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

Sahir Bhatnagar, Yi Yang, Mathieu Blanchette, Luigi Bouchard, (others), Celia MT Greenwood

October 20, 2016

(Incomplete) 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, there are several applications where interaction models can reflect biological phenomena and improve statistical power.

General Comments

Check when abbreviations are first defined and used.

1 Introduction

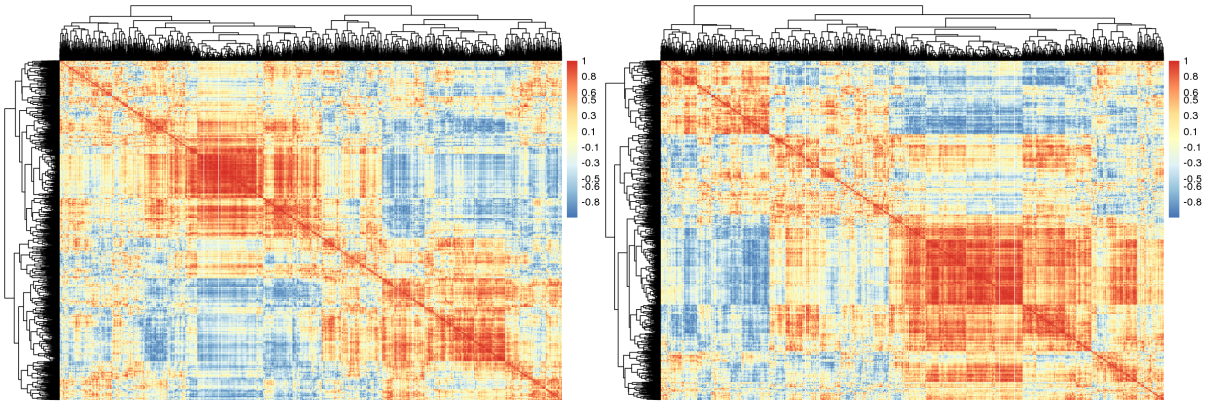
In this article, we consider the prediction of an outcome variable y observed on n individuals from p variables, where p is much larger than n. In addition to problems related to overfitting, it can be very challenging to interpret the results of prediction models with ultra-high dimensional predictor 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.

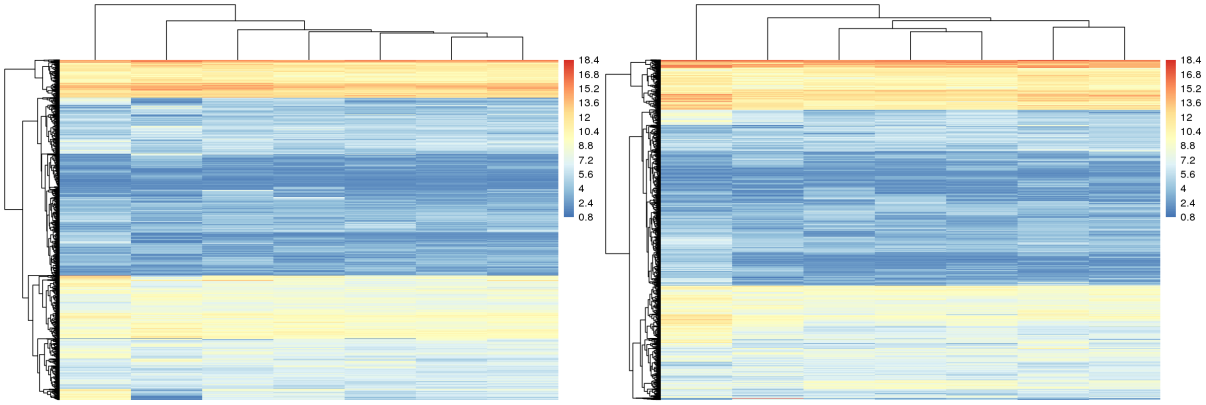
There are several advantages to these two-step methods. Through the reduction of the dimension of the model, the results are often more stable with smaller prediction variance, and through identification of sets of correlated variables, the resulting clusters can provide an easier route to interpretation. 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 from a study of Chronic Obstructive Pulmonary Disease (COPD) (Sathirapongsasuti, 2013) in Figure 1. This study measured gene expression in 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 1d, 1c, 1b and 1a 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 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.

Why did you use income in the pediatric cortical thickness study? here you discuss age

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 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 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 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:

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

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 second term in (3) resembles an interaction model, with $\beta_2 \psi$ being the interaction parameter.

For this second illustration and non linearity: Suppose

$$Y \sim \beta_1 X_1 + \beta_2 X_2 + \beta_C * cor(X_1, X_2) \tag{4}$$

Now substitute for X_2 using equation 1. You should be able to work out the correlation from equation 1. Then what does the equation for Y look like? I think it will be nonlinear.

Furthermore, nonlinear models arise easily from environment-dependent effects on correlations. Suppose there are non-linear effects of $\mathbf{X}$ on the response in a given environment, for example

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 (X_{i1} - c1)_+ \cdot (X_{i2} - c2)_+ \cdot E_i + \varepsilon \tag{5}$$

where $c1, c2$ are constants, and

$$(x - t)_+ = \begin{cases} x - t, & \text{if } x > t \\ 0, & \text{else} \end{cases}$$

A simple illustration of nonlinearity here may be enough. the other text below could move to the introduction or to the motivation of simulation 3

The interaction term in (5) is a product of hinge functions used in multivariate adaptive regression spline (MARS) models (Friedman et al., 2001). Equation (5) is attempting to model the scenario when an entire pathway (e.g. genes X_1 and X_2) becomes disregulated for a given exposure, which has a drastic effect on the response. Indeed, large-scale genomic studies are beginning to show that diseases are a result of networks whose states are affected by a complex interaction of genetic and environmental factors (Schadt, 2009). There is growing evidence to suggest that it is a whole network of genes that combine to influence disease risk (Chen et al., 2008), rather than a small number of them. This speaks to the fact that cellular functions are carried out through a concerted effort of many genes (Segal et al., 2003).

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, 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.

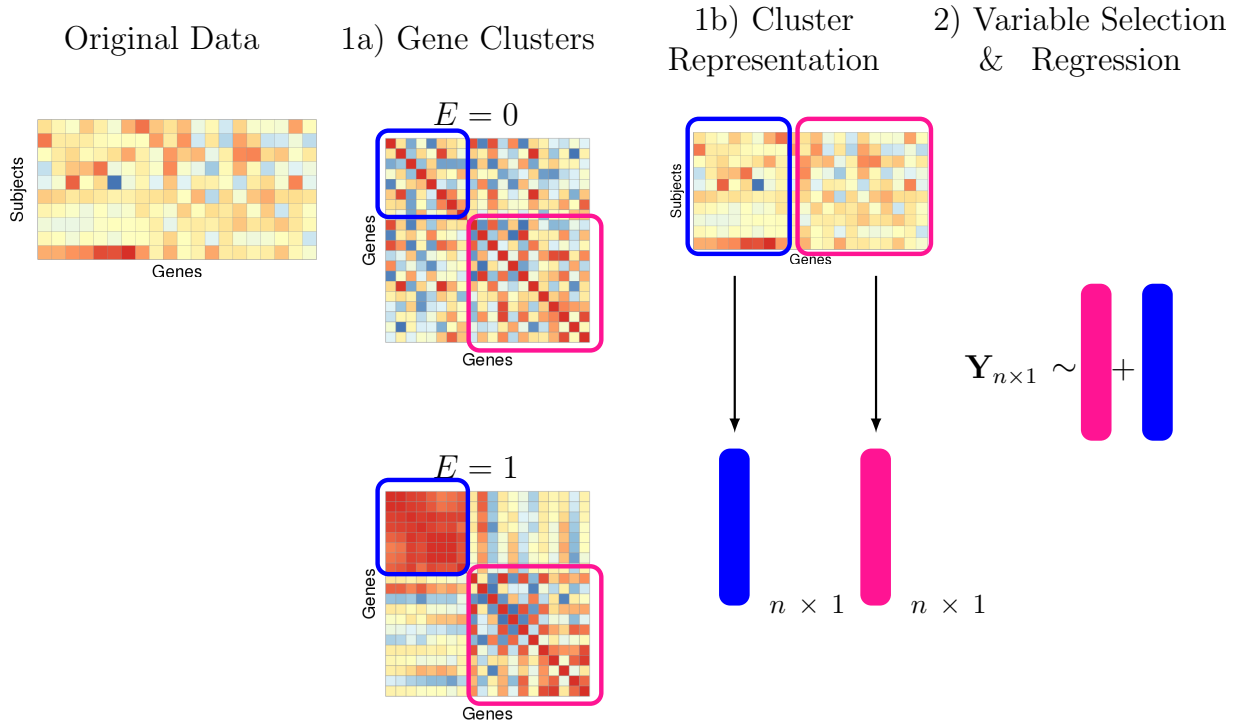


Figure 2: Overview of our proposed method

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 to capture connectedness between two X variables within each environmental condition.

Motivation here to restrict to only the TOM needs strengthening. I think I liked it better when correlations were here as well. This section is kind of wordy for Methods. Can it be more brief and concise?

