables \( x_i ; (i = 1, \ldots, T) \) where \( p \) variables from this set belong to a functional module \( M \) and \( T \gg p \). Using Theorem 2.1, we first estimate \( M' = \text{diag}(\text{Cor}(X', Y)) \text{Var}(X') \text{diag}(\text{Cor}(X', Y)) \) for the full set of \( T \) variables. By the Perron Theorem for positive matrices (Meyer, 2000, ch. 8), the submatrix \( M \) contains the largest absolute eigenvalue \( \lambda > 0 \), with corresponding positive eigenvector \( e \). Subsequently, \( M' \) contains the largest absolute eigenvalue \( \lambda > 0 \) with sparse eigenvector \( (v, 0) \), where \( 0 \) denotes a vector of \( T-p \) zeros.

Our task is to extract a sparse eigenvector from matrix \( W \), a noisy version of \( M' \), where the eigenvector is composed of either zeros or positive values. Simply put, the goal is to identify variables corre-
sponding to large values within \( W \), and these variables corresponding to large values within \( W \) will be considered part of the module. Variables with large values within \( W \) will be identified using sparse principal component analysis (sPCA; Sigg and Buhmann, 2008) which returns an eigenvector of size \( T \) where element \( i \) within the eigenvector corresponds to variable \( i \). Large values within the eigenvector suggest that the variable has larger values within \( W \), and sPCA will truncate small values within the eigenvector to zero. The nonzero eigenvectors are used to identify the module \( M \). We introduce a novel method that follows naturally from the proceeding theoretical development. Note that an alternative to the more explanatory form (2) is \( W = Z'Z \), where \( Z = \text{Hdiag}(\text{Cor}(X', Y)) \) and \( H \) is the column-centered version of \( X' \). As summarized in Algorithm 1, we calculate the transformed data matrix \( Z \), recast the analysis as an sPCA of \( Z \) using the algorithm from Sigg and Buhmann (2008) and then use the nonzero loadings from the sPCA to select module elements.

---

**Algorithm 1 Balanced Functional Module Detection**

**Require** \( X', Y = X' - \text{colmmins}(X') \otimes I \), where \( I \) equals a column vector of 1’s

\[
Z = H \times \text{diag}(\text{Cor}(X', Y))
\]

\[
W = Z'Z
\]

\( q \) = eigenvector from sPCA on \( W 

Module elements = \{i : q_i \neq 0 \}
The sparse principal component method from Sigg and Buhmann (2008) specifies sparsity by inputting the number of non-zero elements, $k$. We calculate the positive eigenvector for all feasible $k$ (e.g. module sizes), and we select the sparsity setting $k$ resulting in the optimally balanced solution. Additional details are described in the Supplementary Appendix.

2.3 Evaluating balance
We provide a metric for evaluating the balance of a computed module based on an equivalent characterization of balance in an undirected signed graph: a signed graph $G$ is balanced if its nodes can be partitioned into subsets $A$ and $B$ where all edges within the subsets are positive and all edges connecting the nodes in different sets are negative (Chartrand, 1977, Ch. 8).

**Lemma 2.2** With balanced $A$ and $B$ as defined above, all paths in the signed graph $G$ connecting nodes in the same set are positive and all paths connecting a node in $A$ to a node in $B$ are negative.

For balanced semidirected graphs, we characterize correlations in Theorem 2.3.

**Theorem 2.3** For a balanced functional network $G \cup Y$, $G$ can be partitioned into two sets of variables $A$ and $B$ such that elements of $A$ are positively correlated, elements of $B$ are positively correlated and correlations between $A$ and $B$ are negative. All elements of the same set have correlations with $Y$ of the same sign, which is opposite for the two sets.

