

RESEARCH ARTICLE

Biclustering by sparse canonical correlation analysis

Harold Pimentel¹, Zhiyue Hu² and Haiyan Huang^{2,*}

¹ Department of Computer Science, University of California, Berkeley, CA 94720, USA

² Department of Statistics, University of California, Berkeley, CA 94720, USA

* Correspondence: hhuang@stat.berkeley.edu

Received July 13, 2017; Revised September 20, 2017; Accepted September 20, 2017

Background: Developing appropriate computational tools to distill biological insights from large-scale gene expression data has been an important part of systems biology. Considering that gene relationships may change or only exist in a subset of collected samples, biclustering that involves clustering both genes and samples has become increasingly important, especially when the samples are pooled from a wide range of experimental conditions.

Methods: In this paper, we introduce a new biclustering algorithm to find subsets of genomic expression features (EFs) (e.g., genes, isoforms, exon inclusion) that show strong “group interactions” under certain subsets of samples. Group interactions are defined by strong partial correlations, or equivalently, conditional dependencies between EFs after removing the influences of a set of other functionally related EFs. Our new biclustering method, named SCCA-BC, extends an existing method for group interaction inference, which is based on sparse canonical correlation analysis (SCCA) coupled with repeated random partitioning of the gene expression data set.

Results: SCCA-BC gives sensible results on real data sets and outperforms most existing methods in simulations. Software is available at <https://github.com/pimentel/scca-bc>.

Conclusions: SCCA-BC seems to work in numerous conditions and the results seem promising for future extensions. SCCA-BC has the ability to find different types of bicluster patterns, and it is especially advantageous in identifying a bicluster whose elements share the same progressive and multivariate normal distribution with a dense covariance matrix.

Keywords: biclustering; SCCA; gene clusters

Author summary: In this paper, we introduce a novel biclustering algorithm to find subsets of genomic expression features (EFs) (e.g., genes, isoforms, exon inclusion) that show strong partial correlations (i.e., conditional dependencies between EFs after removing the influences of other EFs in the same set) under certain subsets of samples. We extend an existing method for inferring such conditional dependencies, which is based on sparse canonical correlation analysis (SCCA) coupled with repeated random partitioning and subsampling of the gene expression data set. We incorporate a binary vector such that it will assist the objective function on deciding exclusion or inclusion of a particular sample to the bicluster. We test our algorithm on both simulation and real datasets, and achieve promising results. In addition, our algorithm is shown to be relatively robust to initialization and small perturbation in hyper-parameters. The algorithm is available at <https://github.com/pimentel/scca-bc>.

INTRODUCTION

The past two decades have evidenced the explosion of gene expression data from high-throughput technologies such as microarray and RNA sequencing. With the rapid

increase of expression databases, distilling biological insights from large-scale gene expression data has been an important part of biomedical research. A popular analysis is to find sets of genomic expression features (EFs)

(genes, isoforms, exon inclusion, etc.) where expression behaves similarly among all experimental conditions. Historically, one-way clustering has been used for this type of analysis [1]. As the number of samples grows accompanied by increasing heterogeneity, it is practical to assume that some EF interactions do not persist, or remain in the same direction (e.g., positive *vs.* negative regulation) across all samples. In such situations, one-way clustering can be insufficient as it will likely miss those EFs as a cluster. As a result, two-way clustering or biclustering methods that find subsets of EFs which behave in some related way under certain subsets of samples/conditions have been desired and hence developed.

Many biclustering methods decompose the expression into an additive or multiplicative structure. Let $W = (w_{ij}) \in \mathbb{R}^{n \times k}$ be the gene expression matrix with EFs on the rows and experimental conditions/samples on the columns. In an additive bicluster model, w_{ij} is assumed to be decomposable into $\mu + \alpha_i + \beta_j$ where μ denotes a constant base value, and α_i and β_j denote a constant row and column value respectively [2]. In a multiplicative bicluster model, $w_{ij} = \mu \times \alpha_i \times \beta_j$, which is equivalent to an additive model after taking log on the data [3].

Linear bicluster models generalize additive and multiplicative models by modeling linear relationships between rows (EFs) under a subset of columns (samples), decomposing w_{ij} into $\mu_i + \alpha_i c_j$. Several linear approaches use Pearson correlation as the similarity metric, but employ different searching algorithms for finding biclusters [4,5]. However, Pearson correlation can often lead to false-positives resulting from lowly expressed features [6]. To address this issue Gao *et al.* [6] utilize a fitness function which incorporates Pearson correlation along with a variance measure.

Coherent evolutions models find a permutation of samples such that every feature is being down-regulated or up-regulated in the same manner. For example, features which follow $W_{i\pi(1)} \leq W_{i\pi(2)} \leq \dots \leq W_{i\pi(c)}$, where $\pi(j)$ is the index with j -th rank. These methods, such as OSPM and OP-cluster [7,8], find the largest submatrix that follows these constraints and is statistically significant.

In this paper we introduce a new algorithm, named SCCA-BC, to find biclusters with strong multivariate normal group interactions. As in Ref. [9], we elucidate the group interactions by evaluating conditional dependencies or partial correlations between EFs, i.e., the relationships between EFs after removing the influences of a set of other functionally related EFs. Capturing such group interactions is critical for seeking cooperative and interacting EFs in a biological pathway. Traditional pairwise interaction measures such as Pearson correlation would likely miss these patterns since EFs that are strongly conditionally dependent could have very weak

pairwise marginal correlations. In particular, SCCA-BC extends the approach in Ref. [9] by allowing evaluating group interactions under a subset of samples. Tan and Witten [10] proposed a related but simpler model where the matrix elements are normally distributed with a bicluster-specific mean term and a common variance, and perform biclustering by maximizing the corresponding log likelihood.

RESULTS

Simulation studies

We compared our method (SCCA-BC) to four other popular methods with R packages: an extended version of the Plaid model in Ref. [2], introduced as IP in Ref. [11]; sparseBC in Ref. [10]; matrixBC in Ref. [10]; SSVD in Ref. [12]. We used mostly the default parameters for each method with the following exceptions. For IP, we disabled overlapping clusters. For sparseBC, we always chose the sparse parameters using BIC, fitting parameters from 0 to 40. For matrixBC, we ran it with two possible regularization parameters (0, 10). For SSVD, we increased the maximum number of iterations from 100 to 500. We always fixed the number of clusters to be the true number clusters.