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 TOM separately for exposed ($E = 1$) and unexposed ($E = 0$) individuals (see Supplemental methods for details on the TOM). We then seek to identify nodes that are very *different* between environments under the assumption that these nodes will be associated with the phenotype. We determine differential coexpression 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. Do you have a reference that discusses many different ways to summarize clusters? If so, it could be cited here, then we say we pick these 2. 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

Try to simplify text here to say (1) we use penalized methods for prediction, specifically methods xx, yy, zz, since even after dimension reduction the number of predictors is large

(2) do you really need equation 6? (3) The primary comparison is models with X_all only versus models with X_{all} and X_{diff}. (4) Given the context of either the simulation or the data set, sometimes we use linear models and someimtes non linear modelsl.

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. Therefore we propose to regress the response on the summary measures via penalized likelihood. These methods are able to simultaneously estimate and select regression parameters. 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.

Consider a regression model for an outcome variable $Y = (Y_1, \ldots, Y_n)$ with normally distributed errors ε. Let $E = (E_1, \ldots, E_n)$ be the binary environment vector and $\mathbf{X} = (X_1, \ldots, X_p)$ be the matrix of high-dimensional data. Consider the regression model with main effects and their interactions with E:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_E E + \alpha_1 (X_1 E) + \cdots + \alpha_p (X_p E) + \varepsilon \tag{6}$$

Our goal is to estimate the parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p, \beta_E) \in \mathbb{R}^{p+1}$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p) \in \mathbb{R}^p$ and to improve prediction of Y. In fact, in the light of our goals to improve prediction and interpretability, we also consider the related model

$$Y = \beta_0^* + \sum_{j=1}^{\ell} \beta_j^* \widetilde{X}_{C_j} + \beta_E^* E + \sum_{j=1}^{\ell} \alpha_j^* \left(\widetilde{X}_{C_j} E \right) + \varepsilon \tag{7}$$

where $\widetilde{X}_{C_j}$ are linear combinations of $\mathbf{X}$ (from Step 1b) designed to reduce the dimension, such that $\ell << p$, and the superscript asterisk on the parameters is just to emphasize that these are different from those in (6). In what follows, we omit the asterisk on the parameters for clarity. For non-linear effects we consider the MARS model (Friedman, 1991).

Our general approach, ECLUST, can therefore be summarized by the algorithm in Table 1.

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 $\|TOM_{E=1} - TOM_{E=0}\|$ 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).Unclear. What is the point here? 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 environment in the clustering step, and finally with ECLUST, which clusters both with and without considering the environment. A detailed description of the methods being compared is summarized in

Table 2. We have designed 3 simulation scenarios that are constructed to illustrate different kinds of relationships between the variables and the response. For all scenarios, we have 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 3 simulation scenarios.

All our simulations consider the case of a continuous response variable. Maybe should have a simulation with a binary response - a fourth?

Let

$$Y = Y^* + k \cdot \varepsilon \tag{8}$$

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

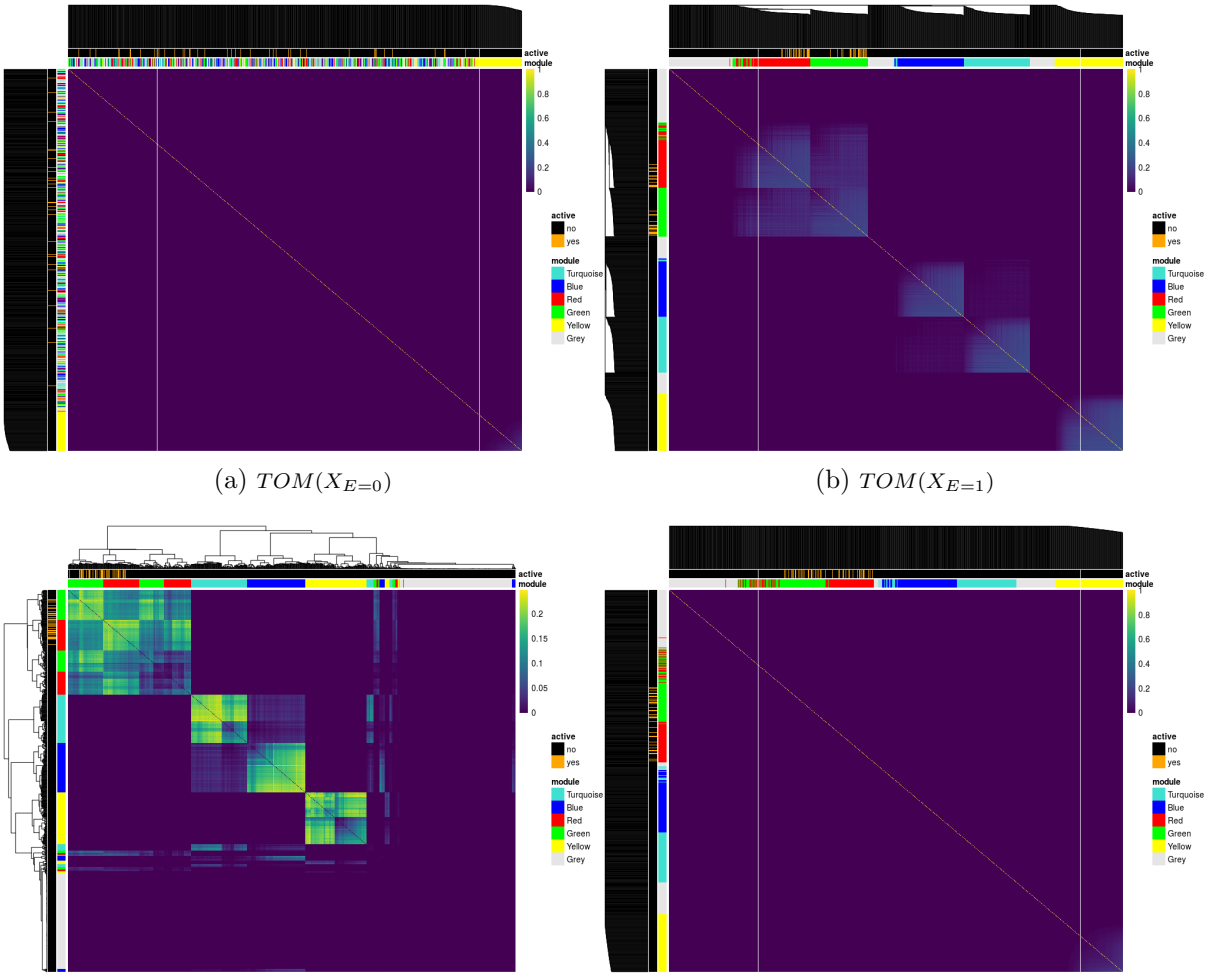
General Approach	Summary Measure of Feature Clusters	Description[a]
SEPARATE	NA	Regression of the original predictors $\{X_1, \ldots, X_p\}$ on the response i.e. no transformation of the predictors is being done here
CLUST	1st principal component, average	Create clusters of predictors without using the 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 $k < \ell < p$. 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

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 by colour (turquoise, blue, red, green and yellow), while the remaining 1250 variables were simulated as independent standard normal vectors (grey). 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. 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 green and red clusters were set to have an average correlation of $\rho = \{0.2, 0.9\}$

and the active set S_0 was distributed evenly between these two blocks. Heatmaps of the TOM for this environment dependent correlation structure are shown in Figure B.1 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 three simulation scenarios differ in how the linear predictor Y^* in (8) is defined and the regression models being used to fit the simulated data. In simulations 1 and 2 we use lasso (Tibshirani, 1996) and elasticnet (Zou and Hastie, 2005) to fit linear models, and use MARS (Friedman, 1991) in simulation 3 to estimate non-linear effects.

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

$$Y^* = \sum_{\substack{j\in\{1,\ldots,250\} \\ j\in \text{ red, green block}}} \beta_j X_j + \beta_E E \tag{9}$$

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

$$Y^* = \sum_{\substack{j\in\{1,\ldots,125\} \\ j\in \text{ red, green block}}} \beta_j X_j + \alpha_j X_j E + \beta_E E \tag{10}$$

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:

$$Y_i^* = \sum_{\substack{j\in\{1,\ldots,250\} \\ j\in \text{ red, green block}}} \beta_j X_{ij} + \beta_E E_i + \alpha_Q E_i \cdot f(Q_i) \tag{11}$$

where

$$Q_i = -\max_{\substack{j\in\{1,\ldots,250\}\\ j\in \text{ red, green block}}} \left(X_{ij} - \bar{X}_i\right)^2 \tag{12}$$

$$f(u_i) = \frac{u_i - \min\limits_{i\in\{1,\ldots,n\}} u_i}{-\min\limits_{i\in\{1,\ldots,n\}} u_i} \tag{13}$$

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

In all simulations? 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 important 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 (12) 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 (13) simply maps Q_i to the [0,1] range. In order to get an idea of the relationship in (11), 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 allow for the possibility of two way interactions between any of the variables.

