bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

An analytic approach for interpretable predictive models in high dimensional data, in the presence of interactions with exposures

Sahir R Bhatnagar[1,2], Yi Yang[4], Mathieu Blanchette[5], Budhachandra Khundrakpam[6], Alan Evans[6], Luigi Bouchard[7], and Celia MT Greenwood[1,2,3]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University
[2]Lady Davis Institute, Jewish General Hospital, Montréal, QC
[3]Departments of Oncology and Human Genetics, McGill University
[4]Department of Mathematics and Statistics, McGill University
[5]Montreal Neurological Institute, McGill University
[6]Department of Computer Science, McGill University
[7]Department of Biochemistry, Université de Sherbrooke

January 23, 2017

Abstract

Computational approaches to variable selection have become increasingly important with the advent of high-throughput technologies in genomics and brain imaging studies, where the data has become massive, yet where it is believed that the number of truly important variables is small relative to the total number of variables. Although many approaches have been developed for main effects, less attention has been paid to interaction models. Here, starting from the hypothesis that a binary exposure variable can alter correlation patterns between clusters of high-dimensional variables, i.e. alter network properties of the variables, we explore whether such exposure-dependent clustering relationships can improve predictive modelling of an outcome or phenotype variable. Hence, we propose a modelling framework called ECLUST to test this hypothesis, and evaluate performance through extensive simulations. We see improved model fit in many scenarios. We further illustrate the framework through the analysis of three data sets from very different fields, each with high dimensional data, a binary exposure, and a phenotype of interest. Our method is available in the *eclust* CRAN package.

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 Introduction

In this article, we consider the prediction of an outcome variable y observed on n individuals from p variables, where p is much larger than n. In addition to problems related to overfitting, it can be very challenging to interpret the results of prediction models with ultra-high dimensional predictor sets. For example, multiple different sets of covariates may provide equivalent measures of goodness of fit (Fan et al., 2014). In consequence, many authors have suggested a two-step procedure where the first step is to cluster or group variables in the design matrix, and the model fitting in the second step uses a summary measure of each group of variables.

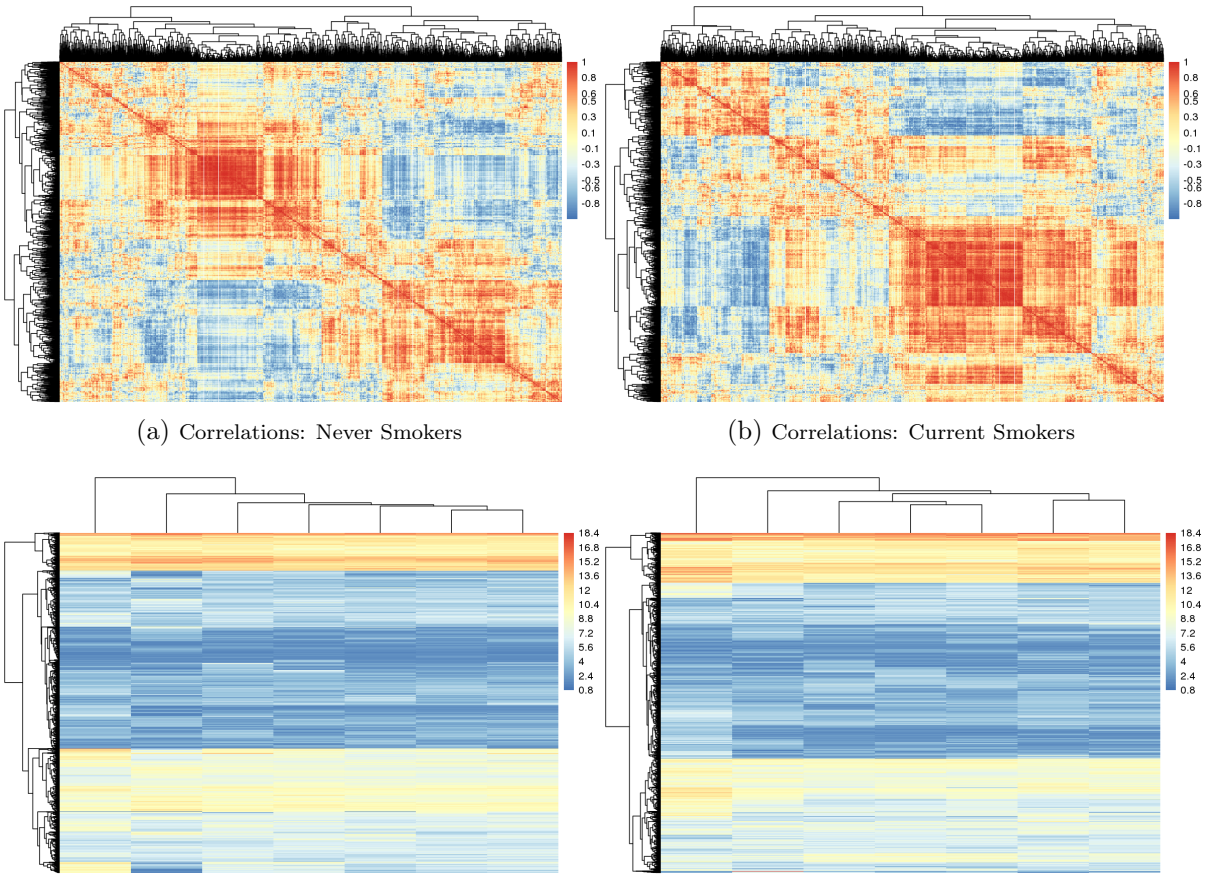
This two-step idea dates back to 1957 when Kendall (Kendall, 1957) first proposed using principal components in regression. Hierarchical clustering based on the correlation of the design matrix has also been used to create groups of genes in microarray studies. For example, at each level of a hierarchy, cluster averages have been used as new sets of potential predictors in both forward-backward selection (Hastie et al., 2001) or the lasso (Park et al., 2007). Bühlmann *et al.* (Bühlmann et al., 2013) proposed a bottom-up agglomerative clustering algorithm based on canonical correlations and used the group lasso on the derived clusters. A more recent proposal performs sparse regression on cluster prototypes (Reid and Tibshirani, 2016), i.e., extracting the most representative gene in a cluster instead of averaging them.

There are several advantages to these two-step methods. Through the reduction of the dimension of the model, the results are often more stable with smaller prediction variance, and through identification of sets of correlated variables, the resulting clusters can provide an easier route to interpretation. From a practical point of view two-step approaches are both flexible and easy to implement because efficient algorithms exist for both clustering (e.g. Müllner (2013)) and model fitting (e.g. Friedman et al. (2010); Yang and Zou (2014); Kuhn (2008)), particularly in the case when the outcome variable is continuous.

These two-step approaches usually group variables based on a matrix of correlations or some transformation of the correlations. However, when there are external factors, such as exposures, that can alter correlation patterns, a dimension reduction step that ignores this information may be suboptimal. Many of the high-dimensional genomic data sets currently being generated capture a possibly dynamic view of how a tissue is functioning, and demonstrate differential patterns of coregulation or correlation under different conditions. We illustrate this critical point with an example of a microarray gene expression dataset available in the *COPDSexualDimorphism.data* package (Sathirapongsasuti, 2013) from Bioconductor. This study measured gene expression in Chronic Obstructive Pulmonary Disease (COPD) patients and controls in addition to their age, gender and smoking status. To see if there was any effect of smoking status on gene expression, we plotted the expression profiles separately for current and never smokers. To balance the covariate profiles, we matched subjects from each group on age, gender and COPD case status, resulting in a sample size of 7 in each group. Heatmaps in Figures 1a, 1b, 1c and 1d show gene expression levels and the corresponding correlation matrices as a function of dichotomized smoking status for 2,900 genes with large variability. Evidently, there are substantial differences in correlation patterns between the smoking groups (Figures 1b and 1a). However, it is difficult to discern any patterns or major differences between the groups when examining the gene expression levels directly (Figures 1d and 1c). This

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

example highlights two key points; 1) environmental exposures can have a widespread effect on regulatory networks and 2) this effect may be more easily discerned by looking at a measure for gene similarity, relative to analyzing raw expression data.



(a) Correlations: Never Smokers

(b) Correlations: Current Smokers

(c) Gene Expression: Never Smokers

(d) Gene Expression: Current Smokers

Figure 1: Heatmaps of gene expression data (top - rows are genes and columns are subjects), and correlations between genes (bottom) stratified by smoking status from a microarray study of COPD (Sathirapongsasuti, 2013). The 20% most variable genes are displayed (2,900 genes). There are 7 subjects in each group, matched on COPD case status, gender and age. Data available on Bioconductor in the `COPDSexualDimorphism.data` package.

Many other examples of altered co-regulation and phenotype associations can be found. For instance, in a pediatric brain development study, very different correlation patterns of cortical thickness within brain regions were observed across age groups , consistent with a process of fine-tuning an immature brain system into a mature one (Khundrakpam et al., 2013). A comparison of gene expression levels in bone marrow from 327 children with acute leukemia found several differentially *coexpressed* genes in philadelphia positive leukemias compared to the cytogenetically normal group (Kostka and Spang, 2004). To give a third example, an analysis of RNA-sequencing data from The Cancer Genome Atlas (TCGA) revealed very different correlation patterns among sets of genes in tumors grouped according to their missense or null mutations in the TP53 tumor suppressor gene (Oros Klein et al., 2016).

Therefore, in this paper, we pose the question whether clustering that incorporates known

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

covariate or exposure information can improve prediction models in high dimensional genomic data settings. Substantial evidence of dysregulation of genomic coregulation has been observed in a variety of contexts, however we are not aware of any work that carefully examines how this might impact the performance of prediction models.