We measured the performance by a modified version of the average module recovery score (ARMS), as introduced in Ref. [13] called extended AMRS (eAMRS). This metric is effectively the Jaccard index on matrix cells. Let V denote the set of matrix cells which belong to a true bicluster. Each cell can be thought of as a tuple of row and column, such as (row #, column #). Thus, $V = \{(r_i, c_i)\}_{i=1}^t$ where r_i and c_i denote the row and columns of cell i and there are t total cells in bicluster V . We similarly define V_{est} as the estimated set of cells that belong to a bicluster. M being the set of all cells contained in the true bicluster while M_{est} being the estimated bicluster; then eAMRS is defined precisely:

$$\text{eAMRS}(M, M_{\text{est}}) = \frac{1}{|M|} \sum_{V \in M} \max_{\{V_{\text{est}} \in M_{\text{est}}\}} \frac{|V_{\text{est}} \cap V|}{|V_{\text{est}} \cup V|}.$$

The eAMRS is bounded between 0 and 1 with a score of 1 being perfect recovery when M is nonempty.

We created four simulation datasets with different bicluster models:

1. Two biclusters: additive and random walk. The additive bicluster is generated from the model $w_{ij} = \mu_i + v_j + \varepsilon_{ij}$ with μ_i and v_j being randomly selected from $\{0, 1, \dots, 5\}$ if EF i and sample j are in the bicluster, and $\mu_i = v_j = 0$ otherwise. ε_{ij} 's are iid following $N(0, 1)$. The bicluster contains 500 EFs and 15 samples.

Next, we describe the random walk bicluster. To

simplify notation, we discuss only one feature in the bicluster. The walk starts at $B = \sum_{i=1}^3 B_i$ where $B_i \sim N(0, 1)$. Then, for each condition c in the bicluster, the expression in condition $c = \{1, 2, \dots, C'\}$ is $W_c = B + \sum_{j=1}^c X_j$ where $X_j \sim N(0, 1)$. This procedure is repeated for each EF in the bicluster. The biclusters generated were of 40 EFs and 15 samples. The total size of the expression matrix is 500 EFs and 40 samples.

2. Moderate correlation multivariate normal bicluster. Each condition is distributed $W_{\cdot j} \sim N(0, \Sigma)$. First, we constructed covariance matrix Σ by simulating $\sigma_{ij}^* \sim \text{Unif}(l, u)$ and $\sigma_{ii}^* = 1$ so that $\rho_{ij}^* = \sigma_{ij}^*$. As the dimension of Σ grows, this procedure is unlikely to produce a positive semi-definite matrix. To correct for this, we found the “nearest” positive-definite matrix [14], which we call Σ^C . We then computed the correlation matrix ρ^C from Σ^C and set $\Sigma = \rho^C$. In this simulation, we choose (l, u) as $(0.5, 0.8)$, which produces a reasonably wide range in correlation. Each bicluster contained 300 EFs and 30 samples. The total size of expression matrix is 1,500 EFs and 100 samples.

3. Higher correlation multivariate normal bicluster. We choose (l, u) as $(0.72, 0.94)$, which produces higher correlation than the previous simulation.

Each simulation was replicated 50 times initialized with different seeds. For any EFs not in a bicluster, expression was generated from a standard normal distribution. The eAMRS of our method and competing methods can be seen in Figures 1–2.

For the additive bicluster simulation (Figure 1A), SCCA-BC and IP perform similarly with sparseBC and matrixBC close behind. SSVD recovered much less of the

bicluster on average. For the random walk bicluster simulation (Figure 1B), we see that our method typically outperforms the others, with SSVD and IP doing similarly on average. SparseBC and matrixBC do slightly better than the other methods, likely due to the starting base values creating a mean in the bicluster deviating from zero.

For both multivariate normal bicluster (Figure 2), IP rarely recovers any correct biclusters as expected. This is because it models each bicluster as a mean, and the mean of the bicluster in this simulation is exactly the same as that of the background distribution. For almost every simulation IP reported “No cluster found”. SSVD recovers more biclusters on average, but our method find even more on average.

Note that in order to balance the computational expense between our methods and other methods, we ran our methods with only 100 random partitions per simulation, hence resulting in larger variance. If variance is a concern during practice, increasing the number of random partitions shall result in less variance.

modENCODE developmental RNA-Seq data

We analyzed the modENCODE fly and worm developmental RNA-Seq data set also analyzed in Ref. [6], in which the details regarding the production of data can be found. The data consists of 44 RNA-Seq experiments in total, 30 in *Drosophila melanogaster* (fly) and 14 in *Caenorhabditis elegans* (worm).

We ran our method using 100 random partitions, with 60% subsampling per partition. Based on the CLiP paper, we set our conditions cluster size to be 27 and 30. We