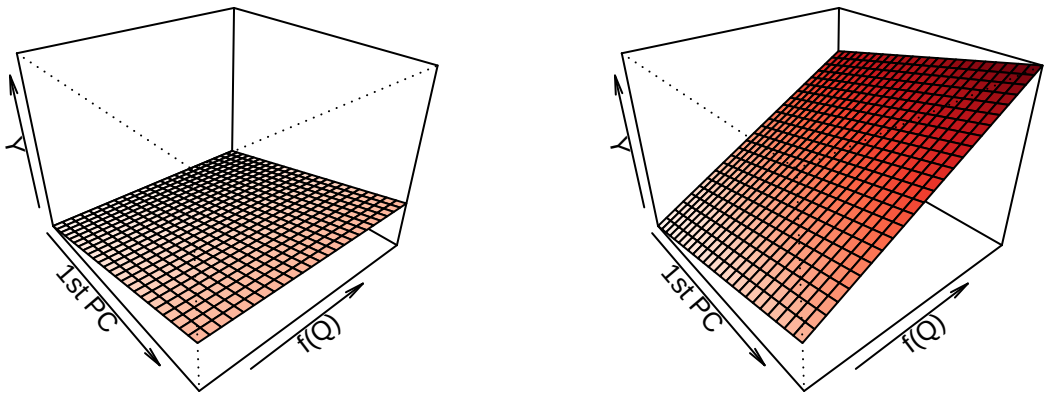


Figure 4: Visualization of the relationship between the response, the first principal component of the main effects and $f(Q_i)$ in (11) for $E = 0$ (left) and $E = 1$ (right) in simulation scenario 3.

Simulation 4

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

$\Pr(z = 1) = p$ and $\Pr(z = 0) = 1 - p$.

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 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 (Tolo̧si 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 correlation 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).

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)	$\lVert \mathbf{Y}_{test} - \widehat{\boldsymbol{\mu}}(\mathbf{X}_{test}) \rVert_2$
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 the three simulation scenarios for model fit (Figure 5) and feature stability (Figure 6). 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. 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 do a much better job at identifying the active variables, though the resulting models are not always sparse. The better performance of ECLUST over CLUST is noticeable as more points lie in the top left part of the plot. The discrete nature of the plots demonstrates the

stability of the clustering algorithm. ECLUST also does better than CLUST in correctly determining whether a feature is zero or nonzero (Figure 5, panel B). 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.

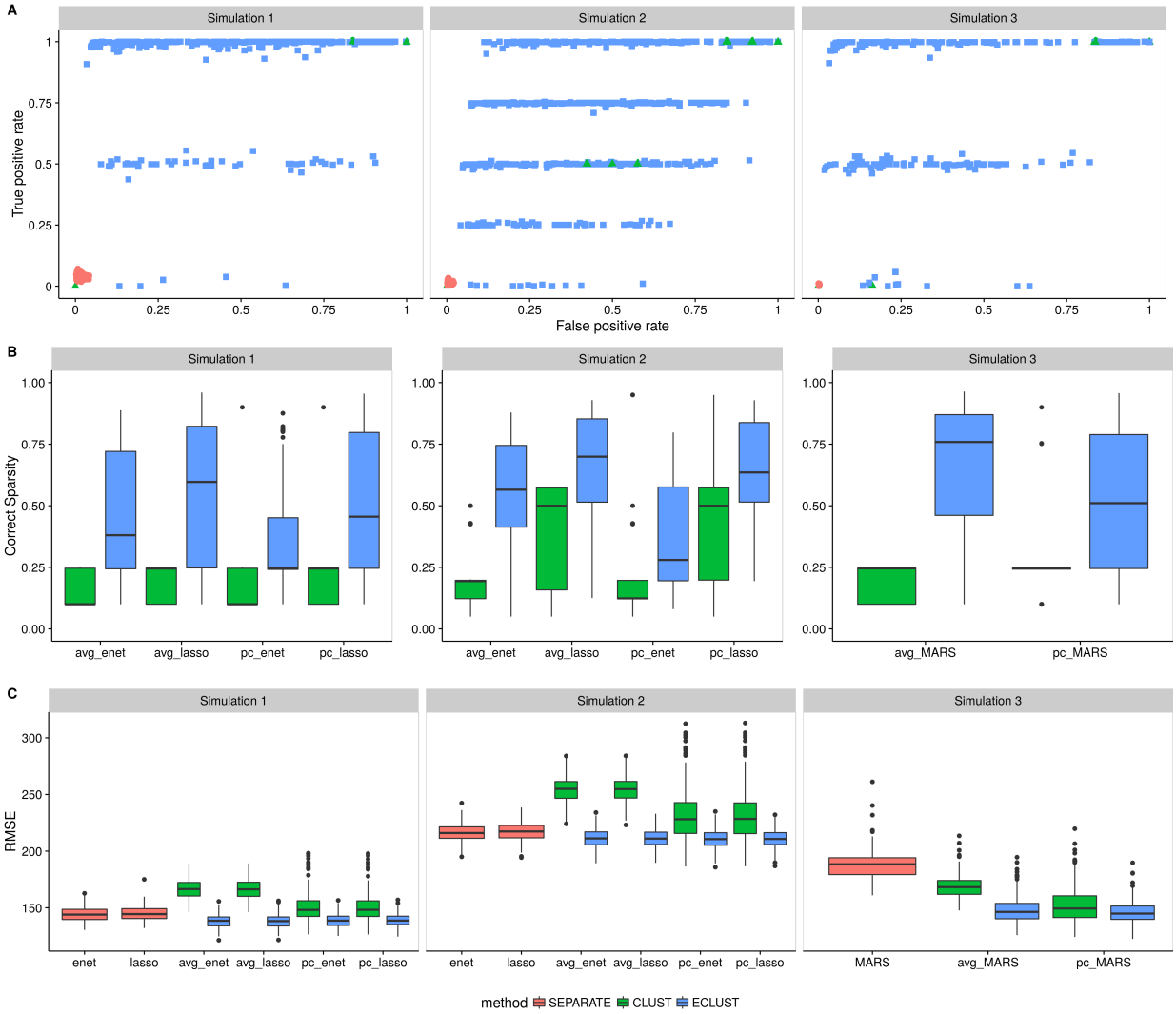


Figure 5: Model fit results in simulations

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. The CLUST approach does slightly better than ECLUST across all stability measures, but can exhibit much higher variability, particularly when looking at the agreement between the value and ranking of the estimated coefficients across CV folds (Figure 6, panel B and C). The number of clusters, and therefore the number of features in the regression model, tends to be much smaller in CLUST compared to ECLUST. This explains its poorer stability measures; there are more coefficients to estimate. Overall, we observe that both the model fit and stability measures are fairly indifferent to the choice of summary measure and penalization procedure. The complete results in the supplemental material show that these conclusions are not sensitive to the SNR, ρ or α_j.