We propose a conceptual analytic strategy called ECLUST for prediction of a continuous outcome in high dimensional contexts while exploiting exposure-sensitive data clusters. We restrict our attention to two-step algorithms in order to implement a covariate-driven clustering. Specifically, we hypothesize that within two-step methods, variable grouping that considers exposure information can lead to improved predictive accuracy and interpretability. In section 2, we describe conceptually the model that is being proposed, particularly focusing on the dimension reduction step (Step 1) of the two step approaches, then in section 3 we use simulations to compare our proposed method to comparable approaches that combine data reduction with predictive modelling. Since we are focusing our attention primarily on the performance of alternative strategies within the first step, we compare performance across a selection of step 2 predictive models that are best adapted to our data. Finally, in section 4 we illustrate these concepts more concretely by analyzing three data sets.

2 Methods

Assume there is a single binary environmental factor E of importance, and an $n \times p$ high dimensional (HD) data set $\mathbf{X}$ (n observations, p features) of relevance. This could be genome-wide epigenetic data, gene expression data, or brain imaging data, for example. Assume there is a continuous or binary phenotype of interest Y and that the environment has a widespread effect on the HD data, i.e., affects many elements of the HD data. The primary goal is to improve prediction of Y by identifying interactions between E and $\mathbf{X}$ through a carefully constructed data reduction strategy that exploits E dependent correlation patterns. The secondary goal is to improve identification of the elements of $\mathbf{X}$ that are involved; we denote this subset by S_0. We hypothesize that a systems-based perspective will be informative when exploring the factors that are associated with a phenotype of interest, and in particular we hypothesize that incorporation of environmental factors into predictive models in a way that retains a high dimensional perspective will improve results and interpretation.

2.1 Potential impacts of covariate-dependent coregulation

Motivated by real world examples of differential coexpression, we first demonstrate that environment-dependent correlations in $\mathbf{X}$ can induce an interaction model. Without loss of generality, let $p = 2$ and the relationship between X_1 and X_2 depend on the environment such that

$$X_{i2} = \psi X_{i1} E_i + \varepsilon_i \tag{1}$$

where ε_i is an error term and ψ is a slope parameter, that is:

$$X_{i2} = \begin{cases} \psi X_{i1} + \varepsilon_i & \text{when } E_i = 1 \\ \varepsilon_i & \text{when } E_i = 0 \end{cases}$$

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Consider the 2-predictor regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i^* \tag{2}$$

where ε_i^* is another error term which is independent of ε_i. At first glance (2) does not contain any interaction terms. However, substituting (1) for X_{i2} in (2) we get

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \psi \left(X_{i1} E_i \right) + \varepsilon_i \beta_2 + \varepsilon_i^* \tag{3}$$

The third term in (3) resembles an interaction model, with $\beta_2 \psi$ being the interaction parameter. We present a second illustration showing how non-linearity can induce interactions. Suppose

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 \max_{j \in \{1,2\}} \left(X_{ij} - \bar{X}_i \right)^2 + \varepsilon_i^* \tag{4}$$

Substituting (1) for X_{i2} in (4) we get a non-linear interaction term.

2.2 Proposed framework and algorithm

We restrict attention to methods containing two phases as illustrated in Figure 2: 1a) a clustering stage where variables are clustered based on some measure of similarity, 1b) a dimension reduction stage where a summary measure is created for each of the clusters, and 2) a simultaneous variable selection and regression stage on the summarized cluster measures. Although this framework appears very similar to any two-step approach, our hypothesis is that allowing the clustering in Step 1a to depend on the environment variable can lead to improvements in prediction after Step 2. Hence, methods in Step 1a are adapted to this end, as decribed in section 2.2.1.

Our focus in this manuscript is on the clustering and cluster representation steps. Therefore, we compare several well known methods for variable selection and regression that are best adapted to our simulation designs and data sets.

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

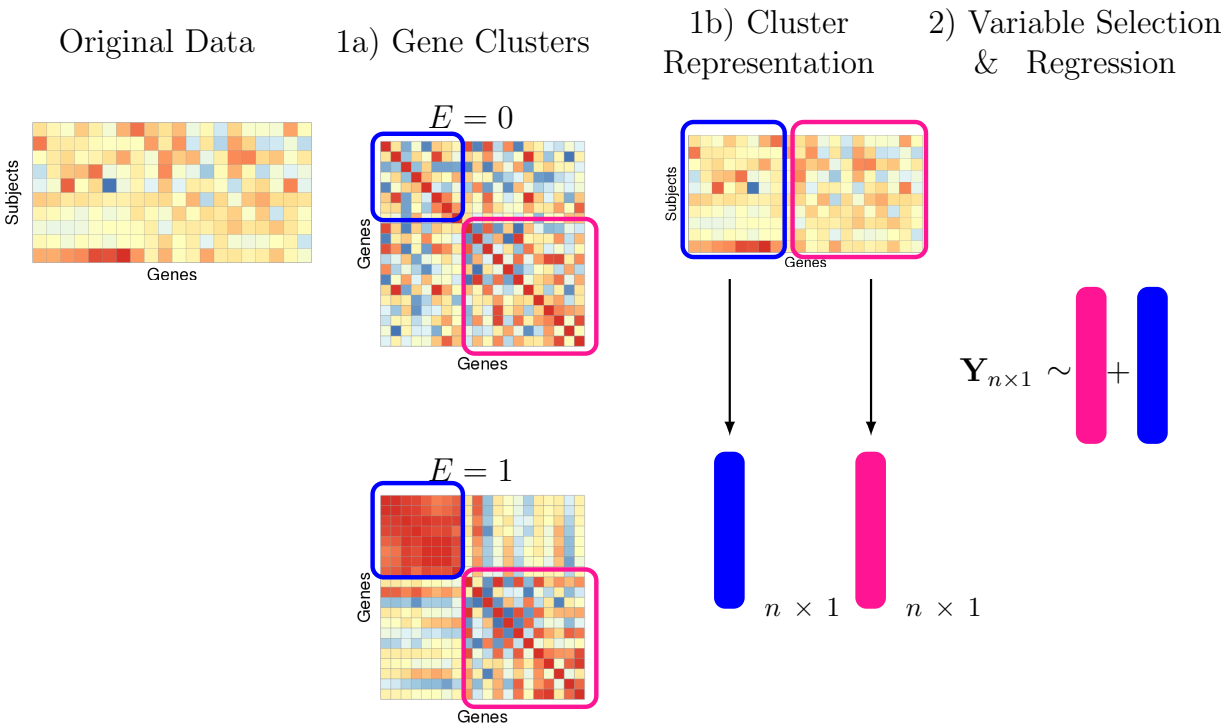


Figure 2: Overview of our proposed method

2.2.1 Step 1a: Clustering using co-expression networks that are influenced by the environment

In agglomerative clustering, a measure of similarity between sets of observations is required in order to decide which clusters should be combined. Common choices include Euclidean, maximum and absolute distance. A more natural choice in genomic or brain imaging data is to use Pearson correlation (or its absolute value) because the derived clusters are biologically interpretable. Indeed, genes that cluster together are correlated and thus likely to be involved in the same cellular process. Similarly, cortical thickness measures of the brain tend to be correlated within pre-defined regions such as the left and right hemisphere, or frontal and temporal regions (Sato et al., 2013). However, the information on the connection between two variables, as measured by the Pearson correlation for example, may be noisy or incomplete. Thus it is of interest to consider alternative measures of pairwise interconnectedness. Gene co-expression networks are being used to explore the system-level function of genes, where nodes represent genes and are connected if they are significantly co-expressed (Zhang and Horvath, 2005), and here we use their overlap measure (Ravasz et al., 2002) to capture connectnedness between two X variables within each environmental condition.

As was discussed earlier, genes can exhibit very different patterns of correlation in one environment versus the other (e.g. Figure 1). Furthermore, measures of similarity that go beyond pairwise correlations and consider the shared connectedness between nodes can be useful in elucidating networks that are biologically meaningful. Therefore, we propose to first look at the topological overlap matrix (TOM) separately for exposed ($E = 1$) and unexposed ($E = 0$) individuals (see Section A.1 for details on the TOM). We then seek to identify nodes that are very different between environments. We determine differential co-

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expression using the absolute difference $TOM(X_{\text{diff}}) = |TOM_{E=1} - TOM_{E=0}|$ (Oros Klein et al., 2016). We then use hierarchical clustering with average linkage on the derived difference matrix to identify these differentially co-expressed variables. Clusters are automatically chosen using the `dynamicTreeCut` (Langfelder et al., 2008) algorithm. Of course, there could be other clusters which are not sensitive to the environment. For this reason we also create a set of clusters based on the TOM for all subjects denoted $TOM(X_{\text{all}})$. This will lead to each covariate appearing in two clusters. In the sequel we denote the clusters derived from $TOM(X_{\text{all}})$ as the set $C_{\text{all}} = \{C_1, \ldots, C_k\}$, and those derived from $TOM(X_{\text{diff}})$ as the set $C_{\text{diff}} = \{C_{k+1}, \ldots, C_\ell\}$ where $k < \ell < p$.

2.2.2 Step 1b: Dimension reduction via cluster representative

Once the clusters have been identified in phase 1, we proceed to reduce the dimensionality of the overall problem by creating a summary measure for each cluster. A low-dimensional structure, i.e. grouping when captured in a regression model, improves predictive performance and facilitates a model's interpretability. We propose to summarize a cluster by a single representative number. Specifically, we chose the average values across all measures (Park et al., 2007; Bühlmann et al., 2013), and the first principal component (Langfelder and Horvath, 2007). These representative measures are indexed by their cluster, i.e., the variables to be used in our predictive models are $\widetilde{\mathbf{X}}_{\text{all}} = \left\{\widetilde{X}_{C_1}, \ldots, \widetilde{X}_{C_k}\right\}$ for clusters that do not consider E, as well as $\widetilde{\mathbf{X}}_{\text{diff}} = \left\{\widetilde{X}_{C_{k+1}}, \ldots, \widetilde{X}_{C_\ell}\right\}$ for E-derived clusters. The tilde notation on the X is to emphasize that these variables are different from the separate variables in the original data.