Theorem 2.3 implies that for module elements $x_i$ and $x_j$, the sign of $z_{ij} = \text{sign} \left( \text{Cov}(x_i, y) \times \text{Cov}(x_j, y) \right) = \text{sign} \left( \text{Cov}(x_i, x_j) \right)$. That is, if $x_i$ and $x_j$ reside in the same set, their correlation is positive and $\text{sign} \left( \text{Cov}(x_i, y) \right) = \text{sign} \left( \text{Cov}(x_j, y) \right)$. If $x_i$ and $x_j$ reside in different sets, their correlation is negative and $\text{sign} \left( \text{Cov}(x_i, y) \right) \neq \text{sign} \left( \text{Cov}(x_j, y) \right)$. Thus, we flatten the upper diagonal elements of $Z = (z_{ij})$ and $\Sigma$ into corresponding vectors of length $(n(n - 1)/2)$ and examine the agreement between the signs of the $z$-vector and the signs of the correlation vector. This can be viewed as the measurement of inter-rater reliability of the two signs. Cohen’s Kappa measures agreement while considering the possibility of chance. A Cohen’s Kappa value less than or equal to 0 corresponds to no agreement/disagreement, and values greater than 0 correspond to agreement with perfect agreement equal to 1. We use Cohen’s Kappa to judge whether the sparse selection has captured agreeable variables according to Theorem 2.3, and hence whether they form a balanced functional module.

We note that the theorem shows that in the case of genetic effects, promotors and suppressors are unambiguously classified by the sign of the simple correlation with outcome.

3 Results
3.1 Simulation
3.1.1 Simulation model
Our method applies to arbitrary module topologies. Our algorithm and its application are independent of the underlying network topology. It is just used to select a causally consistent set of relevant variables. We will simulate its use in the specific case a Single Input Module (SIM), which consists of a set of genes controlled by a single transcription factor (Milo et al., 2002). Experimental evidence shows that SIMs occur frequently (Lee et al., 2002, Milo et al., 2002). For example, consider an SIM represented by the linear model for gene expression

$$x_i = \pi_i \beta_0 + \epsilon_i, i = 1, \ldots, t$$

$$x_0 = \delta_0,$$

where $\beta > 0$, $\pi_i \in \{-1, 1\}$ and the $\epsilon_i, i = 0, 1, \ldots, t$ are independent errors with mean 0 and variance $\sigma^2$. The covariance of all pairs of genes in this system is nonzero. The covariation among the $t$ observed module genes is driven by a latent unobserved hub node $x_0$. We model the functional aspect of the pathway by letting the hub $x_0$ determine an outcome variable Y by the regression function

$$Y = x_0 \delta + \delta,$$

where $\delta > 0$ without loss of generality. Theoretical details on the covariance structure are described in the Supplementary Appendix.

Figure 2 summarizes the components of an SIM. Tracing paths (Wright, 1934) in Figure 2 show that an SIM is a balanced functional module. All predictor nodes positively correlated with $Y$ are positively correlated with each other, and similarly for predictor nodes that are negatively correlated with $Y$. The path connecting a pair of nodes is negative if they have differing correlation with $Y$. Consequently, this simple latent variable model commonly used to illustrate biological phenomenon is actually a balanced module.

3.1.2 Simulation settings
We simulated data containing 5 SIM modules with 20 nodes each and a set of 1000 irrelevant nodes, $\Delta$. Only one of the modules is functional as per (6), and the other four are uncorrelated with the outcome variable Y. We define signal-to-noise ratio (SNR) to be the ratio of $\sigma^2$ to $\sigma_n$, making SNR a ratio of the square root of the variance component in $Y$ attributed to the hub and the variance component in $Y$ attributed to the noise. We set model parameters to produce a specified SNR ($\text{SNR} \in \{3, 5\}$) and intramodular correlation ($r_m \in \{1, 2, 3, 4, 5, 6, 7\}$), for sample sizes of 100 and 400. We additionally included five independent nodes with SNR equivalent to the hub node, for a total of 1105 nodes. The five independent

![Fig. 1. The formation of $W = \text{Cov}(X) - \langle \text{Cov}(X, Y) \text{Cov}(X, Y)^T \rangle$. The leftmost matrix is $\text{Cov}(X)$, the middle matrix is $\text{Cov}(X, Y) \text{Cov}(X, Y)^T$ and the rightmost matrix is their element-wise product $W$](image-url)
nodes represent nonmodular influences on the outcome, which realistically may be present in data.