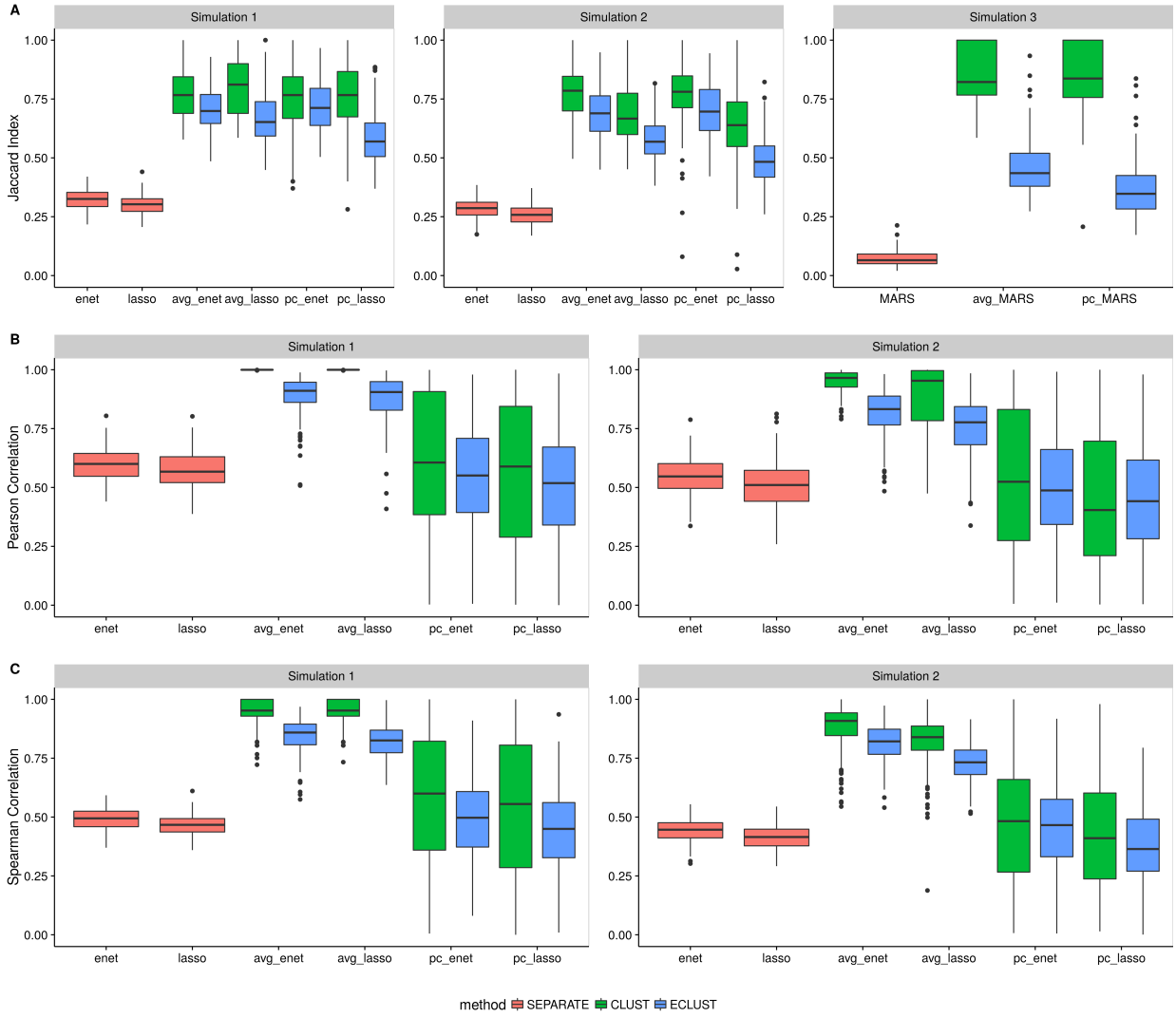


Figure 6: Stability results in simulations

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 one study, the investigators' goal is to examine the impact of gestational diabetes on childhood obesity in a sample of mother-child pairs from a prospective birth cohort. In our second data set, normal brain developmentis examined in conjunction with intelligence scores, and the third data and one that aims to identify molecular subtypes of ovarian cancer with gene expression data. The dataset contain a range of sample sizes, and both the amount of clustering in the HD data and the strength of the effect of the designated exposure variable both 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.

I changed the order.

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. Research on this data has 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 as a negative control for the ECLUST algorithm, i.e., to see its performance when there is little effect of the environment on the correlations in the HD data. We analyzed the 10,000 most variable regions on the cortical surface from 275 brain scans (corresponding to the first sampled time point only). We used a binary income indicator as the environment variable (142 high and 133 low income) and standardized IQ scores were the responses. We identified 86 clusters from $TOM(X_{\text{all}})$ and 49 clusters from $TOM(X_{\text{diff}})$. Results are shown in Figure 7, panel B. The method which uses all individual variables as predictors (pink), has better R^2 but also worse RMSE compared to CLUST and ECLUST. Importantly, we observe very similar performance between CLUST and ECLUST across all models, 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 however not enough is known about the mechanisms involved. However, it is well known that children born to women who had a gestational diabetes (GD) mellitus-affected pregnancy are more likely to be overweight and obese [refs], and evidence suggests epigenetic factors are important piece of the puzzle [refs]. Recently, methylation changes in placenta and cord blood were associated with GD (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 2011]) microarray, in a sample of 28 women, 20 of whom had a GD-affected pregnancy, and here, we used GD 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 in all 28 placentas, $TOM(X_{\text{all}})$, and 75 clusters from $TOM(X_{\text{diff}})$ can you add a phrase on how you ensured stability of the clusters?. The predictive model results from a MARS analysis are shown in Figure 7, panel A. When using R^2 as the measure of performance with the left out bootstrap samples with theh 0.632 estimator??? , 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. Any hypotheses why? 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.

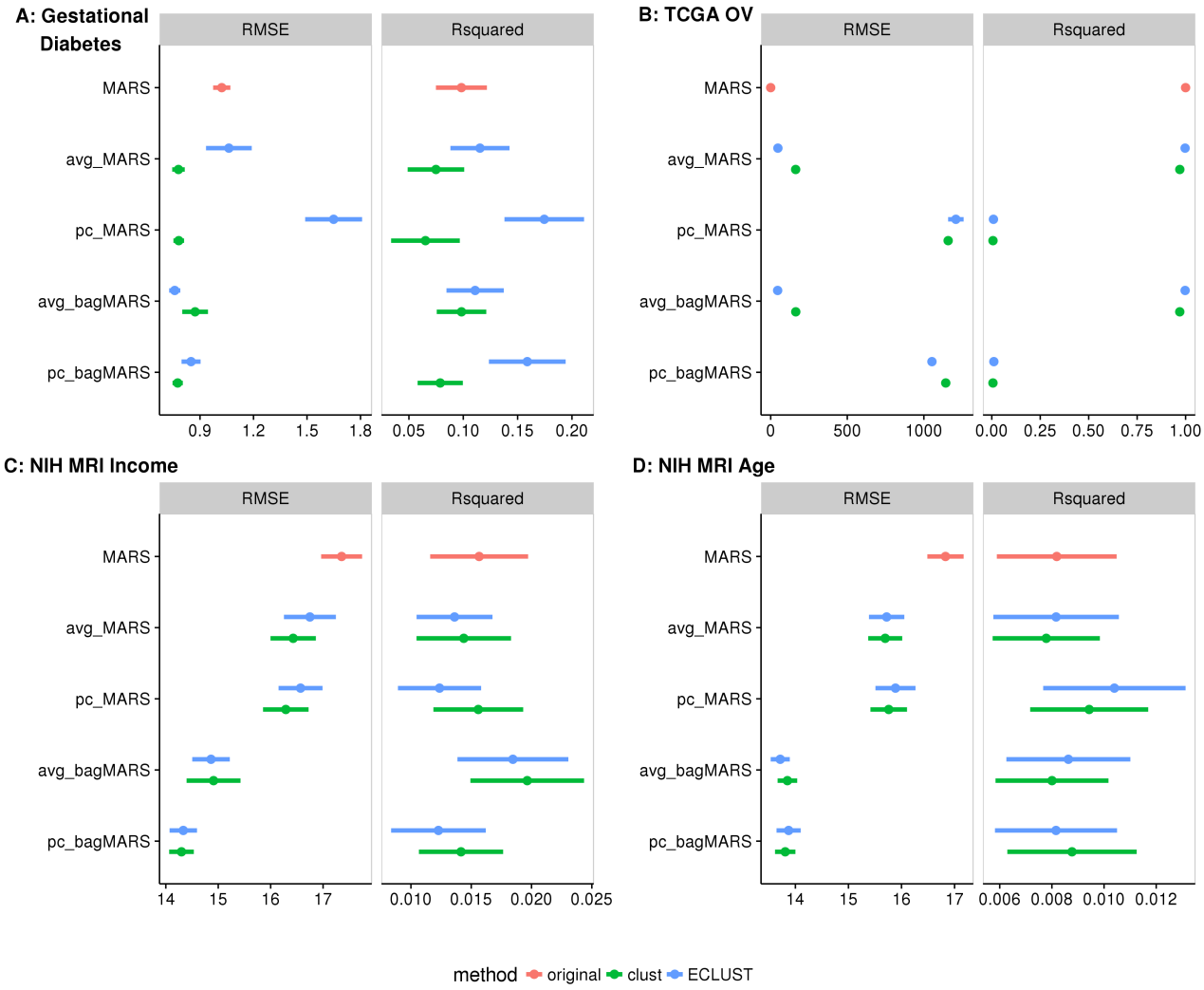


Figure 7: Model fit measures from analysis of three data sets: (A) Gestational diabetes birth-cohort (B) TCGA Ovarian Cancer study (C) NIH MRI Study with income as the environment variable (D) NIH MRI Study with age as the environment variable

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. . To leave this sentence here, it would need to be cited. But it may not belong here. For example, gestational diabetes has been associated with altered methylation levels in placenta and cord blood, and GD is also known to lead to increased risks of overweight in offspring leading to life-long risks of morbidity. Hence, it is essential to have a clear

understanding of how exposures modify HD measures, and how the combination leads to disease. However, 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 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. gestational diabetes 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.

If our measures of interpretation are poor, then it will be necessary to explore this more fully. i.e. can you actually measure whether the exposure-dependent clusters of variables are better identified?

We proposed two key methodological steps necessary to maximize predictive model interpretability 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 algorithm has been implemented in an R package `eclust` publicly available on GitHub at `https://github.com/sahirbhatnagar/eclust`.

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. We are currently exploring ways in which to handle continuous exposures or multiple exposures.

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 environmen-

tal 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.

Bibliography

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

Maurice Kendall. *A Course in Multivariate analysis*. London: Griffin, 1957. 1

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

Mee Young Park, Trevor Hastie, and Robert Tibshirani. Averaged gene expressions for regression. *Biostatistics*, 8(2):212–227, 2007. 1, 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. 1, 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, 9

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, 18

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, 18

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

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001. 5

Eric E Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–223, 2009. 5

Yanqing Chen, Jun Zhu, Pek Yee Lum, Xia Yang, Shirly Pinto, Douglas J MacNeil, Chunsheng Zhang, John Lamb, Stephen Edwards, Solveig K Sieberts, et al. Variations in dna elucidate molecular networks that cause disease. *Nature*, 452(7186):429–435, 2008. 5

Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34(2):166–176, 2003. 5

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 6, 12

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. 6, 12

Jerome H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991. 6, 8, 12

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. 7, 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, 9

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

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

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. 9

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. 9

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/. 9

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

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. 15

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. 15

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, pages 240–242, 1895. 15

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

Paul Jaccard. The distribution of the flora in the alpine zone. *New phytologist*, 11(2): 37–50, 1912. 15

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. 17

Tarald O Kvålseth. Cautionary note about r 2. *The American Statistician*, 39(4):279–285, 1985. 18

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 18

Alan C Evans, Brain Development Cooperative Group, et al. The nih mri study of normal brain development. *Neuroimage*, 30(1):184–202, 2006. 18