2.2.3 Step 2: Variable Selection and Regression

Because the clustering in phase 1 is unsupervised, it is possible that the derived latent representations from phase 2 will not be associated with the response. We therefore use penalized methods for supervised variable selection, including the lasso (Tibshirani, 1996) and elasticnet (Zou and Hastie, 2005) for linear models, and multivariate adaptive regression splines (MARS) (Friedman, 1991) for nonlinear models. We argue that the selected non-zero predictors in this model will represent clusters of genes that interact with the environment and are associated with the phenotype. Such an additive model might be insufficient for predicting the outcome. In this case we may directly include the environment variable, the summary measures and their interaction.

In the light of our goals to improve prediction and interpretability, we consider the following model

$$g(\boldsymbol{\mu}) = \beta_0 + \sum_{j=1}^{\ell} \beta_j \widetilde{X}_{C_j} + \beta_E E + \sum_{j=1}^{\ell} \alpha_j \left(\widetilde{X}_{C_j} E\right) + \varepsilon \tag{5}$$

where $g(\cdot)$ is a known link function, $\boldsymbol{\mu} = \mathsf{E}[Y|\mathbf{X}, E, \boldsymbol{\beta}, \boldsymbol{\alpha}]$ and $\widetilde{X}_{C_j}$ are linear combinations of $\mathbf{X}$ (from Step 1b). The primary comparison is models with $\widetilde{\mathbf{X}}_{\text{all}}$ only versus models with $\widetilde{\mathbf{X}}_{\text{all}}$ and $\widetilde{\mathbf{X}}_{\text{diff}}$. Given the context of either the simulation or the data set, we use either linear models or non linear models. Our general approach, ECLUST, can therefore be summarized by the algorithm in Table 1.

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1: Details of ECLUST algorithm

Step	Description, Software[a] and Reference
1a)	i Calculate TOM separately for observations with $E = 0$ and $E = 1$ using `WGCNA::TOMsimilarityFromExpr` (Langfelder and Horvath, 2008) ii Compute the Euclidean distance matrix of $\lvert TOM_{E=1} - TOM_{E=0} \rvert$ using `stats::dist` iii Run the `dynamicTreeCut` algorithm (Langfelder et al., 2008, 2016) on the distance matrix to determine the number of clusters and cluster membership using `dynamicTreeCut::cutreeDynamic` with `minClusterSize = 50`
1b)	i Calculate the 1st principal component or average for each cluster using `stat::prcomp` or `base::mean` ii For the penalized regression models, create a design matrix of the derived cluster representatives and their interactions with E using `stats::model.matrix` iii For the MARS model, create a design matrix of the derived cluster representatives and E
2)	• For linear models, run penalized regression on design matrix from step 1b) using `glmnet::cv.glmnet` (Friedman et al., 2010).Elasticnet mixing parameter `alpha=1` corresponds to the lasso and `alpha=0.5` corresponds to the value we used in our simulations for elasticnet. The tuning parameter `alpha` is selected by minimizing 10 fold cross-validated mean squared error (MSE). • For non-linear effects, run MARS on the design matrix from step 1b) using `earth::earth` (Milborrow. Derived from mda:mars by T. Hastie and R. Tibshirani., 2011) with pruning method `pmethod = "backward"` and maximum number of model terms `nk = 1000`. The `degree=1,2` is chosen using 10 fold cross validation (CV), and within each fold the number of terms in the model is the one that minimizes the generalized cross validated (GCV) error.

[a] All functions are implemented in `R` (R Core Team, 2016). The naming convention is as follows: `package_name::package_function`. Default settings used for all functions unless indicated otherwise.

3 Simulation Studies

We have evaluated the performance of our ECLUST method in a variety of simulated scenarios. In each, we compared analytic approaches that do not cluster the variables at all, methods that cluster variables but do not account for the environmental exposure in the clustering step, and finally ECLUST, which clusters both with and without considering the environmental exposure. A detailed description of the methods being compared is summarized in Table 2. We have designed 6 simulation scenarios that illustrate different kinds of relationships between the variables and the response. For all scenarios, we have

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

created high dimensional data sets with p predictors ($p = 5000$), and sample sizes of $n = 200$. We also assume that we have two data sets for each simulation - a training data set where the parameters are estimated, and a testing data set where prediction performance is evaluated, each of size $n_{train} = n_{test} = 200$. The number of subjects who were exposed ($n_{E=1} = 100$) and unexposed ($n_{E=0} = 100$) and the number of truly associated parameters ($|S_0| = 500$) remain fixed across the 6 simulation scenarios.

Let

$$Y = Y^* + k \cdot \varepsilon \tag{6}$$

where Y^* is the linear predictor, the error term ε is generated from a standard normal distribution, and k is chosen such that the signal-to-noise ratio $SNR = (Var(Y^*)/Var(\varepsilon))$ is 0.2, 1 and 2 (e.g. the variance of the response variable Y due to ε is $1/SNR$ of the variance of Y due to Y^*).

Table 2: Summary of methods used in simulation study

General Approach	Summary Measure of Feature Clusters	Description[a b]
SEPARATE	NA	Regression of the original predictors $\{X_1, \ldots, X_p\}$ on the response i.e. no transformation of the predictors is being done here
CLUST	1st principal component, average	Create clusters of predictors without using the environment variable $\{C_1, \ldots, C_k\}$. Use the summary measure of each cluster as inputs of the regression model.
ECLUST	1st principal component, average	Create clusters of predictors using the environment variable $\{C_{k+1}, \ldots, C_\ell\}$ where $k < \ell < p$, as well as clusters without the environment variable $\{C_1, \ldots, C_k\}$. Use summary measures of $\{C_1, \ldots, C_\ell\}$ as inputs of the regression model.

[a] Simulations 1 and 2 used lasso and elasticnet for the linear models, and simulation 3 used MARS for estimating non-linear effects

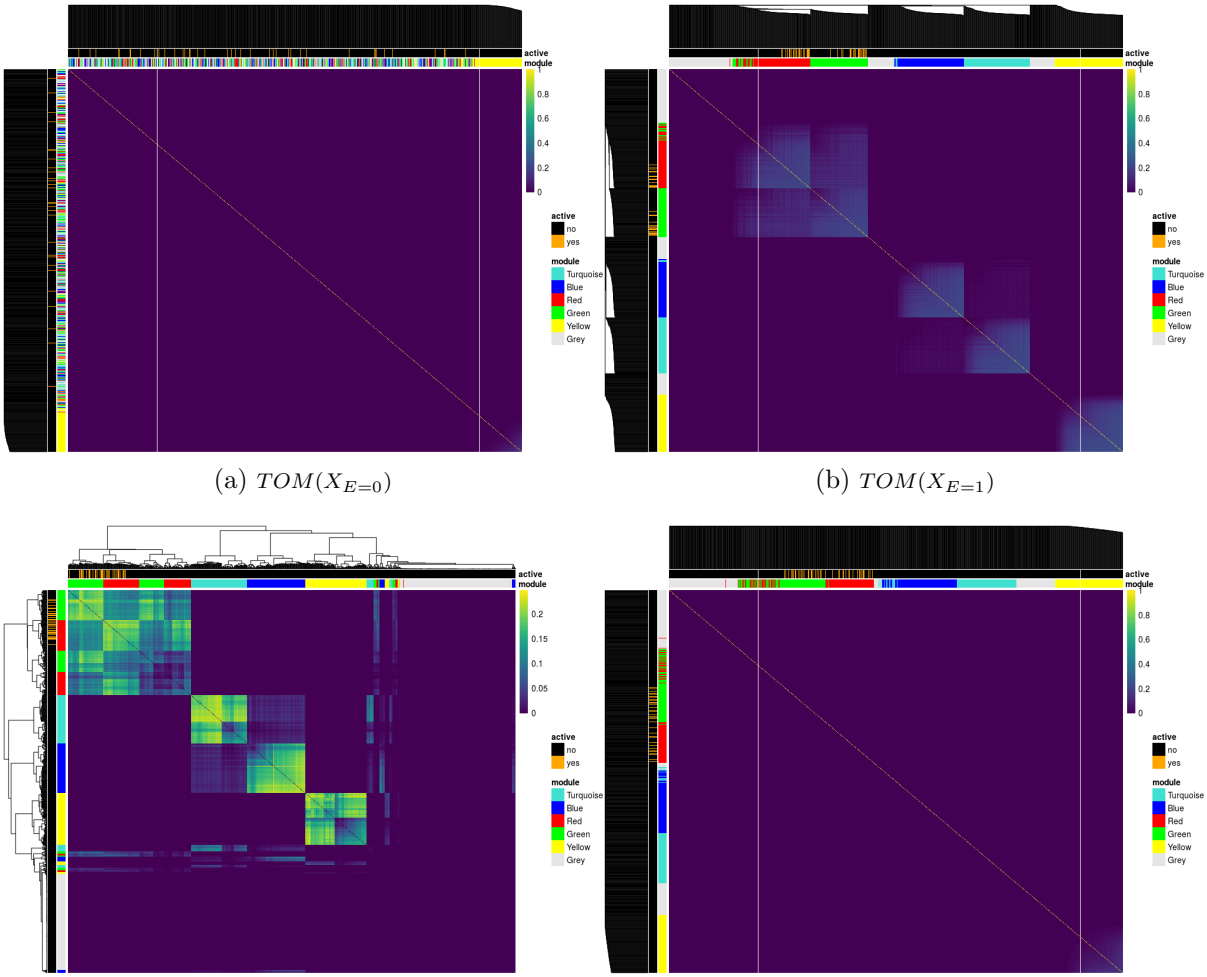
[b] Simulations 4, 5 and 6 convert the continuous response generated in simulations 1, 2 and 3, respectively, into a binary response