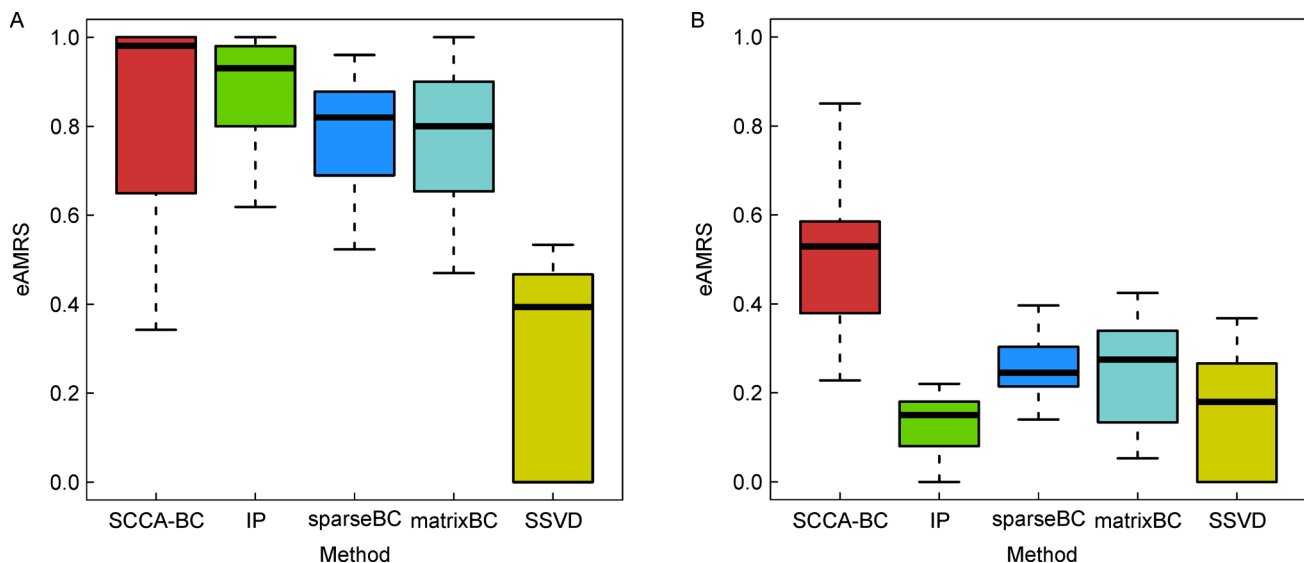


Figure 1. Box plot of eAMRS values for additive and random walk biclusters. (A) Box plot of eAMRS values for additive bicluster under Simulation 1. (B) Box plot of eAMRS values for random walk bicluster under Simulation 2.

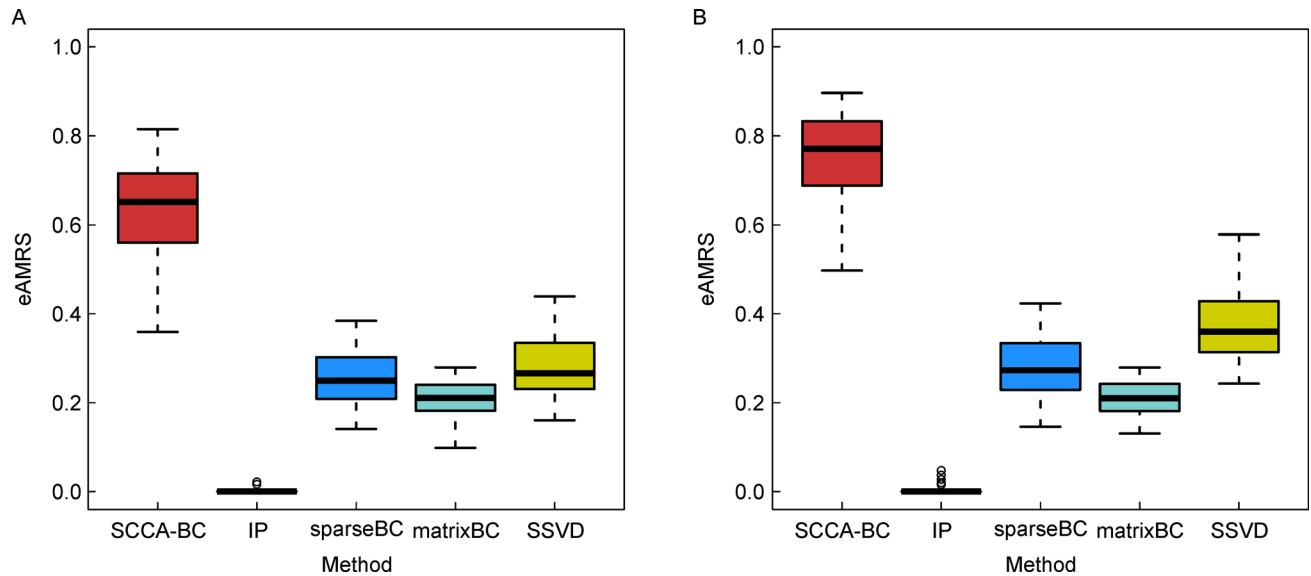


Figure 2. Box plot of eAMRS values for multivariate normal biclusters. (A) Box plots of the eAMRS for moderate correlation multivariate normal bicluster under Simulation 3. (B) Box plots of the eAMRS for higher correlation multivariate normal bicluster under Simulation 4.

present the bicluster with condition size 30 in Figure 3 as we feel it is more easily interpretable. Genes that showed coefficients in the top 90%, 90% of the time they were sampled were chosen as part of the bicluster. This restriction resulted in 196 genes in 30 conditions. Much like the CLiP paper, our method included all the fly embryo conditions, as well as all but one worm embryo condition (late embryo). Upon further inspection, this set of genes for worm late embryo showed a different expression pattern, which explains why it was not included. This bicluster shows that the genes are expressed highly in early embryo and female stages, while they seem to be lowly expressed in later embryo stages and male stages. These results are consistent with the CLiP analysis in Ref. [6], as well as previous studies on maternal effects in early embryo development in Refs. [15, 16].

modENCODE *D. melanogaster* perturbations RNA-Seq data

We also analyzed the modENCODE *D. melanogaster* perturbation data set described in Ref. [17]. The original data set contains 85 conditions with perturbations such as heat shock, cold shock, caffeine feeding, copper feeding, cadmium feeding, and others. We removed the larvae conditions which leaves us with 69 conditions performed in adult flies.

We performed some data cleaning to avoid biclusters largely containing zero values. First, we selected all genes that showed differential expression in at least one condition at or below a false-discovery rate of 30% as

defined in Ref. [17]. From this list, we then removed all genes that did not show expression values greater than zero in at least 90% of their conditions. This filtering resulted in a final list of 2,175 genes. Finally, we standardized each gene's expression as the distributions were very skewed.

We ran SCCA-BC with 100 random partitions and 60% subsampling per partition. We ran maximum condition sizes of 20, 25, and 30 and settled on 30 as qualitatively it selected more conditions and still maintained a high level of correlation. Genes which showed coefficient values in the top 75%, 90% of the time were included into the cluster, and the hierarchical clustering aggregation method was used to select conditions, which is presented in Figure 4.