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. 18

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

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

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. 27

A Supplemental Methods

A.1 Description of Topological Overlap Matrix

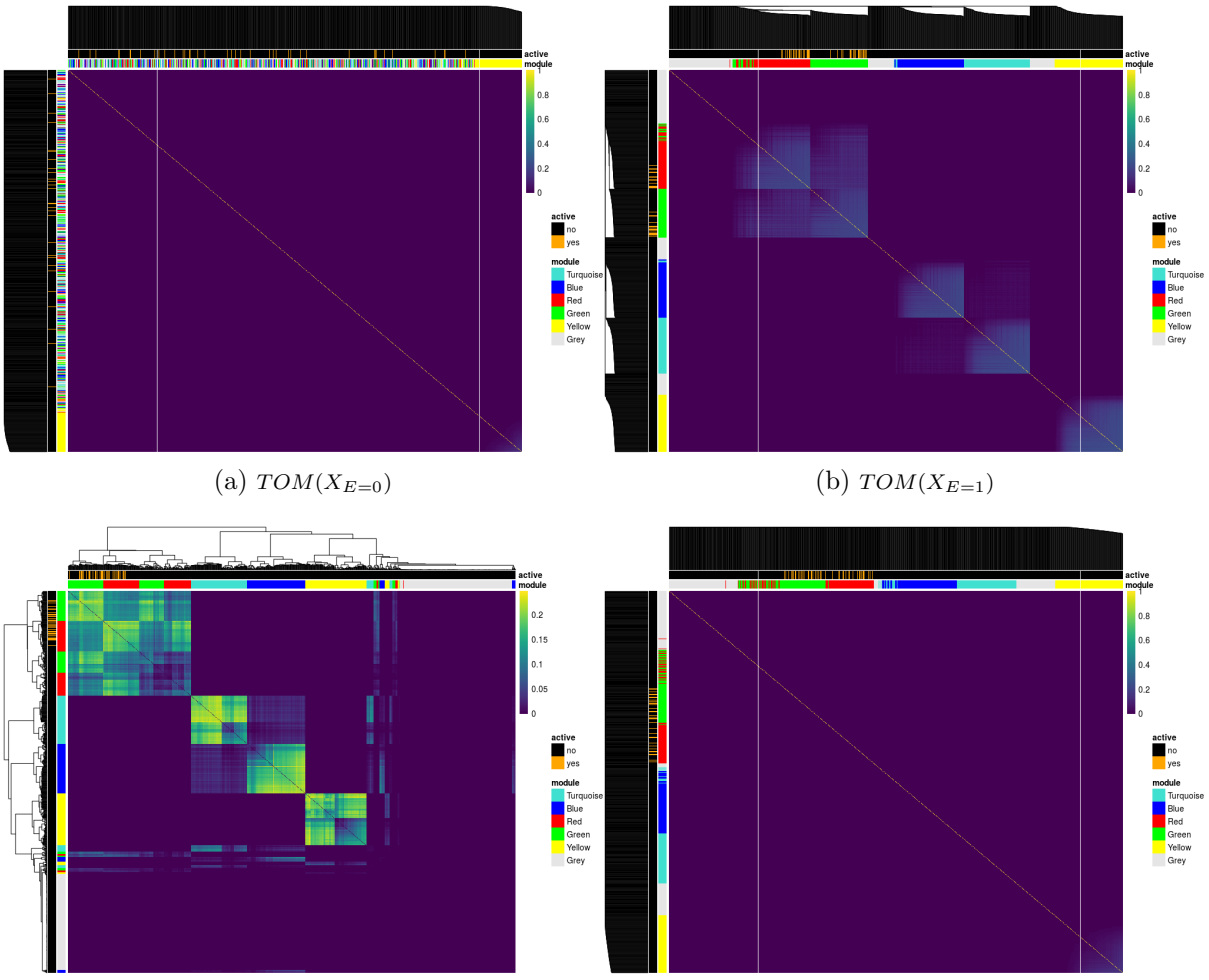
Starting with a similarity measure $s_{ij} = |cor(i, j)|$ between node i and node j, one could apply a hard threshold to determine if this pair is considered connected or not resulting in an un-weighted network (a matrix of 0's and 1's). Instead, Zhang and Horvath (Zhang and Horvath, 2005) propose a soft thresholding framework that assigns a connection weight to each gene pair using a power adjacency function $a_{ij} = |s_{ij}|^{\beta}$. The parameter β determines the sensitivity and specificity of the pairwise connection strengths e.g. a larger β will result in fewer connected nodes which can reduce noise in the network but can also eliminate signal if too large. A measure of similarity is then derived using the symmetric and non-negative topological overlap matrix (Ravasz et al., 2002) (TOM) $\Omega = [\omega_{ij}]$:

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{min\{k_i, k_j\} + 1 - a_{ij}} \tag{14}$$

where $l_{ij} = \sum_u a_{iu}a_{uj}$, $k_i = \sum_u a_{iu}$ is the node connectivity, and the index u runs across all nodes of the network. Basically, ω_{ij} is a measure of similarity in terms of the commonality of the nodes they connect to. If i and j are unconnected and do not share any neighbors then $\omega_{ij} = 0$. An $\omega_{ij} = 1$ means that i and j are connected, and the neighbors of the node with fewer connections are also neighbors of the other node.

B Visual Representation of Similarity Matrices

B.1 Topological Overlap Matrix



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

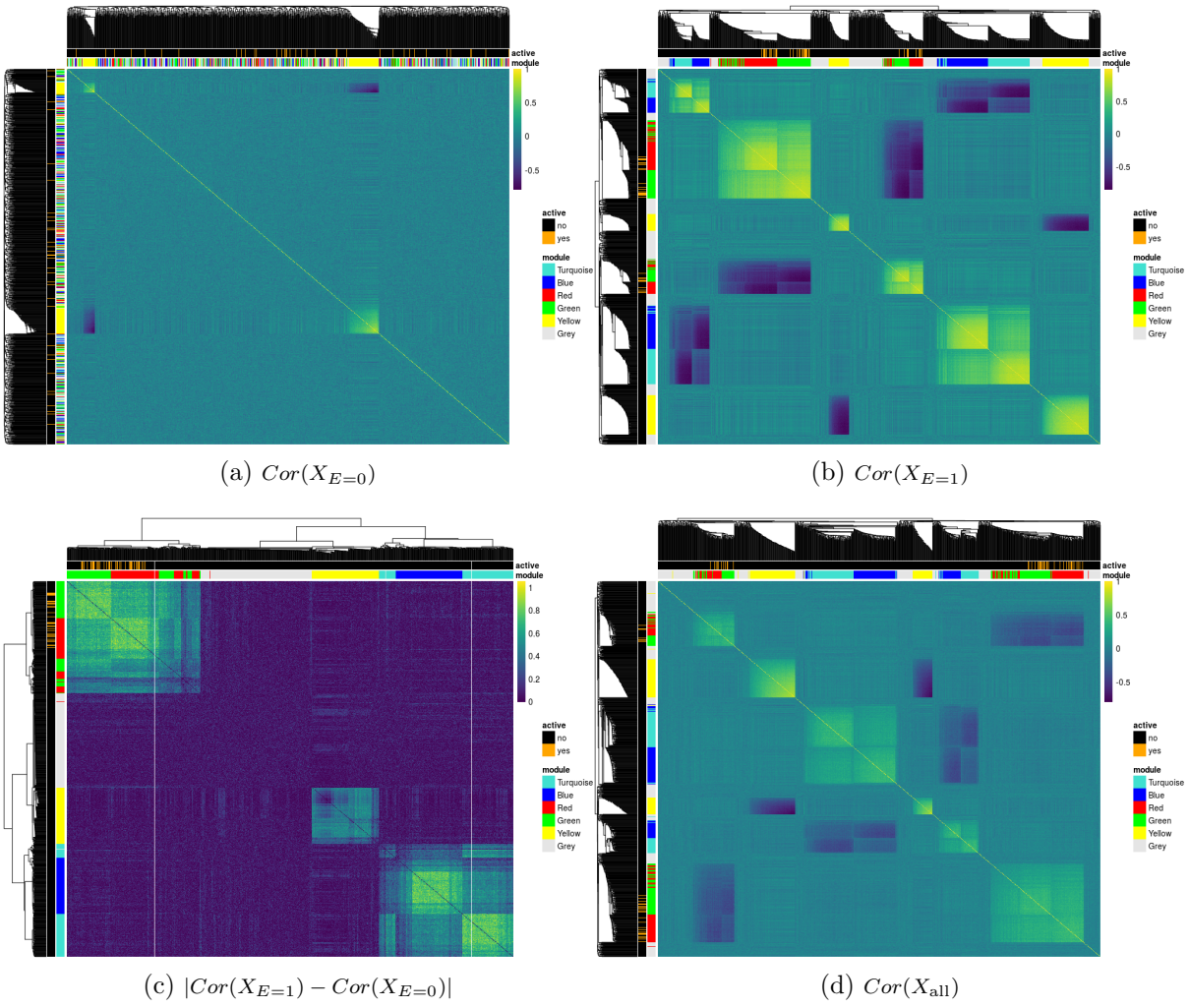
(b) $TOM(X_{E=1})$

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

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

Figure B.1: 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. The *module* annotation represents the true cluster membership for each predictor, and the *active* annotation represents the truly associated predictors with the response.