3.1 The Design Matrix

We generated covariate data in blocks using the `simulateDatExpr` function from the `WGCNA` package in `R` (version 1.51). This generates data from a latent vector: first a seed vector is simulated, then covariates are generated with varying degree of correlation with the seed vector in a given block. We simulated five clusters (blocks), each of size 750 variables, and labeled them by colour (turquoise, blue, red, green and yellow), while the remaining 1250 variables were simulated as independent standard normal vectors (grey) (Figure 3). For the unexposed observations ($E = 0$), only the predictors in the yellow block were simulated with correlation, while all other covariates were independent within and between blocks. The TOM values are very small for the yellow cluster because it is not correlated with any of its neighbors. For the exposed observations ($E = 1$), all 5 blocks contained predictors that are correlated. The blue and turquoise blocks are set to have an average correlation of 0.6. The average correlation was varied for both green and red clusters $\rho = \{0.2, 0.9\}$ and the active set S_0, that are directly associated

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with y, was distributed evenly between these two blocks. Heatmaps of the TOM for this environment dependent correlation structure are shown in Figure 3 with annotations for the true clusters and active variables. This design matrix shows widespread changes in gene networks in the exposed environment, and this subsequently affects the phenotype through the two associated clusters. There are also pathways that respond to changes in the environment but are not associated with the response (blue and turquoise), while others that are neither active in the disease nor affected by the environment (yellow).



(a) $TOM(X_{E=0})$

(b) $TOM(X_{E=1})$

(c) $|TOM(X_{E=1}) - TOM(X_{E=0})|$

(d) $TOM(X_{\text{all}})$

Figure 3: Topological overlap matrices (TOM) of simulated predictors based on subjects with (a) $E = 0$, (b) $E = 1$, (c) their absolute difference and (d) all subjects. Dendrograms are from hierarchical clustering (average linkage) of one minus the TOM for a, b, and d and the euclidean distance for c. Some variables in the red and green clusters are associated with the outcome variable. The *module* annotation represents the true cluster membership for each predictor, and the *active* annotation represents the truly associated predictors with the response.

3.2 The response

The first three simulation scenarios differ in how the linear predictor Y^* in (6) is defined, and also in the choice of regression model used to fit the data. In simulations 1 and

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2 we use lasso (Tibshirani, 1996) and elasticnet (Zou and Hastie, 2005) to fit linear models; then we use MARS (Friedman, 1991) in simulation 3 to estimate non-linear effects. Simulations 4, 5 and 6 use the GLM version of these models, respectively, since the responses are binary.

Simulation 1

Simulation 1 was designed to evaluate performance when there are no explicit interactions between X and E (see (3)). We generated the linear predictor from

$$Y^* = \sum_{\substack{j\in\{1,\ldots,250\} \\ j\in \text{ red, green block}}} \beta_j X_j + \beta_E E \tag{7}$$

where $\beta_j \sim \text{Unif}\,[0.9, 1.1]$ and $\beta_E = 2$. That is, only the first 250 predictors of both the red and green blocks are active. In this setting, only the main effects model is being fit to the simulated data.

Simulation 2

In the second scenario we explicitly simulated interactions. All non-zero main effects also had a corresponding non-zero interaction effect with E. We generated the linear predictor from

$$Y^* = \sum_{\substack{j\in\{1,\ldots,125\} \\ j\in \text{ red, green block}}} \beta_j X_j + \alpha_j X_j E + \beta_E E \tag{8}$$

where $\beta_j \sim \text{Unif}\,[0.9, 1.1]$, $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$ or $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$, and $\beta_E = 2$. In this setting, both the main effects and their interactions with E are being fit to the simulated data.

Simulation 3

In the third simulation we investigated the performance of the ECLUST approach in the presence of non-linear effects of the predictors on the phenotype:

$$Y_i^* = \sum_{\substack{j\in\{1,\ldots,250\} \\ j\in \text{ red, green block}}} \beta_j X_{ij} + \beta_E E_i + \alpha_Q E_i \cdot f(Q_i) \tag{9}$$

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where

$$Q_i = -\max_{\substack{j\in\{1,\ldots,250\} \\ j\in \text{ red, green block}}} \left(X_{ij} - \bar{X}_i\right)^2 \tag{10}$$

$$f(u_i) = \frac{u_i - \min\limits_{i\in\{1,\ldots,n\}} u_i}{-\min\limits_{i\in\{1,\ldots,n\}} u_i} \tag{11}$$

$$\bar{X}_i = \frac{1}{500} \sum_{\substack{j\in\{1,\ldots,250\} \\ j\in \text{ red, green block}}} X_{ij}$$

In this third simulation, we set $\beta_j \sim \text{Unif}\,[0.9, 1.1]$, $\beta_E = 2$ and $\alpha_Q = 1$. We assume the data has been appropriately normalized, and that the correlation between any two features is greater than or equal to 0.

In simulation 3, we tried to capture the idea that an exposure could lead to coregulation or disregulation of a cluster of X's, which in itself directly impacts Y. Hence, we defined coregulation as the X's being similar in magnitude and disregulation as the X's being very different. The Q_i term in (10) is defined such that higher values would correspond to strong coregulation whereas lower values correspond to disregulation. For example, suppose Q_i ranges from -5 to 0. It will be -5 when there is lots of variability (disregulation) and 0 when there is none (strong coregulation). The function $f(\cdot)$ in (11) simply maps Q_i to the [0,1] range. In order to get an idea of the relationship in (9), Figure 4 displays the response Y as a function of the first principal component of $\sum_j \beta_j X_{ij}$ (denoted by `1st PC`) and $f(Q_i)$. We see that lower values of $f(Q_i)$ (which implies disregulation of the features) leads to a lower Y. In this setting, although the clusters do not explicitly include interactions between the X variables, the MARS algorithm allows for the possibility of two way interactions between any of the variables.

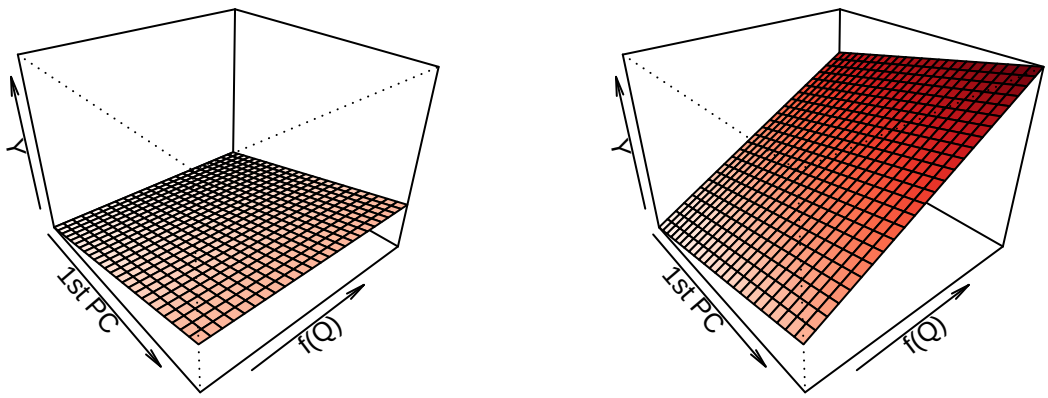


Figure 4: Visualization of the relationship between the response, the first principal component of the main effects and $f(Q_i)$ in (9) for $E = 0$ (left) and $E = 1$ (right) in simulation scenario 3.

Simulation 4, 5 and 6

We used the same simulation setup as above, except that we took the continuous outcome Y, defined $p = 1/(1 + exp(-Y))$ and used this to generate a two-class outcome z with

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$\Pr(z = 1) = p$ and $\Pr(z = 0) = 1 - p$. Simulations 4, 5 and 6 are the binary response versions of simulations 1, 2 and 3, respectively.

3.3 Measures of Performance

Simulation performance was assessed with measures of model fit, prediction accuracy and feature stability. Several measures for each of these categories, and the specific formulae used are provided in Table 3. We simulated both a training data set and a test data set for each simulation: all tuning parameters for model selection were selected using the training sets only. Although most of the measures of model fit were calculated on the test data sets, true positive rate, false positive rate and correct sparsity were calculated on the training set only. The root mean squared error is determined by predicting the response for the test set using the fitted model on the training set. The area under the curve is determined using the trapezoidal rule (Robin et al., 2011).

The stability of feature importance is defined as the variability of feature weights under perturbations of the training set, i.e., small modifications in the training set should not lead to considerable changes in the set of important covariates (Toloşi and Lengauer, 2011). A feature selection algorithm produces a weight (e.g. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$), a ranking (e.g. $rank(\boldsymbol{\beta}) : \mathbf{r} = (r_1, \ldots, r_m)$) and a subset of features (e.g. $\mathbf{s} = (s_1, \ldots, s_p), s_j = \mathbb{I}\{\beta_j \neq 0\}$ where $\mathbb{I}\{\cdot\}$ is the indicator function). In the CLUST and ECLUST methods, we defined a predictor to be non-zero if its corresponding cluster representative weight was non-zero.

Using 10-fold cross validation (CV), we evaluated the similarity between two features and their rankings using Pearson and Spearman correlation, respectively. For each CV fold we re-ran the models and took the average Pearson/Spearman correlations of the $\binom{10}{2}$ combinations of estimated coefficients vectors. To measure the similarity between two subsets of features we took the average of the Jaccard distance in each fold. A Jaccard distance of 1 indicates perfect agreement between two sets while no agreement will result in a distance of 0. For MARS models we do not report the Pearson/Spearman stability rankings due to the adaptive and functional nature of the model (there are many possible combinations of predictors, each of which are linear basis functions).