The bicluster selected all samples enriched for neural tissue, including adult heads and dissected central nervous systems from larvae and pupae. Notably, embryonic samples were also selected, but those were exclusively the 16–24 h samples, after the structural differentiation of the nervous system. These results allow for further interpretation of recent findings on the same dataset reported in Ref. [17], namely, the observation that the vast majority of transcriptionally complex genes generate numerous discrete transcript isoforms exclusively in the developing and adult nervous system, and that transcript isoforms are least diverse in the gonads. Many of the selected genes in the bicluster correspond to RNA binding proteins expressed at medium to high levels in gonadal tissue and low levels in neural tissues. Hence, we observed disparate post-transcriptional regulatory patterns and inverse correlation in the expression of genes encoding

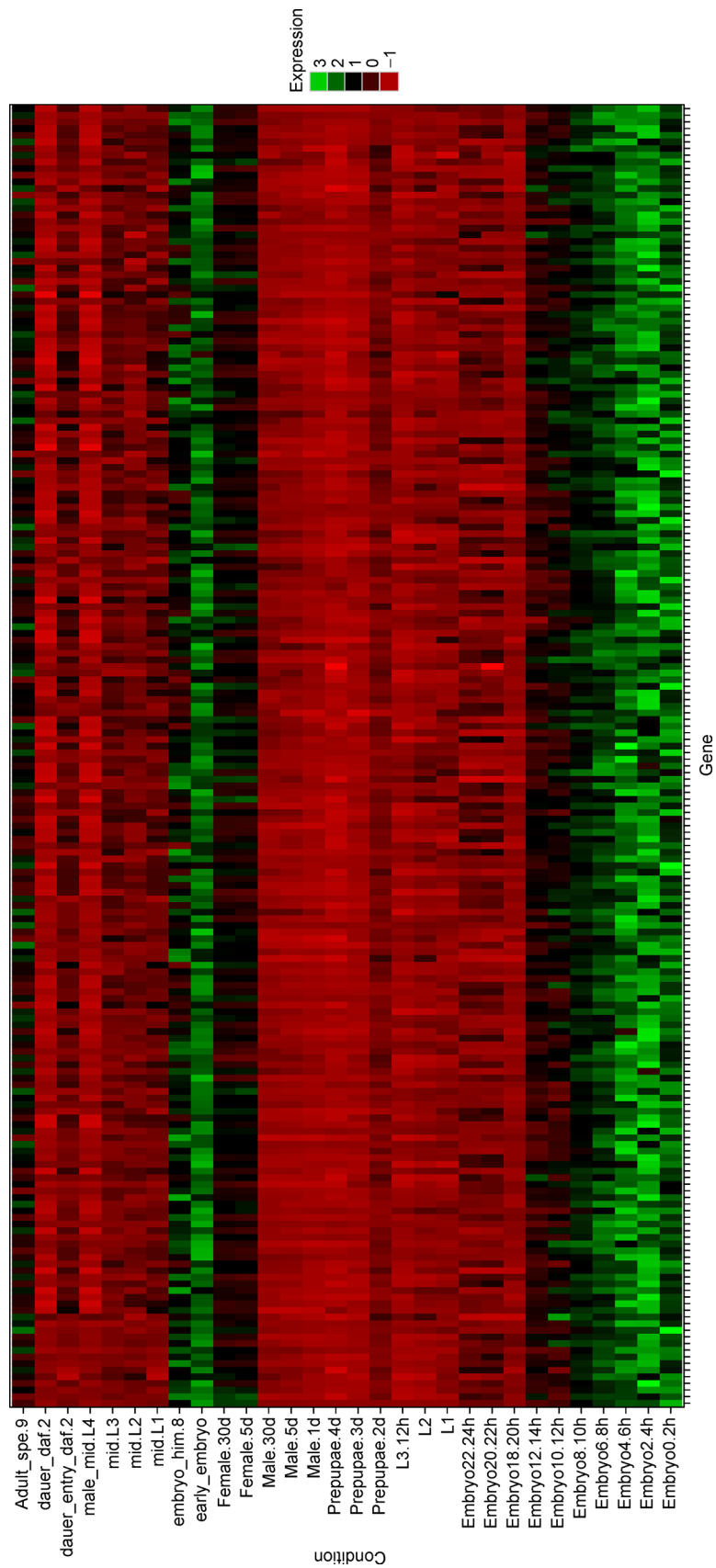


Figure 3. Most highly correlated bicluster found in the fly/worm data set. The majority of fly and worm embryo time stages are included in the bicluster. The overall pattern of each gene is also relatively consistent through each condition.