B.2 Pearson Correlation Matrix



(a) $Cor(X_{E=0})$ (b) $Cor(X_{E=1})$

(c) $|Cor(X_{E=1}) - Cor(X_{E=0})|$ (d) $Cor(X_{\text{all}})$

Figure B.2: Pearson correlation matrices 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 correlation matrix for a, b, and d and the euclidean distance for c. The *module* annotation represents the true cluster membership for each predictor, and the *active* annotation represents the truly associated predictors with the response.

C Analysis of Clusters

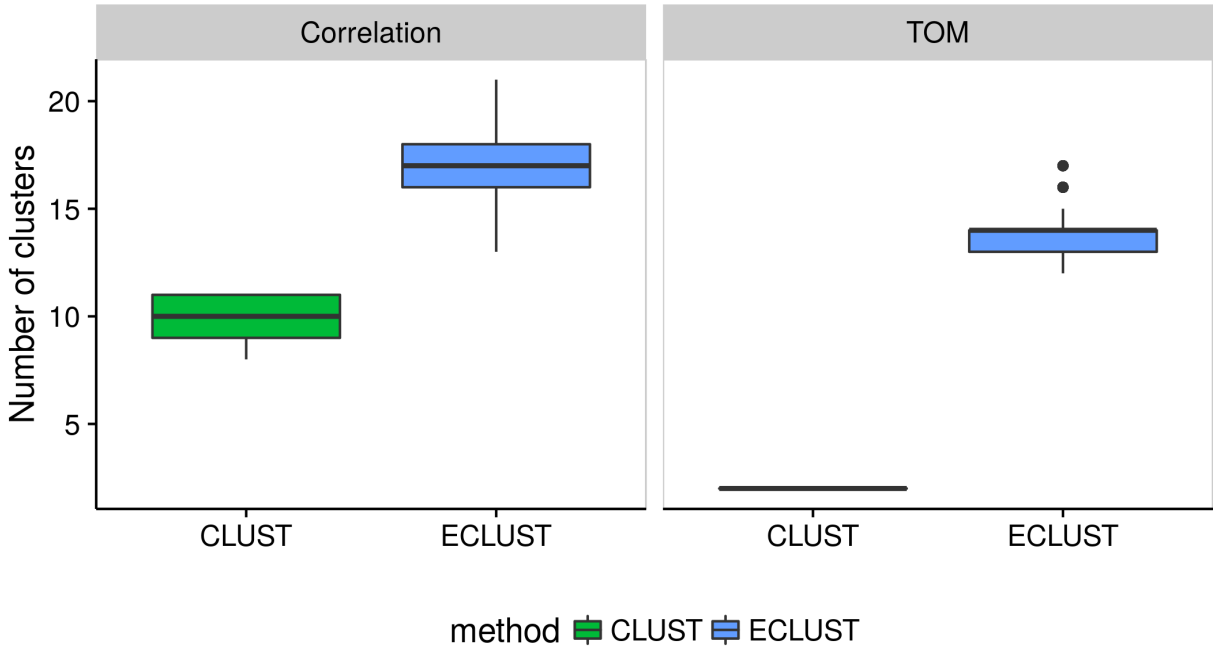


Figure C.1: Number of estimated clusters from applying the `dynamicTreeCut` algorithm to hierarchical clustering of the dissimilarity matrix with average linkage. Left panel: CLUST uses $1 - Cor(X_{all})$ and ECLUST uses the euclidean distance of $Cor(X_{\text{diff}})$ as measures of dissimilarity. Right panel: CLUST uses $1 - TOM(X_{all})$ and ECLUST uses the euclidean distance of $TOM(X_{\text{diff}})$ as measures of dissimilarity. Empirical distributions based on 200 simulation runs.

D Simulation Results Using TOM as a Measure of Similarity

D.1 Simulation 1

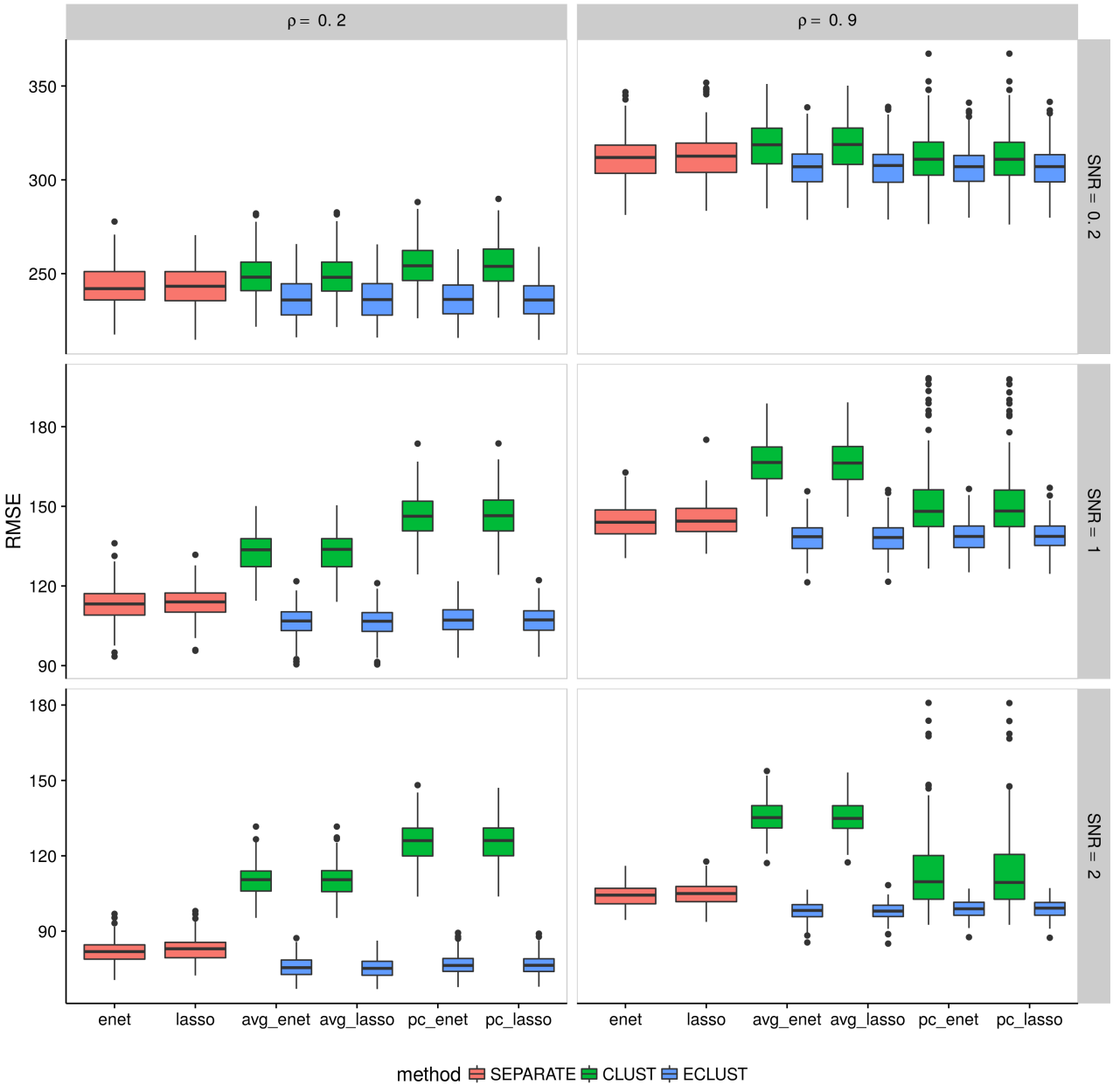


Figure D.1: Simulation 1 – Root mean squared error on an independent test set using the TOM as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