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3: Measures of Performance

Measure	Formula
Model Fit	
True Positive Rate (TPR)	$\lvert\widehat{S} \in S_0\rvert / \lvert S_0\rvert$
False Positive Rate (TPR)	$\lvert\widehat{S} \notin S_0\rvert / \lvert j \notin S_0\rvert$
Correct Sparsity (Witten et al., 2014)	$\frac{1}{p}\sum_{j=1}^{p} A_j$ $A_j = \begin{cases} 1 & \text{if } \hat{\beta}_j = \beta_j = 0 \\ 1 & \text{if } \hat{\beta}_j \neq, \beta_j \neq 0 \\ 0 & \text{if else} \end{cases}$
Prediction Accuracy	
Root Mean Squared Error (RMSE)	$\lVert \mathbf{Y}_{test} - \widehat{\boldsymbol{\mu}}(\mathbf{X}_{test}) \rVert_2$
Area Under the Curve (AUC)	Trapezoidal rule
Feature Stability using K-fold Cross-Validation on training set (Kalousis et al., 2007)	
Pearson Correlation (ρ) (Pearson, 1895)	$\binom{K}{2}^{-1} \sum_{i,j\in\{1,\ldots,K\}, i\neq j} \frac{cov\left(\widehat{\beta}_{(i)}, \widehat{\beta}_{(j)}\right)}{\sigma_{\widehat{\beta}_{(i)}} \sigma_{\widehat{\beta}_{(j)}}}$
Spearman Correlation (r) (Spearman, 1904)	$\binom{K}{2}^{-1} \sum_{i,j\in\{1,\ldots,K\}, i\neq j} \left[1 - 6\sum_m \frac{\left(r_{m_{(i)}} - r_{m_{(j)}}\right)^2}{p(p^2-1)}\right]$
Jaccard Distance (Jaccard, 1912)	$\frac{\lvert\widehat{S}_{(i)} \cap \widehat{S}_{(j)}\rvert}{\lvert\widehat{S}_{(i)} \cup \widehat{S}_{(j)}\rvert}$

[a] $\widehat{\boldsymbol{\mu}}$: fitting procedure on the training set
[b] S_0: index of active set $= \left\{j; \beta_j^0 \neq 0\right\}$
[c] $\widehat{S}$: index of the set of non-zero estimated coefficients $= \left\{j; \widehat{\beta}_j \neq 0\right\}$
[d] $|A|$: is the cardinality of set A

3.4 Results

All reported results are based on 200 simulation runs. We graphically summarized the results across simulations 1-3 for model fit (Figure 5) and feature stability (Figure 6). The results for simulations 4-6 are shown in the supplemental material (Figures C.1 and C.2). We restrict our attention to $SNR = 1$, $\rho = 0.9$, and $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Complete results for different values of ρ, SNR and α_j (when applicable) are available in the supplemental material (Section D). The model names are labeled as `summary measure_model` (e.g. `avg_lasso` corresponds using the average of the features in a cluster as inputs into a lasso regression model). When there is no summary measure appearing in the model name, that indicates that the original variables were used (e.g. `enet` means all separate features were used in the elasticnet model).

In panel A of Figure 5, we plot the true positive rate against the false positive rate for each of the 200 simulations. We see that across all simulation scenarios, the SEPARATE method has extremely poor sensitivity compared to both CLUST and ECLUST, which

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

do much better at identifying the active variables, though the resulting models are not always sparse. The relatively few number of green points in panel A is due to the small number of estimated clusters (see Figure F.1 for details) leading to very little variability in performance across simulations. The better performance of ECLUST over CLUST is noticeable as more points lie in the top left part of the plot. The horizontal banding in panel A reflects the stability of the TOM-based clustering approach; these jumps in sensitivity were not seen when correlation-based measures of distance were used (not shown). ECLUST also does better than CLUST in correctly determining whether a feature is zero or nonzero (Figure 5, panel B). Importantly, across all three simulation scenarios, ECLUST outperforms the competing methods in terms of RMSE (Figure 5, panel C), regardless of the summary measure and modeling procedure.

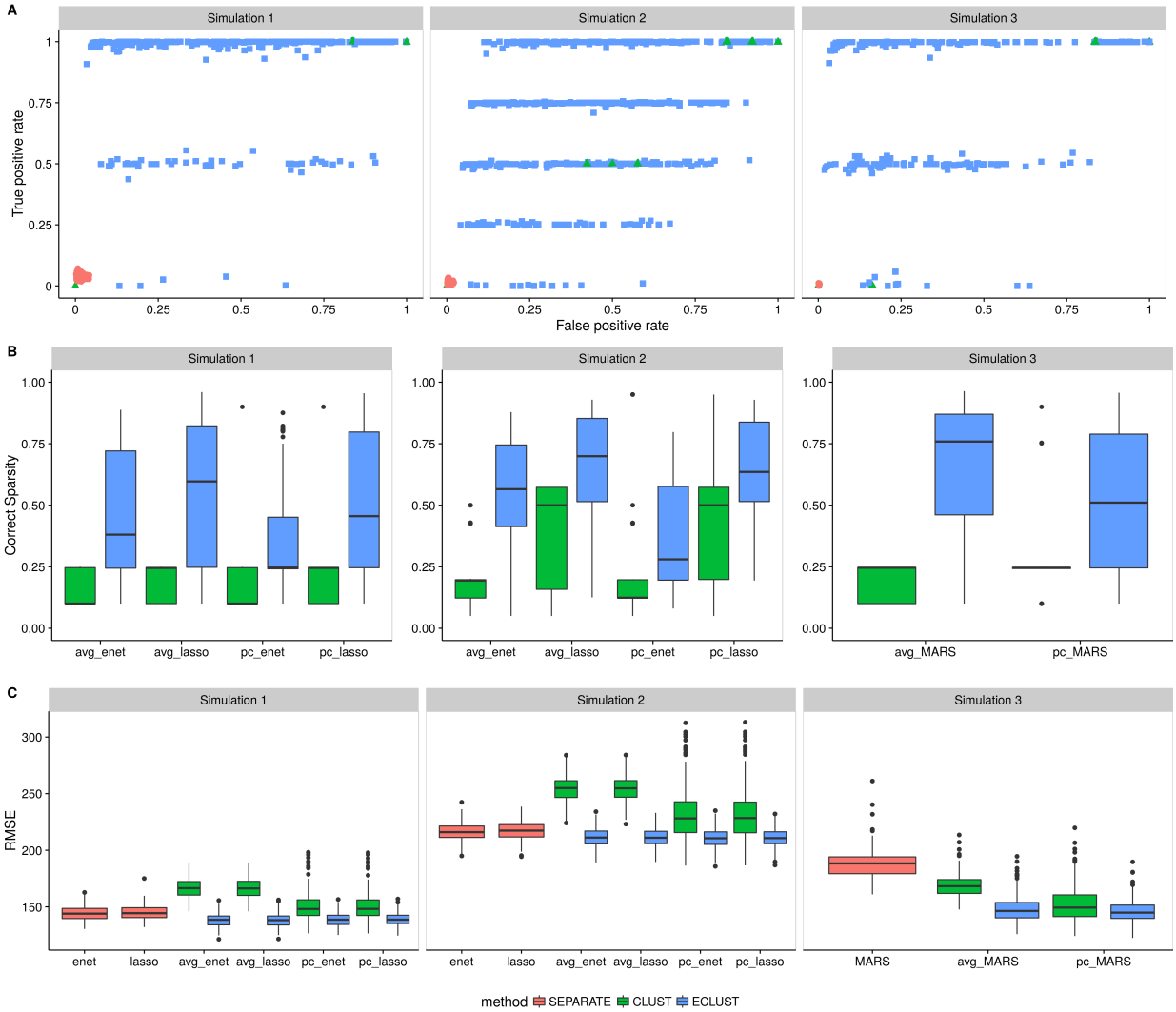


Figure 5: Model fit results in simulations. SEPARATE results are in pink, CLUST in green and ECLUST in blue.

While the approach using all separate original variables (SEPARATE) produce sparse models, they are sensitive to small perturbations of the data across all stability measures (Figure 6), i.e, similar datasets produce very different models. Although the median for the CLUST approach is always slightly better than the median for ECLUST across

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

all stability measures, CLUST results can be much more variable, particularly when stability is measured by the agreement between the value and the ranking of the estimated coefficients across CV folds (Figure 6, panels B and C). The number of estimated clusters, and therefore the number of features in the regression model, tends to be much smaller in CLUST compared to ECLUST, and this explains its poorer performance using the stability measures in Figure 6, since there are more coefficients to estimate. Overall, we observe that the relative performance of ECLUST versus CLUST in terms of stability is consistent across the two summary measures (average or principal component) and across the penalization procedures. The complete results in the supplemental material show that these conclusions are not sensitive to the SNR, ρ or α_j. Similar conclusions are made for a binary outcome using logistic regression versions of the lasso, elasticnet and MARS.

We also ran all our simulations using the Pearson correlation matrix as a measure of similarity (Figure B.1) in order to compare its performance against the TOM. The complete results are in the supplemental material (Section E). In general, we see slightly better performance of CLUST over ECLUST when using Pearson correlations. This result is probably due to the imprecision in the estimated correlations. The exposure dependent similarity matrices are quite noisy, and the variability is even larger when we examine the differences between two correlation matrices. Such large levels of variability have a negative impact on the clustering algorithm's ability to detecting the true clusters.