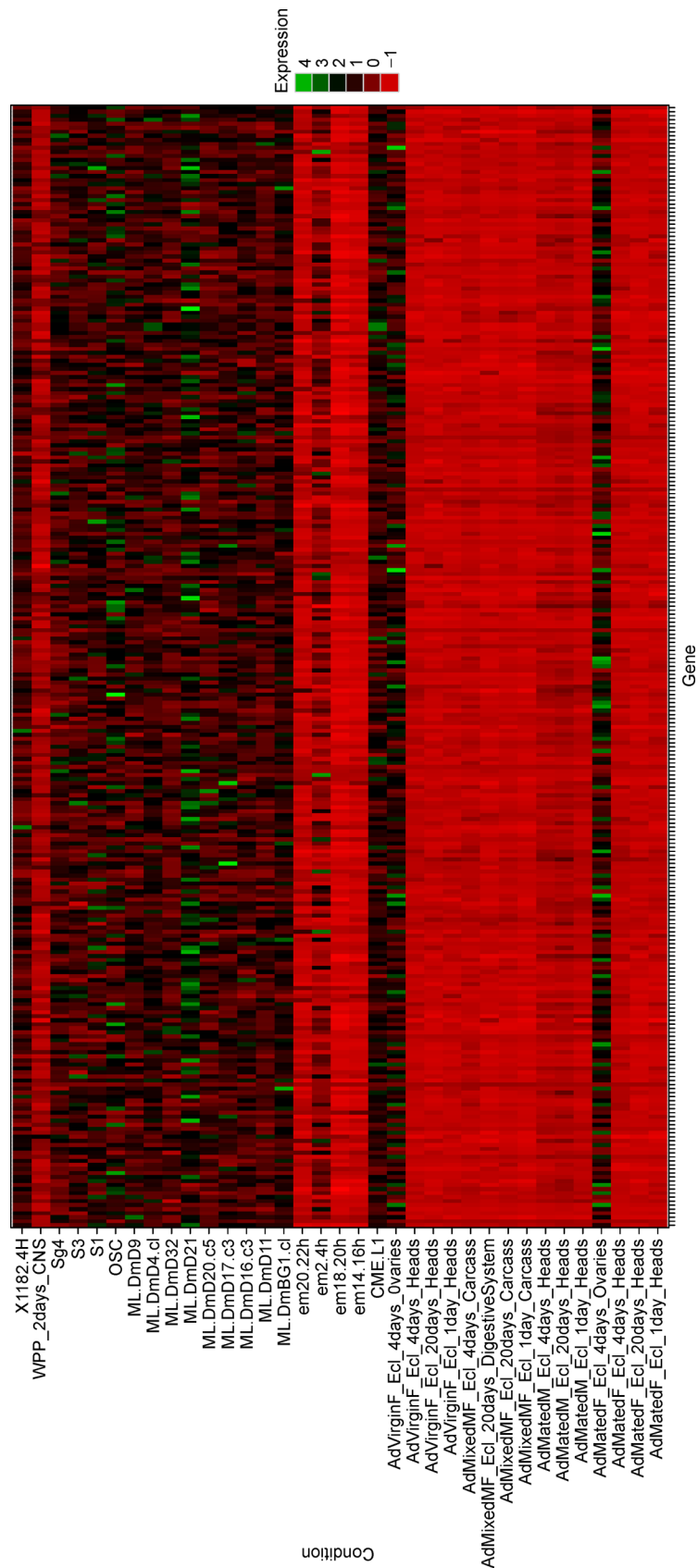


Figure 4. Most highly correlated bicluster found in the fly perturbation data set.

mRNA binding proteins between these tissue types. Indeed, a GO analysis using DAVID reveals that one of the most strongly enriched biological process GO terms (level 5 only) is mRNA Metabolism ($\text{FDR} \approx 4.2\text{e-}10$), and followed by numerous reproductive terms in an annotation cluster such as oogenesis, female gamete generation, sexual reproduction, and others ($\text{FDR} < 0.05$). These results suggest cohorts of RNA binding proteins that may lie at the mechanistic basis of differential RNA processing in gonadal and neuronal tissue, and hence motivate future functional studies.

DISCUSSION

We have presented a new method for biclustering, SCCA-BC, an application of sparse canonical correlation analysis (SCCA). As presented in our simulations, the method has the ability to find numerous types of biclusters, including, but not limited to: constant values, additive values, linearly correlated, and random walk. We found that SCCA-BC performs relatively better than other common bicluster algorithms when the underlying distribution is progressive and multivariate normal.

There is also one useful step in our method — the matrices of coefficients generated from optimizing parameters over the feature and condition parameters, which we presented one method of aggregation in this paper, k -means clustering on F and C separately, though, there are likely other ways to extract the biclusters from the aggregation matrix.

Being a randomized algorithm, SCCA-BC will have some variances in its solution, unlike the deterministic algorithms. In some situations this may be undesirable. However, if the data is not too noisy, running the algorithm several times should provide stable solutions.

Additionally, as the algorithm depends on random partitions and solving a large optimization problem, runtime is relatively slow compared to other methods. Other tools commonly can complete a run with a few thousand genes in less than a minute, whereas SCCA-BC takes on the order of minutes. While this runtime is relatively longer, it is not prohibitively long provided that the user restricts themselves to less than a few thousand genes. Additionally, SCCA-BC partitions are random parallelized so the runtime scales with the number of processors.

The method seems to work in numerous conditions and the results seem promising for future extensions. Automation of the condition cluster size is the most pressing current issue. In practice the user will likely have insight into how many samples might be related. In addition, exploring different types of regularization on the condition cluster matrix what allow for different types of clusters to be found (e.g., a subset of a timeseries). Finally,

while in practice convergence has not been an issue, theoretical bounds on convergence would be helpful.

METHODS

Method overview

Wang *et al.* [9] introduced a method for estimating group interactions using SCCA coupled with repeated random partition and subsampling of the EF expression dataset. By separating the EFs into two groups, Wang *et al.* [9] used SCCA to search for meaningful linear group relationships which, reframed in a linear regression setting, gives estimates proportional to partial correlations conditioned on different sets of signal EFs (with noisy EFs eliminated through sparsity). The subsampling procedure employed in the method allows for the discovery of multiple interacting groups simultaneously by stepping through subsets of the EFs with varying signal strengths. The final interaction measure is obtained by averaging the results from all the random partitions and subsamples, and thus provides an aggregated estimation of partial correlations of different orders. The major advantage of this approach is that it is capable of evaluating conditional dependencies when the correct dependent sets are unknown or only partially known.

SCCA-BC extends the above approach to allow capturing biclusters with strong EF group interactions. This extension is achieved through incorporating an indicator vector for samples in the objective function that is related to SCCA. The elements of the vector are binary, with 0 and 1 indicating exclusion or inclusion of a particular sample to the bicluster, respectively.

Technical details of SCCA-BC

Let $X \in R^{n_1 \times k}$ be a matrix comprised of k observations on n_1 variables, and $Y \in R^{n_2 \times k}$ a matrix comprised of k observations on n_2 variables. Assuming the rows of X and Y have been centered and scaled, the objective of SCCA is then to find sparse $\mathbf{a} \in R^{n_1}$ and $\mathbf{b} \in R^{n_2}$ that maximize $\mathbf{a}^T X Y^T \mathbf{b}$ (see Ref. [18] and the Section of Appendix for more details).

We extend the above SCCA model with the introduction of a diagonal matrix which determines inclusion of a sample to a bicluster. For dimension denoting convenience, we denote our matrix with centralized and scaled rows as $W \in R^{2n \times k}$ (i.e., expression matrix for $2n$ EFs in k samples), though the number of EFs under consideration does not necessarily have to be even. To find one bicluster, first, we randomly partition our matrix $W \in R^{2n \times k}$ by EFs into two matrices of equal (or similar) size: $X \in R^{n \times k}$, $Y \in R^{n \times k}$. We define matrix $D \in R^{k \times k}$ as a

diagonal matrix with diagonal vector \mathbf{d} . The elements in \mathbf{d} (denoted by $d_i, i = 1, \dots, k$) are assumed to be a relaxation of binary with 1 denoting absolute inclusion of the particular sample to the bicluster and 0 denoting absolute exclusion. SCCA-BC then solves the following optimization problem for each partition of \mathbf{W} by rows (EFs):

$$\underset{\mathbf{a}, \mathbf{b}, \mathbf{D}}{\text{maximize}} \quad \mathbf{a}^T \mathbf{X} \mathbf{D} \cdot (\mathbf{b}^T \mathbf{Y} \mathbf{D})^T. \quad (1)$$

subject to

$$\|\mathbf{a}\|^2 \leq 1, \|\mathbf{b}\|^2 \leq 1, \|\mathbf{a}\|_1 \leq c_1, \|\mathbf{b}\|_1 \leq c_2, \quad (2)$$

$$0 \leq D_{ii} \leq 1, \forall i \in \{1, \dots, k\}, c_3 \leq \sum_{i=1}^k d_i \leq c_4. \quad (3)$$

In the above optimization, $c_j > 0$ ($j = 1, \dots, 4$) are constants defining the amount of regularization: c_1 and c_2 control the level of sparseness of \mathbf{a} and \mathbf{b} ; c_3 and c_4 have the interpretation of being the minimum and maximum number of samples included into a bicluster. The elements in $|\mathbf{a}|$ and $|\mathbf{b}|$ can be interpreted as the magnitude of contribution from each of the EFs to maximizing the canonical correlation in Equation (1). \mathbf{D} tells the level inclusion for each sample. After multiple random partitions of \mathbf{W} by the EFs, the EFs and samples related to a bicluster of our interest will regularly have larger absolute values in \mathbf{a} , \mathbf{b} , and \mathbf{d} .