We compare our method (bFMD) to three others. The first (W) was given by Miecznikowski et al. (2016) and starts with the matrix W defined previously. They set all negative elements of the matrix W to zero since the negative elements will correspond to unbalanced node pairs. The transformed matrix will then be matrix of elements corresponding to theoretical zeros and a highly positive block. Since the highly positive block will have a maximal sparse positive eigenvector according to the Perron Theorem, they use it to identify the module variables. The theoretical eigenvector has $p$ positive elements and $T−p$ zeros, so they partition their approximate eigenvector elements into two clusters by k-means and take the cluster with a higher mean to be the module. This maneuver models sparsity only indirectly so their method does not require specifying a sparsity parameter.

We also contrast our approach with Weighted Correlation Network Analysis (WGCNA), which is a two-step strategy of cluster analysis (modularity) followed by evaluating the identified clusters for the average correlation of the cluster members with the outcome variable ( functionality; Langfelder and Horvath, 2008). Like bFMD and W, WGCNA obtains sets of interrelated variables, but in contrast, the functional information captured by the individual gene-outcome associations is not incorporated in the search for clusters at the first step. The basic approach can be implemented in a variety of ways by varying the similarity metric and the clustering method used, and it has been extended and highly developed (Horvath 2011; Zhang and Horvath, 2005) as part of a broad approach to network analysis. To implement it for our simulation study, we expressed similarity as absolute correlation and used partition around medoids (PAM) for the clustering method (Reynolds et al., 2006). The clustering methods were set to obtain six clusters: one around medoids (PAM) for the clustering method (Reynolds et al., 2006) finds components composed of genes that are predictive of the outcome. In the first step, SPC selects a set genes highly associated with the outcome. Principal components are then computed for just those genes, and the resulting components are used for prediction. The number of selected genes is tuned using cross-validation to minimize the estimated out-of-sample prediction error using the principal component of the selected genes as the predictor. The components extracted by SPC are intended for prediction and not module identification. However, it is reasonable to consider using SPC for functional module discovery in the specific case that a single functional module influences the outcome since the genes selected would include the module genes and the principal component would load on the mutually correlated module genes. In this case, SPC exemplifies a two-step method that considers association with the outcome in the first step and coexpression in the second step, making it the reverse order compared to the WGCNA approach. For SPC, we used the package ‘superpc’ version 1.09 (Bair and Tibshirani, 2004).

Four simulation scenarios were considered: sample size of 100 with same sign coefficients, sample size of 100 with random sign coefficients, sample size of 400 with same sign coefficients and sample size of 400 with random sign coefficients. In the simple case of coefficients with the same sign, we let $\gamma = 1$ in (4) for all module genes so that all intramodular correlations resulted in the same sign. For random signs, we randomly selected each $\gamma_i$ from $\{-1, 1\}$. Each simulation scenario was repeated 200 times and bFMD, W, WGCNA and SPC were applied and compared under each simulation.

3.1.3 Simulation results
We describe results for each method in terms of sensitivity, false discovery rate (FDR, where FDR was computed as 1 minus the average of the number of module variables discovered divided by the total number of variables selected), the distribution of detected module size and the Hamming distance between the detected module and the true module. The Hamming distance was normalized by the vector lengths so that 0 indicates perfect concordance and 100 indicates complete discordance.

Figures 3 and 4 were generated from a sample size of 100, $SNR = 0.3$, and had randomly signed edges. Figure 3 shows the sensitivity and FDR of the methods. The method W from Miecznikowski et al. (2016) selected the largest number of variables in the true module, followed by WGCNA and bFMD. The sensitivity of SPC plummets for higher intramodule correlations and bFMD by far results in the lowest FDR. The FDR for W is
the most similar to bFMD, but it remains higher for all correlation values. The unacceptable FDR for WGCNA and SPC barely improves with higher intramodule correlation.

Figure 4 compares the Hamming distances and the distributions of the module sizes found by the methods. bFMD has by far the closest distance from the true module while also achieving smaller module sizes. Also, we see that module size for bFMD is tightly distributed about the true module size of 20. This is analogous to the case of a statistical estimator yielding low bias and low variance. Since the Hamming distance reflects both false positives and false negatives, the superiority of bFMD suggests that false discoveries heavily impact the distance measure due to the high dimensionality. WGCNA remains relatively unimproved by increasing the intramodule correlation, and SPC outperformed W.