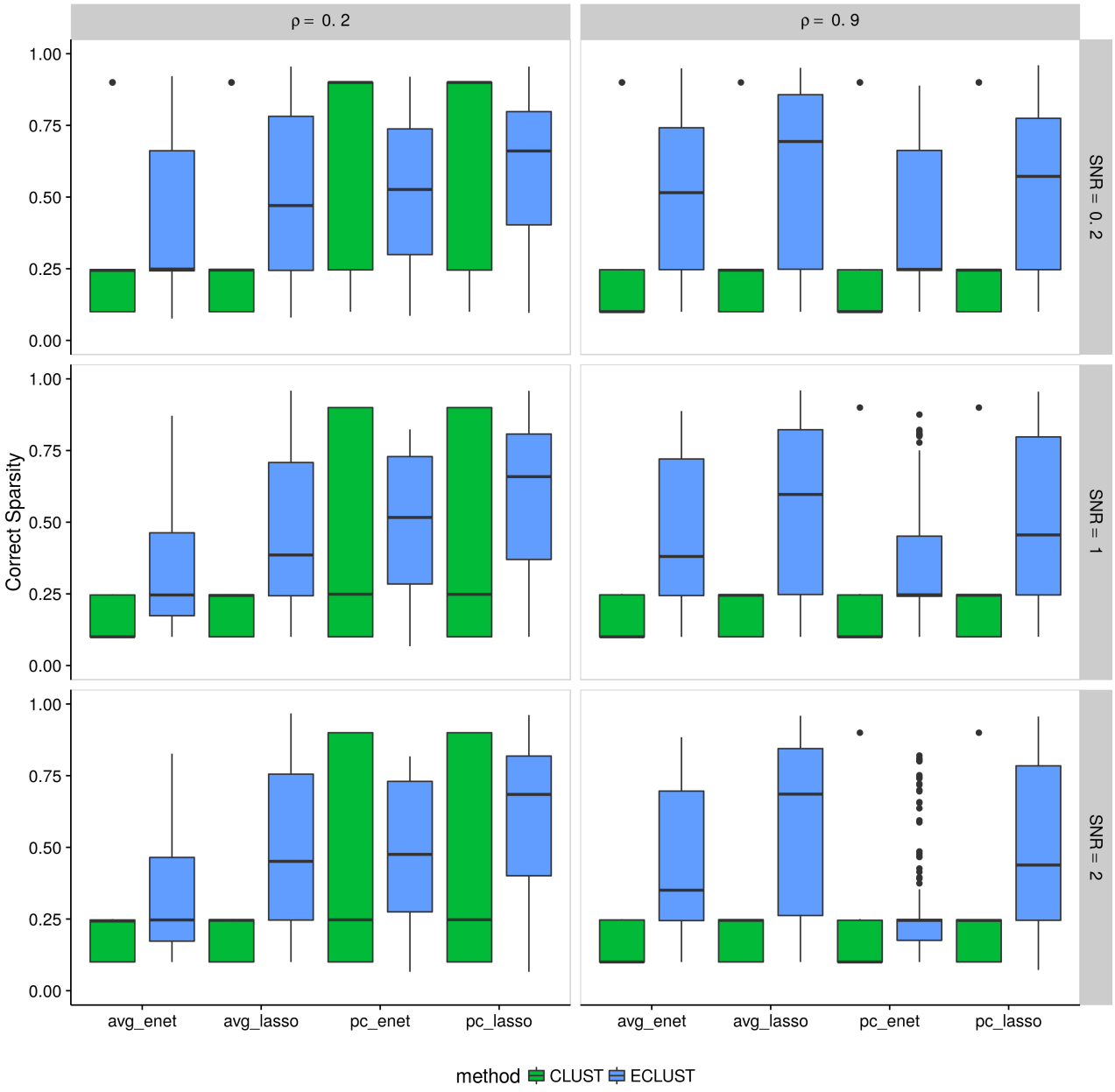


Figure D.2: Simulation 1 – Correct Sparsity based on the training set using the TOM as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

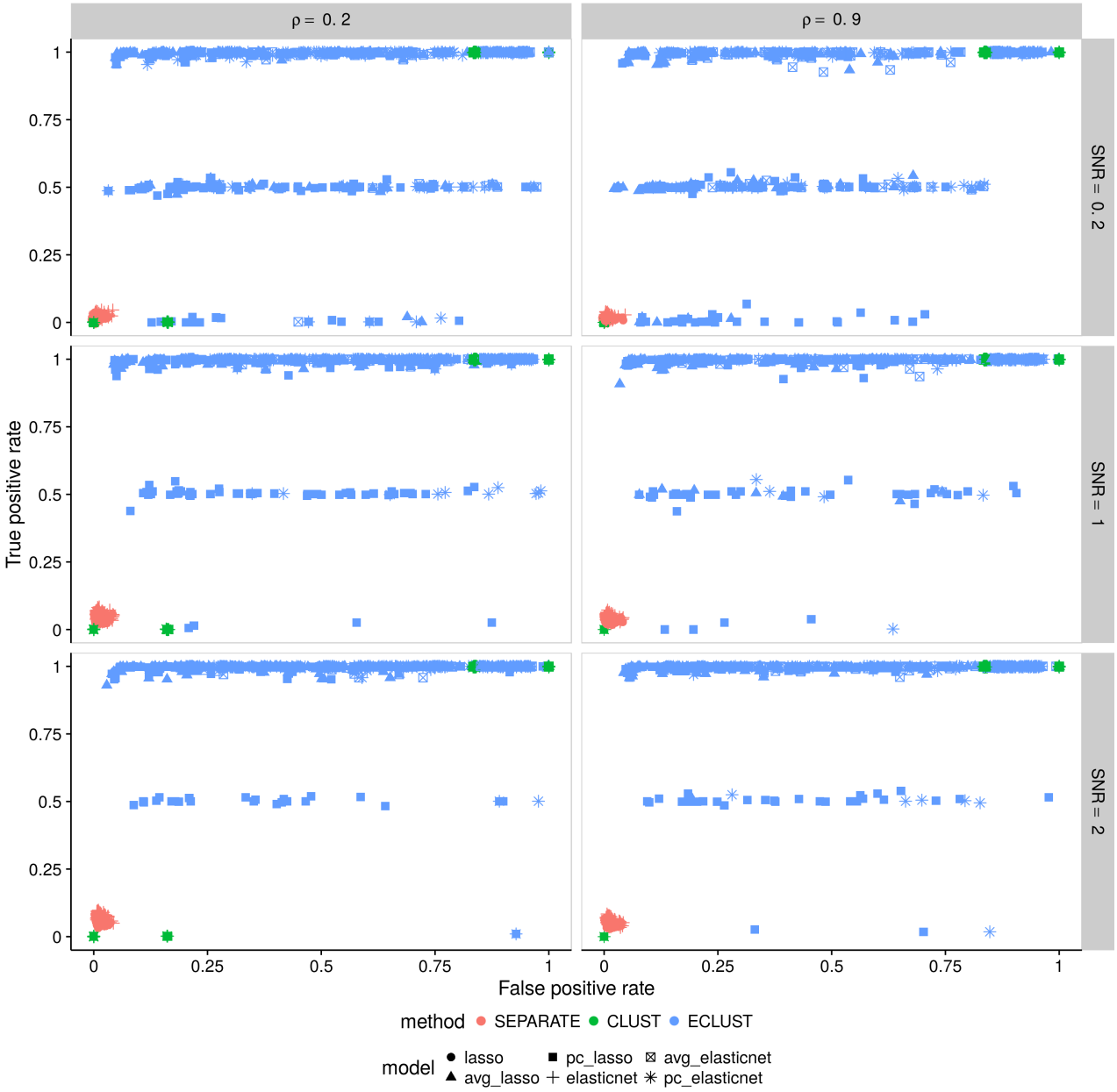


Figure D.3: Simulation 1 – True positive rate vs. false positive rate based on the training set using the TOM as a measure of similarity. Each point represents 1 simulation run (there are a total of 200 simulation runs). Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

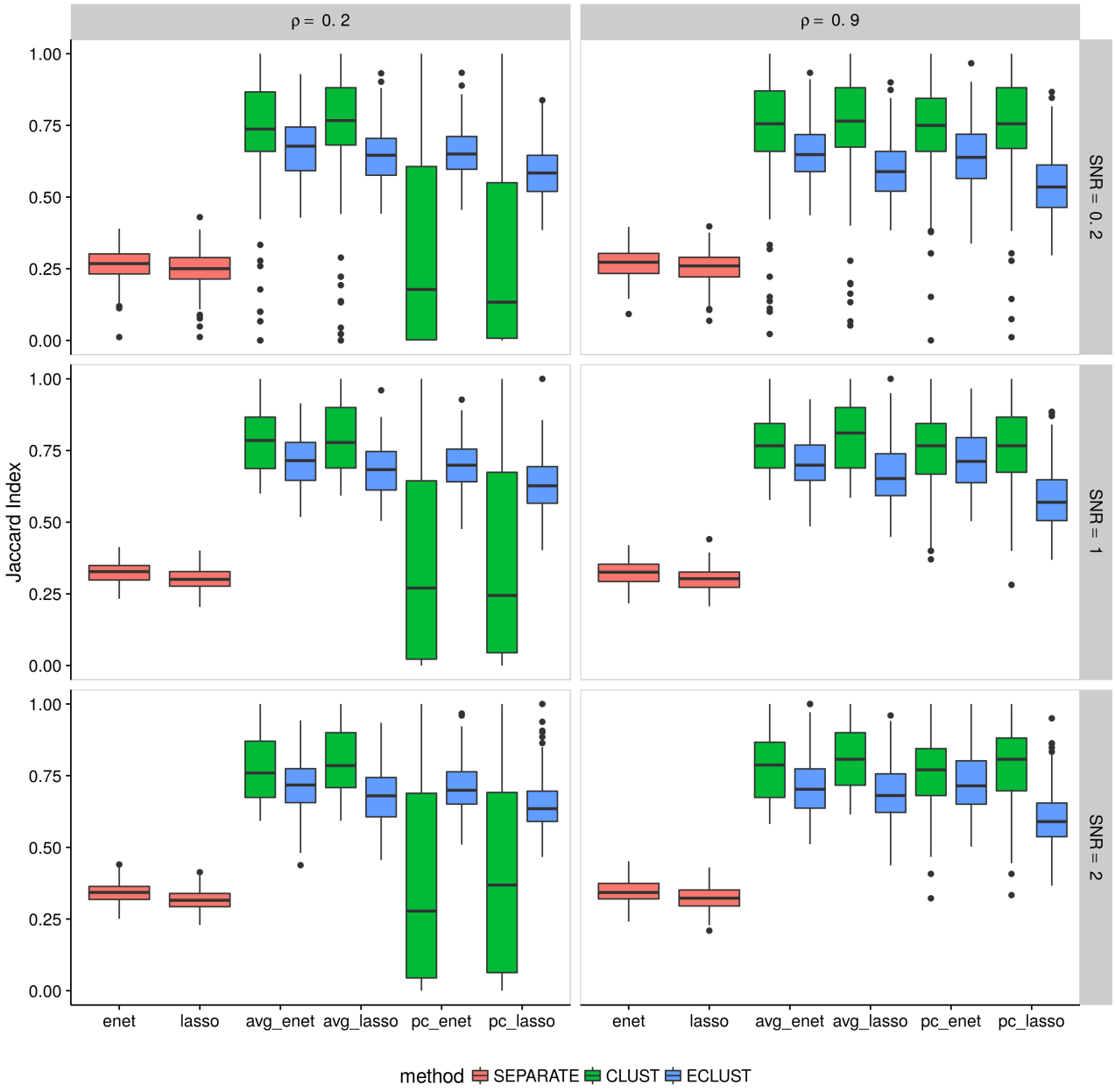


Figure D.4: Simulation 1 – Average Jaccard Index from 10 CV folds of the training set using the TOM as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of selected predictors. We then calculate the Jaccard Index between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