Since the entire function is known convex, we propose a two-step iterative procedure in SCCA-BC to solve the maximization problem. Briefly, in step (i), we fix \mathbf{d} and solve for \mathbf{a} and \mathbf{b} , and in step (ii), using the solutions from (i), we fix \mathbf{a} and \mathbf{b} and solve for \mathbf{d} . These two steps are iterated until convergence.

Step (i): If we fix \mathbf{D} to a feasible solution, we can rewrite the objective function in Equation (1) as a SCCA problem: Let $\tilde{\mathbf{X}} = (\mathbf{X}\mathbf{D})^T$ and $\tilde{\mathbf{Y}} = (\mathbf{Y}\mathbf{D})^T$ then the objective function becomes

$$\underset{\mathbf{a}, \mathbf{b}}{\text{maximize}} \quad \mathbf{a}^T \tilde{\mathbf{X}} \tilde{\mathbf{Y}} \mathbf{b}. \quad (4)$$

If we maximize Equation (4) subject to Equation (2), we have the SCCA problem in Ref. [19]. We relegate the choice of c_1 and c_2 to the Lee *et al.* [19] method which performs cross-validation.

Step (ii): Now we fix \mathbf{a} and \mathbf{b} and solve for $\mathbf{d} = \text{diag}(\mathbf{D})$. We rewrite the objective function in Equation (1) as:

$$\underset{\mathbf{d}}{\text{minimize}} \quad -\sum_{j=1}^k (\mathbf{a}^T \mathbf{X})_j D_{jj}^2 (\mathbf{Y}^T \mathbf{b})_j = \underset{\mathbf{d}}{\text{minimize}} \quad \sum_{j=1}^k q_j d_j^2, \quad (5)$$

Where $(\mathbf{a}^T \mathbf{X})_j = \sum_{i=1}^n \mathbf{a}_i \mathbf{X}_{ij}$, and $q_j = -(\mathbf{a}^T \mathbf{X})_j \cdot (\mathbf{b}^T \mathbf{Y})_j$. While the objective function in Equation (5) presents a

quadratic problem, the values of q_j are typically a mix of positive and negative values, resulting in a nonconvex objective function. Even though in general the minimum is unbounded, our constraints in Equation (3) allow us to always find at least one minimum.

Note that the smaller the d_i value, the larger the objective function in Equation (5) and the smaller of the canonical correlation in Equation (1). We thus call d_i the “contribution score” for certain sample i to the bicluster of our interest. The algorithm to find \mathbf{d} is seen in Algorithm 1. Remarks on Algorithm 1:

Algorithm 1 Maximizing the sample coefficients. The function ORDER(x) returns the index of the elements of x in increasing order.

```

1:  procedure SAMPLEFUNCMax ( $q$ , lower, upper)
2:       $sum \leftarrow 0$ 
3:       $\mathbf{d} \leftarrow \mathbf{0}$ 
4:      for  $i$  in ORDER( $q$ ) do
5:          if (lower  $\leq$  sum) and  $q_i \geq 0$  then
6:              break
7:          if  $q_i < 0$  then
8:               $d_i \leftarrow \min(\text{upper} - \text{sum}, 1)$ 
9:          else
10:              $d_i \leftarrow \min(\text{lower} - \text{sum}, 1)$ 
11:              $sum \leftarrow sum + d_i$ 
12:             if  $sum = \text{upper}$  then
13:                 break
14:             return  $\mathbf{d}$ 
    
```

1. If $c_3 = 1$, to minimize the objective function in Equation (5), we would take $d_i = 1$ for the smallest q_i with all other $d_j = 0$. To avoid this meaningless solution, c_3 should be ≥ 1 .

2. If we can saturate the upper bound (c_4) without using any nonnegative q_i , then we have indeed reached a minimum. If we have to use some nonnegative q_i , then the lower bound constraint (c_3) is forcing us to include some conditions that will actually weaken the linear relationship in the bicluster. While this is true, in practice c_3 should be a small number simply to avoid the case where only two or three conditions are chosen.

3. In the case where q_i and q_j have tied ranks, one can choose either value. If the index is the final index chosen before termination, then choosing the other index simply corresponds to another local minima. However, after aggregating the results from multiple random partitions of \mathbf{W} , we expect the effects from such local minimas would be minimized.

Post processing of results from SCCA-BC

After multiple random partitions of \mathbf{W} by EFs (into

submatrices X and Y), we can define two matrices F and C to record the estimated contritions of EFs and samples to a bicluster, respectively. Specifically, each column of F will be the list of absolute element values from a and b inferred based on one partition of W . The element values are ordered according to the EF list. The dimension of F is thus $2n$ by T , where T is the number of random partitions applied to W . The matrix C can be defined similarly using the element values in d , ordered according to the sample list for each column. The dimension of C is k by T . Now we are posed with new problems: using F and C to find the EFs and samples that regularly have relatively large values over many partitions.

Since we expect EFs and conditions which belong in the bicluster to have larger F and C on average, we propose to run k -means clustering on F and C separately. The procedure is the same for EFs and conditions. First, k is decided by the gap statistic [20]. We then choose the cluster with the greatest mean centroid, and all EFs in that cluster become part of the bicluster. Since EFs and conditions are chosen independently, there is a possibility of non-correspondence between the EFs chosen and conditions chosen. However, we have noticed that the selection of conditions chosen is quite robust and usually results in one clear non-zero cluster.

After the first bicluster is found, we can remove the identified EFs and update the expression matrix and repeat the analysis to the updated expression matrix to find additional biclusters. This means that the identified biclusters would have non-overlapping EFs but may have overlapping samples.