The performance of the methods improves when SNR is increased to 0.5 due to the stronger signal, but the relative patterns remain similar, with the exception that the Hamming distance for W is greatly reduced. The interquartile ranges of module sizes become bFMD (21, 25), W (21, 168), WGCNA (177, 191) and SPC (30, 317).

For increased sample size of 400 and SNR = 0.3, the characteristics of bFMD and W tend to converge. The sensitivity of bFMD matches that of W for intramodule correlations over 0.2, and the FDR and Hamming distance of W match that of bFMD for correlations greater than 0.3. The Hamming distance for bFMD remains superior for all intramodule correlations.

To assess other module sizes, we ran similar simulations with the size of all modules increased to 50 with random signs and SNR = 0.3. Aside from the improved sensitivity in WGCNA, the patterns are consistent as in the simulations with module sizes of 20. Additionally, conclusions remain unchanged when repeating all of the simulations with entirely positive edges.

In summary, bFMD and W are superior to WGCNA and SPC. WGCNA suffers by not taking the association with Y into account when calculating clusters, which tend to be too large. SPC will select variables associated with Y, but those which are not correlated with other predictors will be missed. Only bFMD and W simultaneously consider both clustering and prediction criteria. W is more sensitive than bFMD, although the difference diminishes greatly for larger intramodule correlations. This is due to W producing larger modules, thus decreasing the probability of missing an important variable. However, the higher sensitivity in W also increases its FDR compared to bFMD since W overselects irrelevant variables. The k-means partition used by W is a crude surrogate for sparsity, but W and bFMD tend to converge with increased sample size and SNR, i.e. easier problems, since the underlying data structure operated on by the two methods is similar. Investigators should use the W approach if they are adverse to missing any important variable, and they should use bFMD if they are interested in getting the most accurate representation of the module.

### 3.2 Example

The fourth leading cancer amongst women in the USA may be attributed to Uterine Corpus Endometrial Carcinoma (UCEC), according to Levine et al. (2013). In 2013, there were 50,000 new diagnosed UCEC cases and the cause of approximately 8000 deaths in the USA. UCEC was selected for assessment in the Pan-Cancer project within The Cancer Genome Atlas (TCGA) since achieving a sufficient sample size would be made easy due to its high incidence.
Mutation, copy number, gene expression, DNA methylation, MicroRNA and reverse-phase protein array were obtained from tumor samples across 545 donors, and the data were made publicly available in the TCGA database (Weinstein et al., 2013).

For this example, we will focus on discovering the subset of genes from the gene expression dataset \( (X) \) that form a balanced module relating to percent tumor invasion \( (Y) \). The entire gene expression dataset within the TCGA-UCEC project contains 56,457 variables. To help reduce the amount of data in \( X \), we will leverage prior research that shows mutations in MSH1, MSH2, MSH6 and PMS2 have been linked to Lynch Syndrome and ultimately Endometrial Cancers (Kempers et al., 2011). Since MSH2 and MSH6 both reside on chromosome 2 and have been shown to be associated with one another (Schweizer et al., 2001), we will focus on chromosome 2 for the analysis. Percent tumor invasion was selected as the outcome for this example since it associates with tumor grade and patient survival (Boronow et al., 1984).

Ultimately, the analytical set resulted in 469 subjects, each with a percent tumor invasion outcome and 3811 genes to consider for the functional module. In notational form, we now have a matrix \( X \) with 469 rows and 3811 variables (e.g. genes) and an outcome vector \( Y \) of length 469 (e.g. percent tumor invasion).