4 Analysis of three data sets

In this section we demonstrate the performance of ECLUST on three high dimensional datasets with contrasting motivations and features. In the first data set, normal brain development is examined in conjunction with intelligence scores. In the second data set we aim to identify molecular subtypes of ovarian cancer using gene expression data. The investigators' goal in the third data set is to examine the impact of gestational diabetes mellitus (GDM) on childhood obesity in a sample of mother-child pairs from a prospective birth cohort. The datasets comprise a range of sample sizes, and both the amount of clustering in the HD data and the strength of the effects of the designated exposure variables vary substantially. Due to the complex nature of these datasets, we decided to use MARS models for step 2 of our algorithm for all 3 datasets, as outlined in Table 1.

In order to assess performance in these data sets, we have computed the 0.632 estimator (Efron, 1983) and the 95% confidence interval of the R^2 and RMSE from 100 bootstrap samples. The R^2 reported here is defined as the squared Pearson correlation coefficient between the observed and predicted response (Kvålseth, 1985), and the RMSE is defined as in Table 3. Because MARS models can result in unstable predictors (Kuhn, 2008), we also report the results of bagged MARS from $B = 50$ bootstrap samples, where bagging (Breiman, 1996) refers to averaging the predictions from each of the MARS models fit on the B bootstrap samples.

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

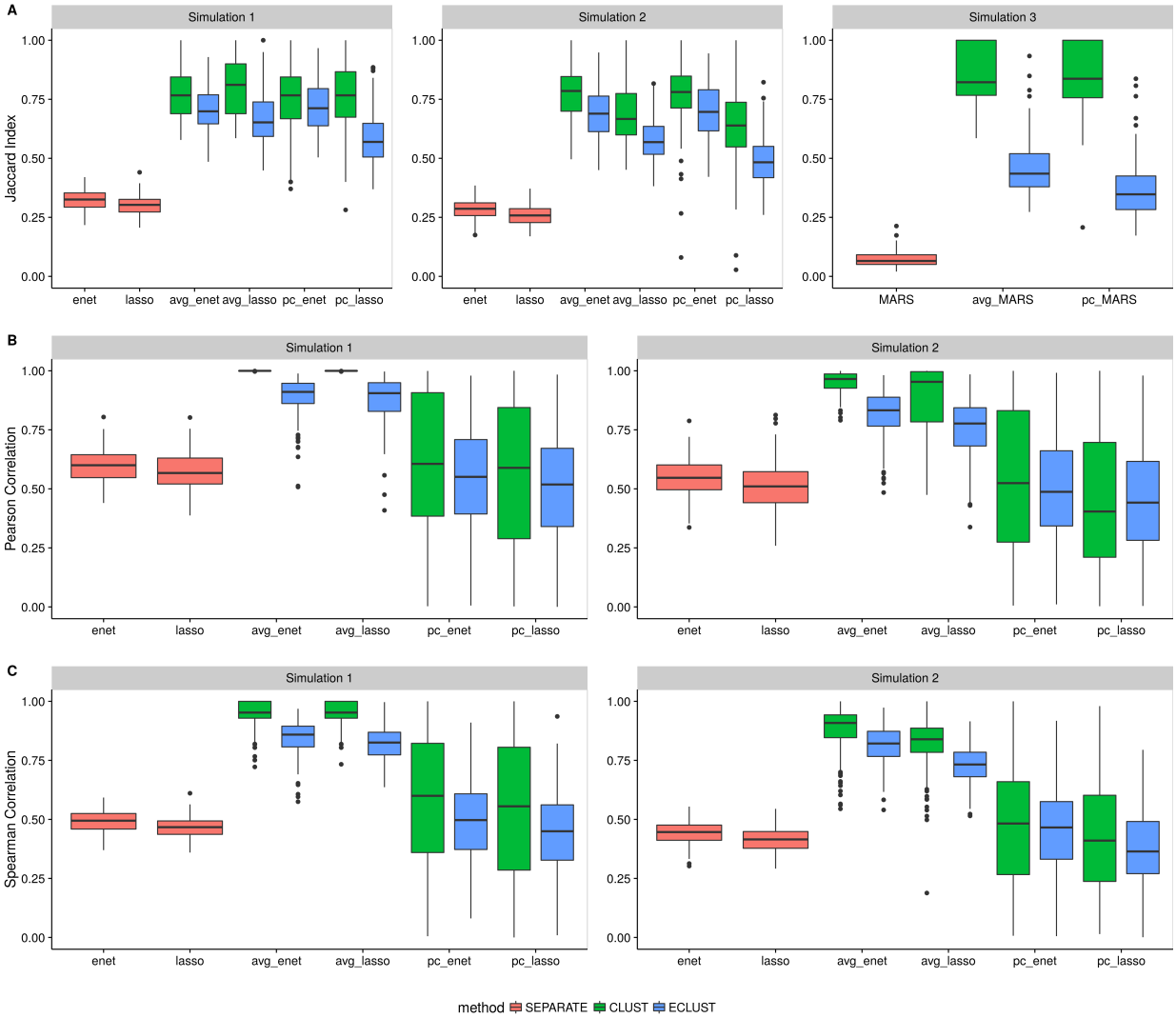


Figure 6: Stability results in simulations. SEPARATE results are in pink, CLUST in green and ECLUST in blue.

4.1 NIH MRI Study of Normal Brain Development

The NIH MRI Study of Normal Brain Development, started in 2001, was a 7 year longitudinal multi-site project that used magnetic resonance technologies to characterize brain maturation in 433 medically healthy, psychiatrically normal children aged 4.5-18 years (Evans et al., 2006). The goal of this study was to provide researchers with a representative and reliable source of healthy control subject data as a basis for understanding atypical brain development associated with a variety of developmental, neurological, and neuropsychiatric disorders affecting children and adults. Brain imaging data (e.g. cortical surface thickness, intra-cranial volume), behavioural measures (e.g. IQ scores, psychiatric interviews, behavioral ratings) and demographics (e.g. socioeconomic status) were collected at two year intervals for three time points and are publically available upon request. Previous research using these data found that level of intelligence and age correlate with cortical thickness (Shaw et al., 2006; Khundrakpam et al., 2013), but to our knowledge no such relation between income and cortical thickness has been observed. We therefore used this data to see the performance of ECLUST in the presence (age) and absence (income)

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of an effect on the correlations in the HD data. We analyzed the 10,000 most variable regions on the cortical surface from brain scans corresponding to the first sampled time point only. We used binary age (166 age ≤ 11.3 and 172 > 11.3) and binary income (142 high and 133 low income) indicator as the environment variables and standardized IQ scores as the response. We identified 22 clusters from $TOM(X_{\text{all}})$ and 57 clusters from $TOM(X_{\text{diff}})$ when using age as the environment, and 86 clusters from $TOM(X_{\text{all}})$ and 49 clusters from $TOM(X_{\text{diff}})$ when using income as the environment. Results are shown in Figure 7, panels C and D. The method which uses all individual variables as predictors (pink), has better R^2 but also worse RMSE compared to CLUST and ECLUST, likely due to over-fitting. There is a slight benefit in performance for ECLUST over CLUST when using age as the environment (panel D). Importantly, we observe very similar performance between CLUST and ECLUST across all models (panel C), suggesting very little impact on the prediction performance when including features derived both with and without the E variable, in a situation where they are unlikely to be relevant.

4.2 Gene Expression Study of Ovarian Cancer

Differences in gene expression profiles have led to the identification of robust molecular subtypes of ovarian cancer; these are of biological and clinical importance because they have been shown to correlate with overall survival (Tothill et al., 2008). Improving prediction of survival time based on gene expression signatures can lead to targeted therapeutic interventions (Helland et al., 2011). The proposed ECLUST algorithm was applied to gene expression data from 511 ovarian cancer patients profiled by the Affymetrix Human Genome U133A 2.0 Array. The data were obtained from the TCGA Research Network: http://cancergenome.nih.gov/ and downloaded via the TCGA2STAT R library (Wan et al., 2015). Using the 881 signature genes from Helland et al. (2011) we grouped subjects into two groups based on the results in this paper, to create a "positive control" environmental variable expected to have a strong effect. Specifically, we defined an environment variable in our framework as: $E = 0$ for subtypes C1 and C2 ($n = 253$), and $E = 1$ for subtypes C4 and C5 ($n = 258$). Overall survival time (log transformed) was used as the response variable. Since these genes were ascertained on survival time, we expected the method using all genes without clustering to have the best performance, and hence one goal of this analysis was to see if ECLUST performed significantly worse as a result of summarizing the data into a lower dimension. We found 3 clusters from $TOM(X_{\text{all}})$ and 3 clusters from $TOM(X_{\text{diff}})$; results are shown in Figure 7, panel C. Across all models, ECLUST performs slightly better than CLUST. Furthermore it performs almost as well as the separate variable method, with the added advantage of dealing with a much smaller number of predictors (881 with SEPARATE compared to 6 with ECLUST).

4.3 Gestational diabetes, epigenetics and metabolic disease

Events during pregnancy are suspected to play a role in childhood obesity development but only little is known about the mechanisms involved. Indeed, children born to women who had GDM in pregnancy are more likely to be overweight and obese (Wendland et al., 2012), and evidence suggests epigenetic factors are important piece of the puzzle (Bouchard et al., 2010, 2012). Recently, methylation changes in placenta and cord

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

blood were associated with GDM (Ruchat et al., 2013), and here we explore how these changes are associated with obesity in the children at the age of about 5 years old. DNA methylation in placenta was measured with the Infinium HumanMethylation450 BeadChip (Illumina, Inc (Bibikova et al., 2011)) microarray, in a sample of 28 women, 20 of whom had a GDM-affected pregnancy, and here, we used GDM status as our E variable, assuming that this has widespread effects on DNA methylation and on its correlation patterns. Our response, Y, is the standardized body mass index (BMI) in the offspring at the age of 5. In contrast to the previous two examples, here we had no particular expectation of how ECLUST would perform.