In the case that too many features are getting selected for one bicluster, it is possible to subsample the data. We do this by selecting $p\%$ of the data randomly and running the algorithm as usual on this subsample. Due to the subsampling, hierarchical clustering is no longer valid as there are many missing values in our matrix and their position is not necessarily meaningful. A possible heuristic for selecting the features is choosing those which are in some upper percentile u , at least $q\%$ of the times it is sampled. In practice, these thresholds u and q can be adjusted for more or less sensitive feature selection. The thresholds can also be adjusted to find a desired number of features.

ACKNOWLEDGEMENTS

The work is partially supported by NSF DMS-1160319.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Harold Pimentel, Zhiyue Hu and Haiyan Huang declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

APPENDIX

Basic SCCA

Sparse canonical correlation analysis (SCCA) is a large component of our method that allows us to find features that show a strong linear relationship among a fixed subset of conditions. SCCA is a penalized extension of canonical correlation analysis (CCA) developed by Ref. [18]. CCA has many interpretations, but perhaps the one best suited for our purposes is the most common, which we will now explain.

Suppose there are two matrices of data, $X \in R^{n \times p}$ and $Y \in R^{n \times q}$. We wish to find two vectors $a \in R^p$, $b \in R^q$ to maximize $\text{Cor}(Xa, Yb)$. Thus, we are interested in vectors a and b which maximize the linear correlation between X and Y . CCA is not scale invariant, so X and Y should be scaled and centered. It follows that $\text{argmax}_{a,b} \text{Cor}(Xa, Yb) = \text{argmax}_{a,b} a^T X^T Y b$.

There has been much work in solving for a and b in the classical 1, terature such as Ref. [21]. Unfortunately, in biology sample sizes tend to be quite small relative to the number of features. Conditions are often on the order of 30–100, While features are often in the thousands. A classical formulation of CCA is then problematic with far too many free parameters as the number of genes is often in the thousands. Naturally, one solution to the problem is to regularize the objective function. In general, the SCCA objective function is:

$$\text{maximize}_{a,b} a^T X^T Y b \quad (\text{A1})$$

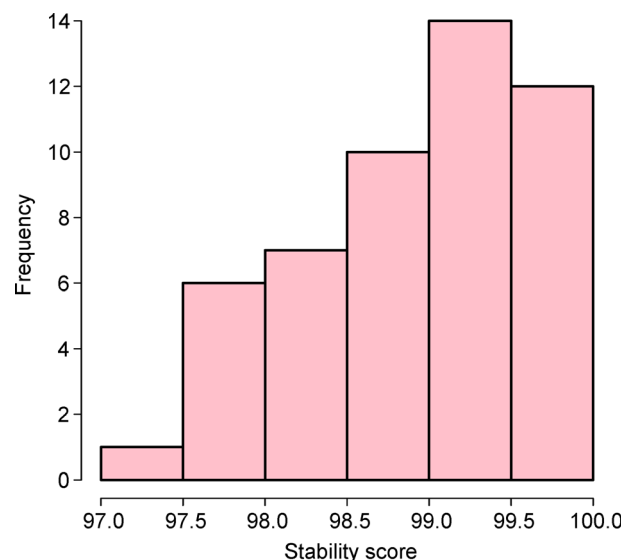


Figure A1. Histogram of the stability score in 50 trials.

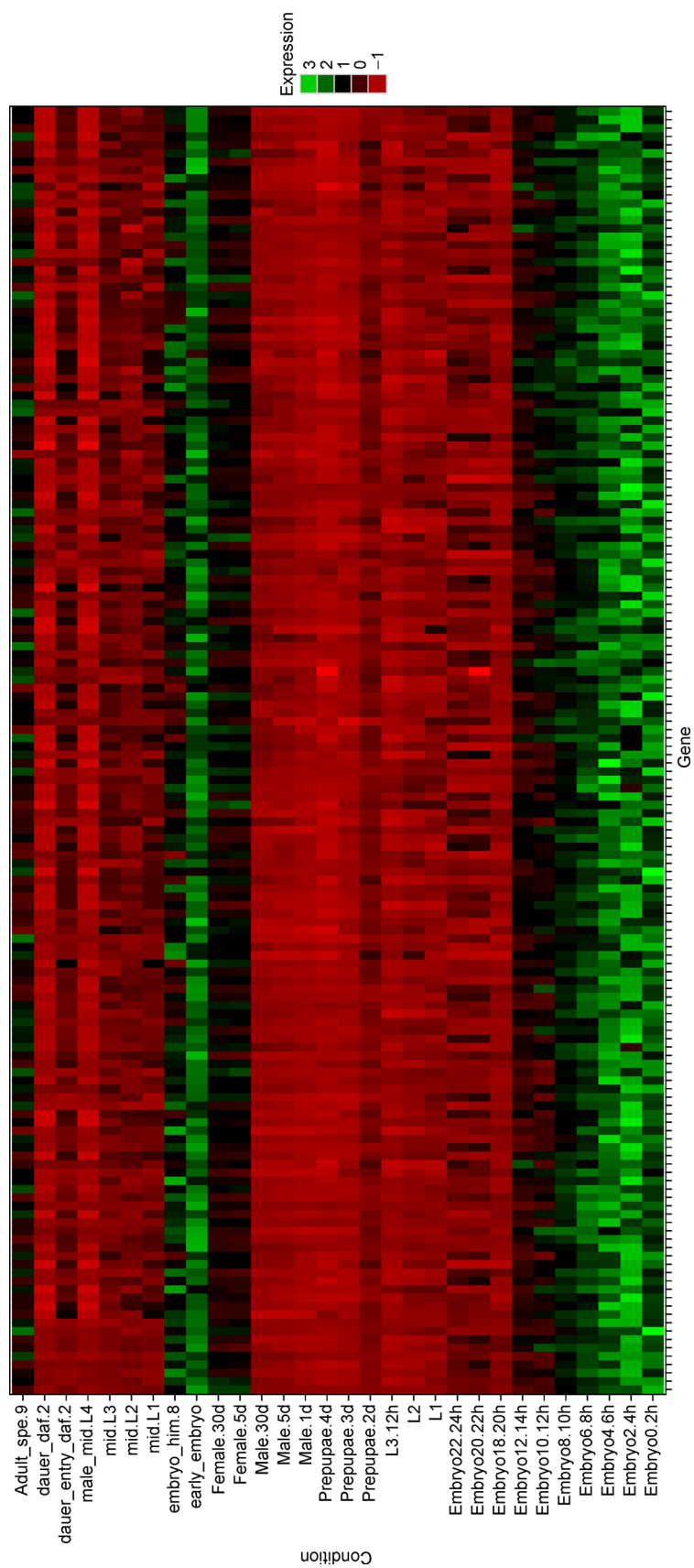


Figure A2. Most highly correlated bicluster found in the fly/worm data set with a slightly different parameter setting.