Analyses were done using bFMD, W, WGCNA and SPC. bFMD took the longest to execute at 20.1 min as compared to 1.8, 0.2 and 1.8 min for W, WGCNA and SPC, respectively. We describe the connectivity of estimated networks by the average correlation between module members, and the effect of the module on the outcome by the average correlation of module members with \( Y \). For WGCNA, we estimated that there are six clusters by inspecting the eigenvalues of the correlation matrix (Fallah et al., 2008). Table 1 summarizes the results, and we note that bFMD outperforms the other methods in every metric. While achieving the largest average module correlation and module effect, bFMD also returns the smallest module size, indicating a more tightly connected network. The value of Kappa for bFMD is 0.961, suggesting that the sparse eigenvector algorithm has produced a highly balanced module.

Table 2 shows the normalized Hamming distances between the variable selection vectors of the four methods. WGCNA produced the most different results compared to the other methods, which are likely attributed to its higher module size. bFMD produced the most similarly to SPC, likely due to the smaller module size from those two methods.

The 187 features selected by bFMD included the MSH2 and MSH6 genes, which provide biological relevancy to the results. The W and WGCNA methods also selected the MSH2 and MSH6 genes, but they also selected many more genes which likely included noise due to the decreased module correlation and module effect. Despite the second smallest module size, moderate module correlation and module effects, SPC failed to select the MSH2 and MSH6 genes, weakening the biological interpretation of its selections. Figure 5 displays the selected module genes that are most negatively and positively correlated with MSH2 and MSH6.

From Figure 5, we see that MSH2 and MSH6 positively correlate with each other and with % Tumor Invasion, suggesting they are tumor promoters for UCEC. Conversely, we see that the selected gene TEKT4 was found to be negatively correlated with MSH2.

### Table 1. TCGA-UCEC Transcription data: properties of estimated networks

| Method  | Module size | Module correlation | Module effect |
|---------|-------------|--------------------|---------------|
| bFMD    | 187         | 0.23               | 0.10          |
| W       | 680         | 0.19               | 0.08          |
| WGCNA   | 795         | 0.05               | 0.03          |
| SPC     | 248         | 0.12               | 0.09          |

### Table 2. TCGA-UCEC Transcription data: normalized Hamming distance between network variable selection vectors of bFMD, W, WGCNA and SPC

| Method  | bFMD | W     | WGCNA | SPC  |
|---------|------|-------|-------|------|
| bFMD    |      | 6.5   | 8.3   | 5.6  |
| W       | 12.4 |       | 9.4   |      |
| WGCNA   | 13.0 |       |       |      |
MSH6 and % Tumor Invasion suggesting that TEKT4 is a tumor suppressor. This is consistent with prior research showing that TEKT4 loss relates to breast cancer metastasis (Ge et al., 2021). The BOK gene has been shown to be a tumor suppressor for nonsmall-cell lung carcinoma (Moravcikova et al., 2017), and it was selected by bFMD as negatively correlated with MSH2, MSH6 and % Tumor Invasion which also suggests it is a tumor suppressor for UCEC.

bFMD selected a much narrower set of genes as compared to the W method. To compare the gene ontology between these methods, we input the selections into DAVID (Huang et al., 2009; Sherman et al., 2009). Only 8 of the 187 genes selected by bFMD were unannotated within DAVID as compared to the 286 unannotated genes of the 680 selected by W. We clustered the annotated gene selections using DAVID’s Gene Functional Classification Tool which clusters genes based on similar gene-to-gene functional annotations. The goal of DAVID clustering is to combine biologically similar genes into a cluster. From this, DAVID identified 4 clusters across 18 genes from the selections by bFMD as compared to 7 clusters across 53 genes from the selections by W. Sixteen genes from the bFMD genes from the selections by bFMD as compared to 7 clusters across 18 genes from the selections by bFMD were also contained within the W clusters. Additionally, the 4 clusters within bFMD corresponded to the top 4 clusters found within W where the ranking is according to enrichment score, a measure of cluster quality. In summary, bFMD identified a more biologically succinct set of genes as designed by the balance criteria.

4 Summary and discussion

The elements of an efficient biological mechanism may be expected to exhibit balanced effects which complement each other. The methods bFMD and W both operate on a matrix that highlights balanced sets of variables affecting an outcome variable as a positive submatrix. Of the two, bFMD most accurately identifies the set of module variables, as measured by the Hamming distance. Once the variables are extracted from a high-dimensional superset, detailed analysis of individual associations can be performed using existing methods for graphical model analysis (e.g. Edwards, 2000; Werneth, 2012; Werneth and Cox, 2013; Whittaker, 2009).