Using the 10,000 most variable probes, we found 2 clusters from $TOM(X_{\text{all}})$, and 75 clusters from $TOM(X_{\text{diff}})$. The predictive model results from a MARS analysis are shown in Figure 7, panel A. When using R^2 as the measure of performance, ECLUST outperforms both SEPARATE and CLUST methods. When using RMSE as the measure of model performance, performance tended to be better with CLUST rather than ECLUST perhaps in part due to the small number of clusters derived from $TOM(X_{\text{all}})$ relative to $TOM(X_{\text{diff}})$. Overall, the ECLUST algorithm with bagged MARS and the 1st PC of each cluster performed best, i.e., it had a better R^2 than CLUST with comparable RMSE. The sample size here is very small, and therefore the stability of the model fits is limited stability.

The probes in these clusters mapped to 164 genes and these genes were selected to conduct pathway analyses using the Ingenuity Pathway Analysis (IPA) software (Ingenuity System). IPA compares the selected genes to a reference list of genes included in many biological pathways using a hypergeometric test. Smaller p values are evidence for over-represented gene ontology categories in the input gene list. The results are summarized in Table 4 and provide some biological validation of our ECLUST method. For example, the Hepatic system is involved with the metabolism of glucose and lipids (Saltiel and Kahn, 2001), and behavior and neurodevelopment are associated with obesity (Epstein et al., 2004). Furthermore, it is interesting that embryonic and organ development pathways are involved since GDM is associated with macrosomia (Ehrenberg et al., 2004).

5 Discussion

The challenge of precision medicine is to appropriately fit treatments or recommendations to each individual. Data such as gene expression, DNA methylation levels, or magnetic resonance imaging (MRI) signals are examples of HD measurements that capture multiple aspects of how a tissue is functioning. These data often show patterns associated with disease, and major investments are being made in the genomics research community to generate such HD data. Analytic tools increasing prediction accuracy are needed to maximize the productivity of these investments. However, the effects of exposures have usually been overlooked, but these are crucial since they can lead to ways to intervene. Hence, it is essential to have a clear understanding of how exposures modify HD measures, and how the combination leads to disease. Existing methods for prediction (of disease), that are based on HD data and interactions with exposures, fall far short of being able to obtain this clear understanding. Most methods have low power and poor interpretability, and furthermore, modelling and interpretation problems are exacerbated when there is

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

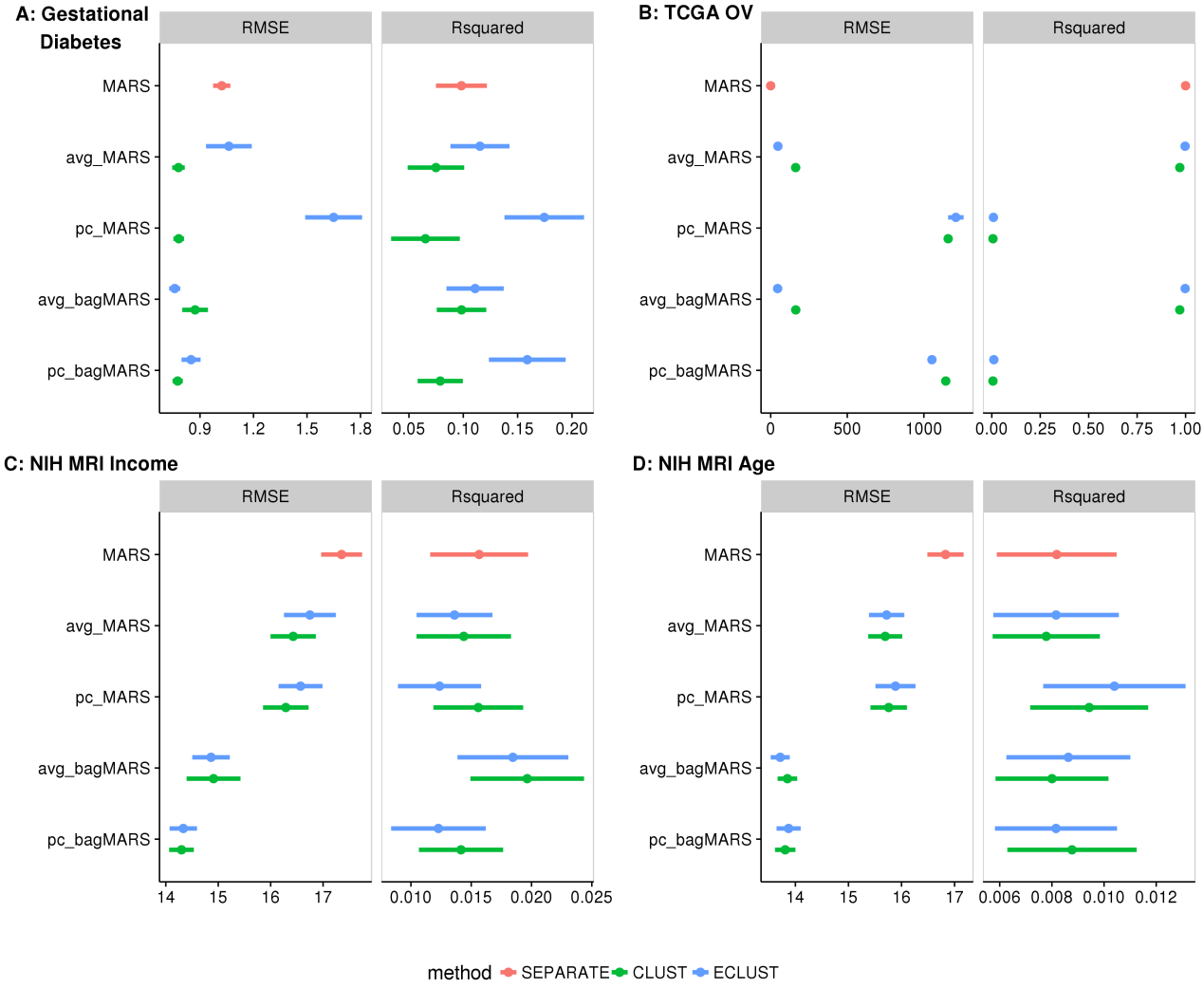


Figure 7: Model fit measures from analysis of three data sets: (A) Gestational diabetes birth-cohort (B) TCGA Ovarian Cancer study (C) NIH MRI Study with income as the environment variable (D) NIH MRI Study with age as the environment variable

interest in interactions. In general, power to estimate interactions is low, and the number of possible interactions could be enormous.

Therefore, here we have proposed a strategy to leverage situations where a covariate (e.g. an exposure) has a wide-spread effect on one or more HD measures, e.g. GDM on methylation levels. We have shown that this expected pattern can be used to construct dimension-reduced predictor variables that inherently capture the systemic covariate effects. These dimension-reduced variables, constructed without using the phenotype, can then be used in predictive models of any type. In contrast to some common analysis strategies that model the effects of individual predictors on outcome, our approach makes a step towards a systems-based perspective that we believe will be more informative when exploring the factors that are associated with disease or a phenotype of interest. We have shown, through simulations and real data analysis, that incorporation of environmental factors into predictive models in a way that retains a high dimensional perspective can improve results and interpretation for both linear and non linear effects.

We proposed two key methodological steps necessary to maximize predictive model in-

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 4: Ingenuity Pathway Analysis Results – top-ranked diseases and disorders, and physiological system development and function epigentically affected by gestational diabetes mellitus and associated with childhood body mass index

Category	Name	p values	n^a
Diseases and Disorders	Hepatic System Disease	9.61e−7 – 5.17e−7	75
Physiological System Development and Function			
	Behavior	1.35e−2 – 7.82e−8	33
	Embryonic Development	1.35e−2 – 2.63e−8	26
	Nervous System Development and Function	1.35e−2 – 2.63e−8	43
	Organ Development	1.35e−2 – 2.63e−8	20
	Organismal Development	1.35e−2 – 2.63e−8	34

[a] number of genes involved in each pathway

terpretability when using HD data and a binary exposure: (1) dimension reduction of HD data built on exposure sensitivity, and (2) implementation of penalized prediction models. In the first step, we proposed to identify exposure-sensitive HD pairs by contrasting the TOM between exposed and unexposed individuals; then we cluster the elements in these HD pairs to find exposure-sensitive co-regulated sets. New dimension-reduced variables that capture exposure-sensitive features (e.g. the first principal component of each cluster) were then defined. In the second step we implemented linear and non-linear variable selection methods using the dimension-reduced variables to ensure stability of the predictive model. The ECLUST method has been implemented in the `eclust` (Bhatnagar) R package publicly available on CRAN. Our method along with computationally efficient algorithms, allows for the analysis of up to 10,000 variables at a time on a laptop computer.

The methods that we have proposed here are currently only applicable when three data elements are available. Specifically a binary environmental exposure, a high dimensional dataset that can be affected by the exposure, and a single phenotype. When comparing the TOM and Pearson correlations as a measure of similarity, our simulations showed that the performance of ECLUST was worse with correlations. This speaks to the potential of developing a better measure than the difference of two matrices. For example, we are currently exploring ways in which to handle continuous exposures or multiple exposures. The best way to construct an exposure-sensitive distance matrix that can be used for clustering is not obvious in these situations. One possible solution relies on a non-parametric smoothing based approach where weighted correlations are calculated. These weights can be derived from a kernel-based summary of the exposure covariates (e.g. (Qiu et al., 2016)). Then, contrasting unweighted and weighted matrices will allow construction of covariate-sensitive clusters.