subject to :

$$\|a\|^2 \leq 1, \|b\|^2 \leq 1, \quad (A2)$$

$$P_1(a) \leq c_1, P_2(b) \leq c_2. \quad (A3)$$

The relaxation of Equation (A2) to an inequality (as opposed to equality) comes from framing the problem as a convex maximization problem (the maximum usually happens around 1). Functions P_1 and P_2 are arbitrary (though usually convex) penalty functions. The L1 norm is often used ($P_i(a) = \sum_j |a_j|$), which enforces sparsity. As in Ref. [22], the L1 norm has nice properties when used to regularize CCA. The L1 norm results in a sparse solution. Sparsity makes sense in biology where we expect only a small amount of features to be involved in the interactions. For our purposes, we use the L1 penalty function.

There are many methods for solving the SCCA problem. A notable method introduced by Ref. [19] was able to recast SCCA into an existing problem, non-linear iterative partial least squares (NIPALS), resulting in a relatively efficient solution. This formulation is the one we follow in our implementation, though the choice is mostly out of convenience and computational efficiency.

Stability of SCCA-BC on real datasets

We now test the stability of our algorithm on the flyworm dataset as provided in the section of modENCODE development RNA-Seq data. Each trial shares the same parameter, but with a different seed. We examine the stability of our algorithm based on the following rule:

$$S = \frac{\mathbb{1}_{\in B}}{|B|}$$

S denotes the score of one single trial. $\mathbb{1}_{\in B}$ denotes the number of elements in the biclusters such iteration overlaps with the base example, which is the result we provided in the section of modENCODE development RNA-Seq data. And $|B|$ denotes the size of biclusters provided in the base example. The histogram of the stability score in 50 trials is provided in Figure A1.

From the figure we can conclude that our algorithm is fairly stable with respect to the initialization state despite the fact that there is random partition and subsampling involved.

We also demonstrate that our method is relatively robust regarding small parameter perturbation. On the same data set, we modify the parameters such that “genes that showed coefficients in the top 95%, 95% of the time they were sampled were chosen as part of the bicluster”. This resulted in slightly less genes selected but ended up

with a relatively similar graph, as shown in Figure A2, to the one provided in the Section of modENCODE Developmental RNA-Seq Data. We also test the output of our algorithm under other small parameter perturbation combinations, for example, 90% and 95%, and these combinations end up producing an almost identical result, hence the graphs are not attached.

REFERENCES

1. Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95, 14863–14868
2. Lazzeroni, L. and Owen, A. (2002) Plaid models for gene expression data. *Stat. Sin.*, 12, 61–86
3. Kluger, Y., Basri, R., Chang, J. T. and Gerstein, M. (2003) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.*, 13, 703–716
4. Bhattacharya, A. and De, R. K. (2009) Bi-correlation clustering algorithm for determining a set of co-regulated genes. *Bioinformatics*, 25, 2795–2801
5. Nepomuceno, J. A., Troncoso, A. and Aguilar-Ruiz, J. S. (2011) Biclustering of gene expression data by correlation-based scatter search. *BioData Min.*, 4, 3
6. Gao, Q., Ho, C., Jia, Y., Li, J. J. and Huang, H. (2012) Biclustering of linear patterns in gene expression data. *J. Comput. Biol.*, 19, 619–631
7. Ben-Dor, A., Chor, B., Karp, R. and Yakhini, Z. (2003) Discovering local structure in gene expression data: the order-preserving submatrix problem. *J. Comput. Biol.*, 10, 373–384
8. Liu, J. and Wang, W. (2003) Op-cluster: Clustering by tendency in high dimensional space. In *Third IEEE International Conference on Data Mining, 2003. ICDM 2003*, pp. 187–194 IEEE
9. Wang, Y. X. R., Jiang, K., Feldman, L. J., Bickel, P. J., and Huang, H. (2015) Inferring gene-gene interactions and functional modules using sparse canonical correlation analysis. *Ann. Appl. Stat.* 9, 300–323
10. Tan, K. M. and Witten, D. M. (2014) Sparse biclustering of transposable data. *J. Comput. Graph. Stat.*, 23, 985–1008
11. Turner, H., Bailey, T. and Krzanowski, W. (2005) Improved biclustering of microarray data demonstrated through systematic performance tests. *Comput. Stat. Data Anal.*, 48, 235–254
12. Lee, M., Shen, H., Huang, J. Z. and Marron, J. S. (2010) Biclustering via sparse singular value decomposition. *Biometrics*, 66, 1087–1095
13. Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bhlmann, P., Gruissem, W., Hennig, L., Thiele, L. and Zitzler, E. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22, 1122–1129
14. Higham, N. J. (2002) Computing the nearest correlation matrix—a problem from finance. *IMA J. Numer. Anal.*, 22, 329–343
15. St Johnston, D. (2002) The art and design of genetic screens: *Drosophila melanogaster*. *Nat. Rev. Genet.*, 3, 176–188
16. Jorgensen, E. M. and Mango, S. E. (2002) The art and design of

- genetic screens: *Caenorhabditis elegans*. Nat. Rev.Genet., 3, 356–369
17. Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., *et al.* (2014) Diversity and dynamics of the *Drosophila* transcriptome. Nature, 512, 393–399
 18. Hotelling, H. (1936) Relations between two sets of variates. Biometrika, 28, 321–377
 19. Lee, W., Lee, D., Lee, Y. and Pawitan, Y. (2011) Sparse canonical covariance analysis for high-throughput data. Stat. Appl. Genet. Mol. Biol., 10
 20. Tibshirani, R., Walther, G. and Hastie, T. (2001) Estimating the number of clusters in a dataset via the gap statistic. J. R. Stat. Soc. B, 63, 411–423
 21. Anderson, T. W. (1958) An Introduction to Multivariate Statistical Analysis. New York: Wiley
 22. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B, 267–288