Our method is appropriate when linear relationships represent relevant aspects of the data. The theorems in this paper rely on the assumption of moderate conditional independence among variables and should be considered approximate for typical data sets. As with many statistical methods, strong multicollinearity potentially degrades performance. Future work could explore using a balance penalty in likelihood methods. For example, Net-Cox uses a predictability matrix with application to microarray data. Stat. Appl. Genet. Mol. Biol., 7, Article24.

Ge,L.-P. et al. (2021) Tektn4 loss promotes triple-negative breast cancer metastasis through HDAC6-mediated tubulin deacetylation and increases sensitivity to HDAC6 inhibitor. Oncogene, 40, 2323–2334.

Harary,F. (1953) On the notion of balance of a signed graph. Michigan Math. J., 2, 143–146.

Horn, V.S. (2011). Weighted Network Analysis. Springer, New York.

Huang,D. et al. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res., 37, 1–13.

Kempers,M.J. et al. (2011) Risk of colorectal and endometrial cancers in EPICAM deletion-positive Lynch syndrome: a cohort study. Lancet Oncol., 12, 49–53.

Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9, 559.

Lee,T. et al. (2002) Transcriptional regulatory networks in Saccharomyces Cerevisiae. Science, 298, 799–804.

Levine,D.A. et al.; Cancer Genome Atlas Research Network. (2013) Integrated genomic characterization of endometrial carcinoma. Nature, 497, 67–73.

Mackey,L. (2009). Deflation methods for sparse PCA. In: Koller,D. et al., editors, Advances in Neural Information Processing Systems 21. Curran Associates, Inc, pp. 1017–1024. San Diego, CA.

Meyer,C. (2000). Matrix Analysis and Applied Linear Algebra. Siam, Philadelphia, PA.

Miecznikowski,J.C. et al. (2016) Identification of consistent functional genetic modules. Stat. Appl. Genet. Mol. Biol., 15, 1–18.

Milo,R. et al. (2002) Network motifs: simple building blocks of complex networks. Science, 298, 824–827.

Moravcikova,E. et al. (2017) Bok displays cell death-independent tumor suppressor activity in non-small-cell lung carcinoma. Int. J. Cancer, 141, 2050–2061.

Reynolds, A.P. et al. (2006) Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. J. Math. Modell. Algorithms, 5, 473–504.

Schweizer,P. et al. (2001) Lack of MSH2 and MSH6 characterizes endometrial but not colon carcinomas in hereditary nonpolyposis colorectal cancer. Cancer Res., 61, 2813–2817.

Sherman, R.T. et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc., 4, 44–57.

Funding

This work was supported by the New York State Department of Health [Empire Clinical Research Investigator Program to L.M.T.-M.].

Conflict of Interest

None declared.

References

Anufriev,M. and Panwenko,V. (2015) Connecting the dots: econometric methods for uncovering networks with an application to the Australian Financial Institutions. J. Banking Finance, 61, S231–S235.

Bair,E. and Tibshirani,R. (2004) Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol., 2, e108.

Bair,E. et al. (2006) Prediction by supervised principal components. J. Am. Stat. Assoc., 101, 119–137.

Bianconi,G. (2018). Multilayer Networks: Structure and Function. Oxford University Press, Oxford, United Kingdom.

Boronow,R. et al. (1984) Surgical staging in endometrial cancer: clinical-pathologic findings of a prospective study. Obstet. Gynecol., 63, 825–832.

Bourgon,R. et al. (2010) Independent filtering increases detection power for high-throughput experiments. Proc. Natl. Acad. Sci. USA, 107, 9546–9551.

Chartrand,G. (1977). Introductory Graph Theory. Dover, New York.

Edwards,D. (2000). Introduction to Graphical Modeling. 2nd edn. Springer, New York.

Edwards,D. et al. (2010) Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. BMC Bioinformatics, 11, 18.

Fallah,S. et al. (2008) Estimating number of clusters based on a general similarity matrix with application to microarray data. Stat. Appl. Genet. Mol. Biol., 7, Article24.