The choice of summary measure for each cluster also warrants further study. While principal components and averages are well understood and easy to implement, the main shortcoming is that they involve all original variables in the group. As the size of the groups

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

increase, the interpretability of these measures decreases. Non-negative matrix factorization (Lee and Seung, 2001) and sparse principal component analysis (SPCA) (Witten et al., 2009) are alternatives which find sparse and potentially interpretable factors. Furthermore, structured SPCA (Jenatton et al., 2009) goes beyond restricting the cardinality of the contributing factors by imposing some a priori structural constraints deemed relevant to model the data at hand.

We are all aware that our exposures and environments impact our health and risks of disease, however detecting how the environment acts is extremely difficult. Furthermore, it is very challenging to develop reliable and understandable ways of predicting the risk of disease in individuals, based on high dimensional data such as genomic or imaging measures, and this challenge is exacerbated when there are environmental exposures that lead to many subtle alterations in the genomic measurements. Hence, we have developed an algorithm and an easy-to use software package to transform analysis of how environmental exposures impact human health, through an innovative signal-extracting approach for high dimensional measurements. Evidently, the model fitting here is performed using existing methods; our goal is to illustrate the potential of improved dimension reduction in two-stage methods, in order to generate discussion and new perspectives. If such an approach can lead to more interpretable results that identify gene-environment interactions and their effects on diseases and traits, the resulting understanding of how exposures influence the high-volume measurements now available in precision medicine will have important implications for health management and drug discovery.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

SRB and CMTG conceived the idea. SRB performed the analysis. BK and LB provided the data and interpretations. SRB and CMTG wrote a draft of the manuscript then all authors edited, read and approved the final manuscript.

Acknowledgements

SRB was supported by the Ludmer Centre for Neuroinformatics and Mental Health.

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bibliography

Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014. 2

Maurice Kendall. *A Course in Multivariate analysis*. London: Griffin, 1957. 2

Trevor Hastie, Robert Tibshirani, David Botstein, and Patrick Brown. Supervised harvesting of expression trees. *Genome Biology*, 2(1):1–0003, 2001. 2

Mee Young Park, Trevor Hastie, and Robert Tibshirani. Averaged gene expressions for regression. *Biostatistics*, 8(2):212–227, 2007. 2, 7

Peter Bühlmann, Philipp Rütimann, Sara van de Geer, and Cun-Hui Zhang. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835–1858, 2013. 2, 7

Stephen Reid and Robert Tibshirani. Sparse regression and marginal testing using cluster prototypes. *Biostatistics*, 17(2):364–376, 2016. 2

Daniel Müllner. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software*, 53(9):1–18, 2013. URL http://www.jstatsoft.org/v53/i09/. 2

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010. 2, 8

Yi Yang and Hui Zou. gglasso: Group lasso penalized learning using a unified bmd algorithm. 2014. URL http://CRAN.R-project.org/package=gglasso. R package version 1.3. 2

Max Kuhn. Caret package. *Journal of Statistical Software*, 28(5), 2008. 2, 16

J Fah Sathirapongsasuti. *COPDSexualDimorphism.data: Data to support sexually dimorphic and COPD differential analysis for gene expression and methylation.*, 2013. R package version 1.4.0. 2, 3

Budhachandra S Khundrakpam, Andrew Reid, Jens Brauer, Felix Carbonell, John Lewis, Stephanie Ameis, Sherif Karama, Junki Lee, Zhang Chen, Samir Das, et al. Developmental changes in organization of structural brain networks. *Cerebral Cortex*, 23(9): 2072–2085, 2013. 3, 17

Dennis Kostka and Rainer Spang. Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, 20(suppl 1):i194–i199, 2004. 3

Kathleen Oros Klein, Karim Oualkacha, Marie-Hélène Lafond, Sahir Bhatnagar, Patricia N Tonin, and Celia MT Greenwood. Gene coexpression analyses differentiate networks associated with diverse cancers harboring tp53 missense or null mutations. *Frontiers in Genetics*, 7, 2016. 3, 7

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

João Ricardo Sato, Marcelo Queiroz Hoexter, Pedro Paulo de Magalhães Oliveira, Michael John Brammer, Declan Murphy, Christine Ecker, MRC AIMS Consortium, et al. Inter-regional cortical thickness correlations are associated with autistic symptoms: a machine-learning approach. *Journal of psychiatric research*, 47(4):453–459, 2013. 6

Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005. 6, 27

Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586): 1551–1555, 2002. 6, 27

Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719–720, 2008. 7, 8

Peter Langfelder and Steve Horvath. Eigengene networks for studying the relationships between co-expression modules. *BMC systems biology*, 1(1):54, 2007. 7

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 7, 11

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. 7, 11

Jerome H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991. 7, 11

Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1, 2008. 8

Peter Langfelder, Bin Zhang, and with contributions from Steve Horvath. *dynamicTreeCut: Methods for Detection of Clusters in Hierarchical Clustering Dendrograms*, 2016. URL https://CRAN.R-project.org/package=dynamicTreeCut. R package version 1.63-1. 8

S. Milborrow. Derived from mda:mars by T. Hastie and R. Tibshirani. *earth: Multivariate Adaptive Regression Splines*, 2011. URL http://CRAN.R-project.org/package=earth. R package. 8

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/. 8

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011. 13

Laura Toloşi and Thomas Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011. 13

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Daniela M Witten, Ali Shojaie, and Fan Zhang. The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics*, 56(1):112–122, 2014. 14

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1):95–116, 2007. 14

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, pages 240–242, 1895. 14

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904. 14

Paul Jaccard. The distribution of the flora in the alpine zone. *New phytologist*, 11(2): 37–50, 1912. 14

Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983. 16

Tarald O Kvålseth. Cautionary note about r 2. *The American Statistician*, 39(4):279–285, 1985. 16

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 16

Alan C Evans, Brain Development Cooperative Group, et al. The nih mri study of normal brain development. *Neuroimage*, 30(1):184–202, 2006. 17

Philip Shaw, Deanna Greenstein, Jason Lerch, Liv Clasen, Rhoshel Lenroot, N et al Gogtay, Alan Evans, J Rapoport, and J Giedd. Intellectual ability and cortical development in children and adolescents. *Nature*, 440(7084):676–679, 2006. 17

Richard W Tothill, Anna V Tinker, Joshy George, Robert Brown, Stephen B Fox, Stephen Lade, Daryl S Johnson, Melanie K Trivett, Dariush Etemadmoghadam, Bianca Locandro, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, 14(16):5198–5208, 2008. 18

Åslaug Helland, Michael S Anglesio, Joshy George, Prue A Cowin, Cameron N Johnstone, Colin M House, Karen E Sheppard, Dariush Etemadmoghadam, Nataliya Melnyk, Anil K Rustgi, et al. Deregulation of mycn, lin28b and let7 in a molecular subtype of aggressive high-grade serous ovarian cancers. *PloS one*, 6(4):e18064, 2011. 18

Ying-Wooi Wan, Genevera I. Allen, Matthew L. Anderson, and Zhandong Liu. *TCGA2STAT: Simple TCGA Data Access for Integrated Statistical Analysis in R*, 2015. URL https://CRAN.R-project.org/package=TCGA2STAT. R package version 1.2. 18

Eliana M Wendland, Maria Regina Torloni, Maicon Falavigna, Janet Trujillo, Maria Alice Dode, Maria Amélia Campos, Bruce B Duncan, and Maria Inês Schmidt. Gestational diabetes and pregnancy outcomes-a systematic review of the world health organization (who) and the international association of diabetes in pregnancy study groups (iadpsg) diagnostic criteria. *BMC pregnancy and childbirth*, 12(1):1, 2012. 18

bioRxiv preprint doi: https://doi.org/10.1101/102475; this version posted January 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Luigi Bouchard, Stéphanie Thibault, Simon-Pierre Guay, Marta Santure, Alexandre Monpetit, Julie St-Pierre, Patrice Perron, and Diane Brisson. Leptin gene epigenetic adaptation to impaired glucose metabolism during pregnancy. *Diabetes care*, 33(11):2436–2441, 2010. 18

Luigi Bouchard, Marie-France Hivert, Simon-Pierre Guay, Julie St-Pierre, Patrice Perron, and Diane Brisson. Placental adiponectin gene dna methylation levels are associated with mothers' blood glucose concentration. *Diabetes*, 61(5):1272–1280, 2012. 18

Stephanie-May Ruchat, Andrée-Anne Houde, Grégory Voisin, Julie St-Pierre, Patrice Perron, Jean-Patrice Baillargeon, Daniel Gaudet, Marie-France Hivert, Diane Brisson, and Luigi Bouchard. Gestational diabetes mellitus epigenetically affects genes predominantly involved in metabolic diseases. *Epigenetics*, 8(9):935–943, 2013. 19

Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, et al. High density dna methylation array with single cpg site resolution. *Genomics*, 98(4):288–295, 2011. 19

Alan R Saltiel and C Ronald Kahn. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature*, 414(6865):799–806, 2001. 19

Leonard H Epstein, Rocco A Paluch, Colleen K Kilanowski, and Hollie A Raynor. The effect of reinforcement or stimulus control to reduce sedentary behavior in the treatment of pediatric obesity. *Health Psychology*, 23(4):371, 2004. 19

Hugh M Ehrenberg, Brian M Mercer, and Patrick M Catalano. The influence of obesity and diabetes on the prevalence of macrosomia. *American journal of obstetrics and gynecology*, 191(3):964–968, 2004. 19

Sahir Rai Bhatnagar. *eclust: Environment Based Clustering for Interpretable Predictive Models in High Dimensional Data*. URL https://cran.r-project.org/package=eclust. R package version 0.1.0. 21

Huitong Qiu, Fang Han, Han Liu, and Brian Caffo. Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(2):487–504, 2016. 21

Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001. 22

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009. 22

Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Structured sparse principal component analysis. *arXiv preprint arXiv:0909.1440*, 2009. 22