Ge,L.-P. et al. (2021) Tektn4 loss promotes triple-negative breast cancer metastasis through HDAC6-mediated tubulin deacetylation and increases sensitivity to HDAC6 inhibitor. Oncogene, 40, 2323–2334.

Harary,F. (1953) On the notion of balance of a signed graph. Michigan Math. J., 2, 143–146.

Horn, V.S. (2011). Weighted Network Analysis. Springer, New York.

Huang,D. et al. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res., 37, 1–13.

Kempers,M.J. et al. (2011) Risk of colorectal and endometrial cancers in EPICAM deletion-positive Lynch syndrome: a cohort study. Lancet Oncol., 12, 49–53.

Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9, 559.

Lee,T. et al. (2002) Transcriptional regulatory networks in Saccharomyces Cerevisiae. Science, 298, 799–804.

Levine,D.A. et al.; Cancer Genome Atlas Research Network. (2013) Integrated genomic characterization of endometrial carcinoma. Nature, 497, 67–73.

Mackey,L. (2009). Deflation methods for sparse PCA. In: Koller,D. et al., editors, Advances in Neural Information Processing Systems 21. Curran Associates, Inc, pp. 1017–1024. San Diego, CA.

Meyer,C. (2000). Matrix Analysis and Applied Linear Algebra. Siam, Philadelphia, PA.

Miecznikowski,J.C. et al. (2016) Identification of consistent functional genetic modules. Stat. Appl. Genet. Mol. Biol., 15, 1–18.

Milo,R. et al. (2002) Network motifs: simple building blocks of complex networks. Science, 298, 824–827.

Moravcikova,E. et al. (2017) Bok displays cell death-independent tumor suppressor activity in non-small-cell lung carcinoma. Int. J. Cancer, 141, 2050–2061.

Reynolds, A.P. et al. (2006) Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. J. Math. Modell. Algorithms, 5, 473–504.

Schweizer,P. et al. (2001) Lack of MSH2 and MSH6 characterizes endometrial but not colon carcinomas in hereditary nonpolyposis colorectal cancer. Cancer Res., 61, 2813–2817.

Sherman, R.T. et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc., 4, 44–57.

Acknowledgements

The author wishes to thank Dr. Xiwei Chen for code used for the simulations for this paper.

Software and data availability

The code to simulate functional modules and perform the bFMD method may be found on Github at https://github.com/orinimm/bFMD. The data from the Example section may be downloaded from the TCGA-UCEC project.
Sigg,C.D. and Buhmann,J.M. (2008). Expectation-maximization for sparse and non-negative PCA. In Proceedings of the 25th International Conference on Machine Learning. ACM, New York, NY, pp. 960–967.

Sulaimanov,N. and Koepp,H. (2016) Graph reconstruction using covariance-based methods. EURASIP J. Bioinformatics Syst. Biol., 2016, 19–20.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B, 58, 267–288.

Towle-Miller,L.M. et al. (2021) SuMO-Fil: Supervised multi-omic fildasdsfring prior to performing network analysis. Plos One., 16.8, e0255579.

Weinstein,J.N. et al.; Cancer Genome Atlas Research Network. (2013) The Cancer Genome Atlas Pan-cancer Analysis Project. Nature Genetics, 45, 1113–1120.

Wermuth,N. (1980) Linear recursive equations, covariance selection, and path analysis. J. Am. Stat. Assoc., 75, 963–972.

Wermuth,N. (2012) Traceable regressions. Int. Stat. Rev., 80, 415–438.

Wermuth,N. and Cox,D. (2013). Concepts and a case study for a flexible class of graphical Markov models. Robustness and Complex Data Structures. Springer, Berlin, Heidelberg, pp. 331–350.

Whitaker,J. (2009). Graphical Models in Applied Multivariate Statistics. Wiley, New York, NY.

Wright,S. (1934) The method of path coefficients. Ann. Math. Stat., 5, 61–215.

Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol., 4, Article17.

Zhang,W. et al. (2013) Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. PLoS Comput. Biol., 9, e1002975.