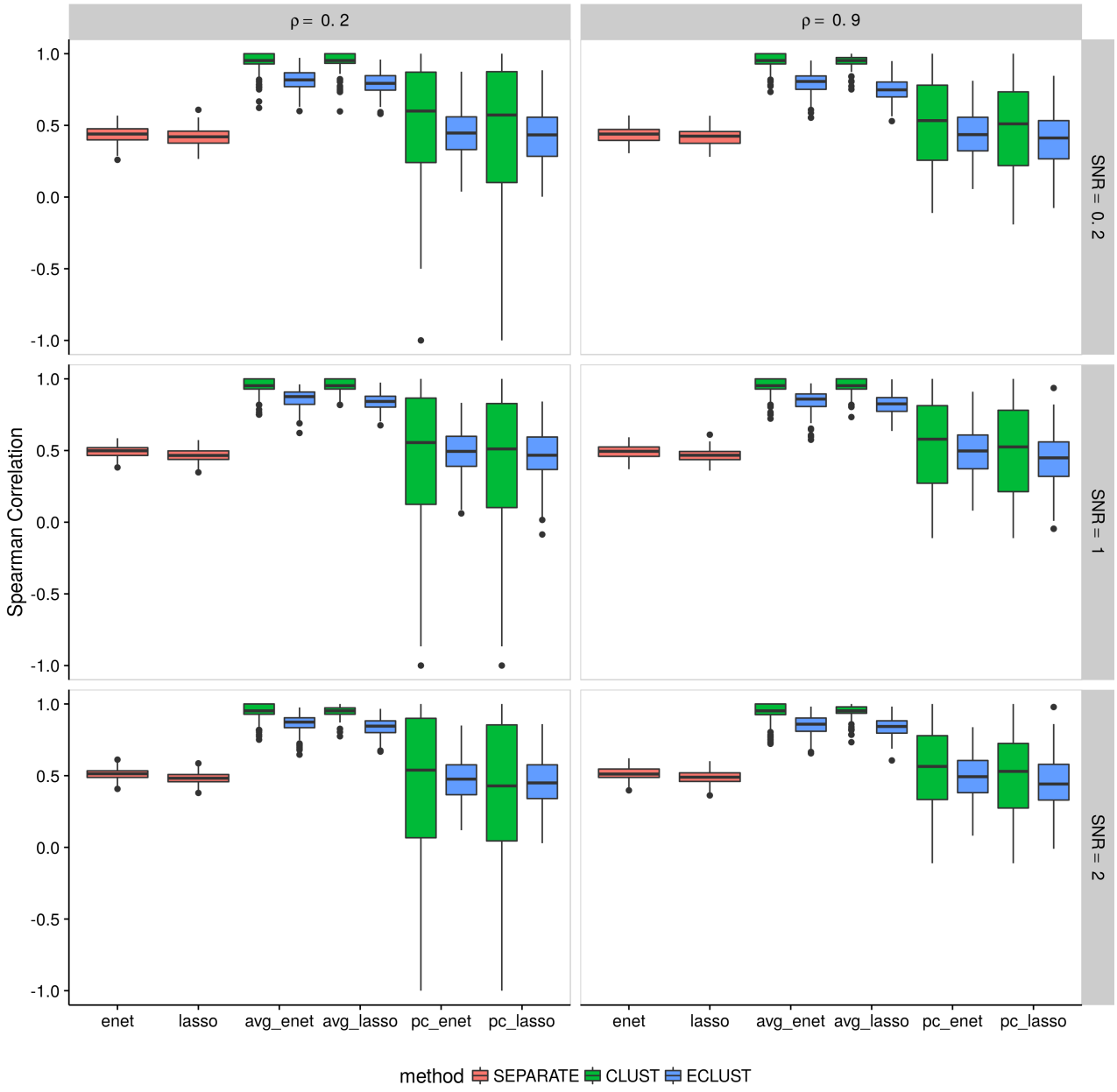


Figure D.5: Simulation 1 – Average Spearman correlation from 10 CV folds of the training set using the TOM as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated regression coefficients. We then calculate the Spearman correlation between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

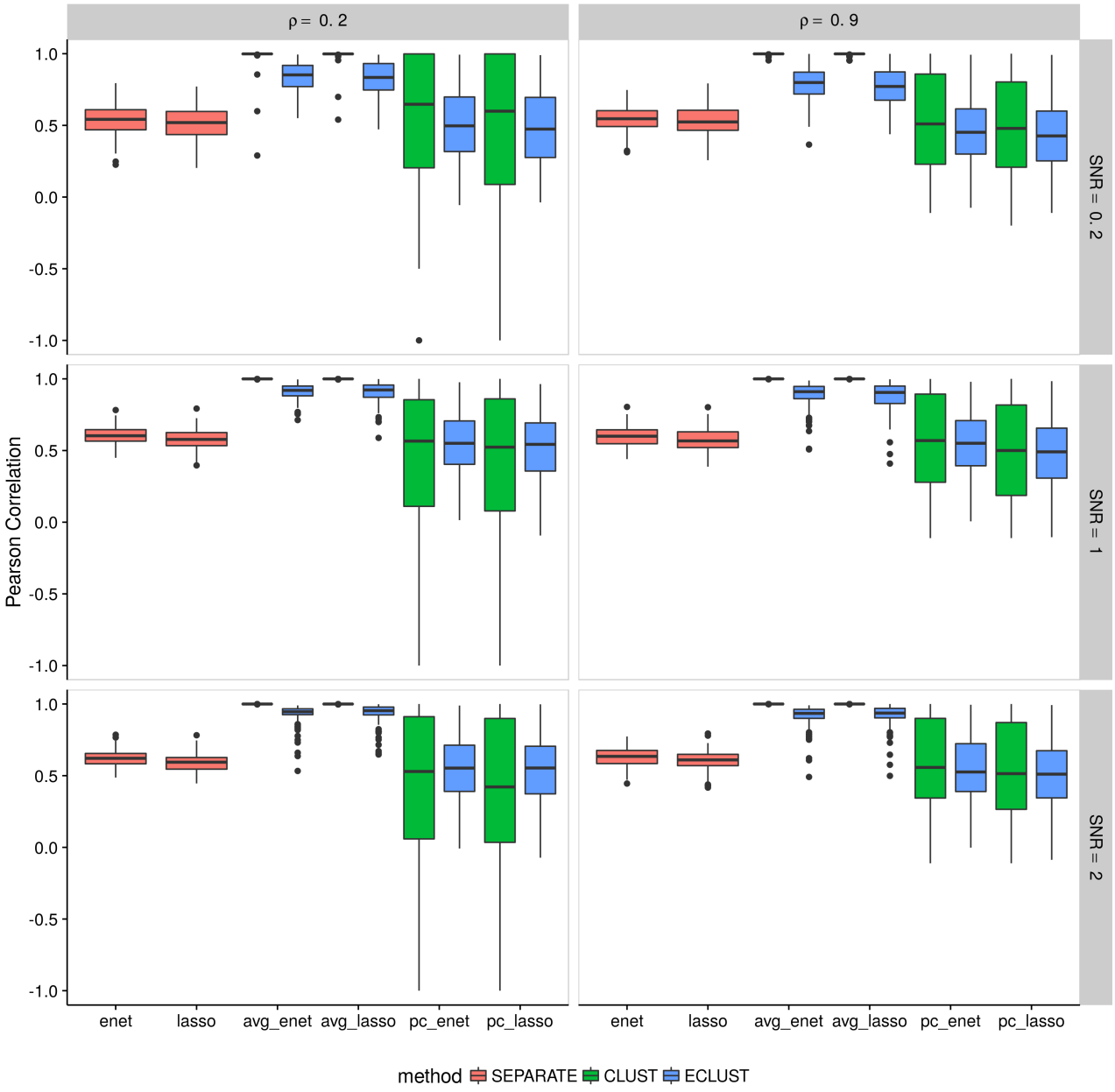


Figure D.6: Simulation 1 – Average Pearson correlation from 10 CV folds of the training set using the TOM as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated regression coefficients. We then calculate the Pearson correlation between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

D.2 Simulation 2

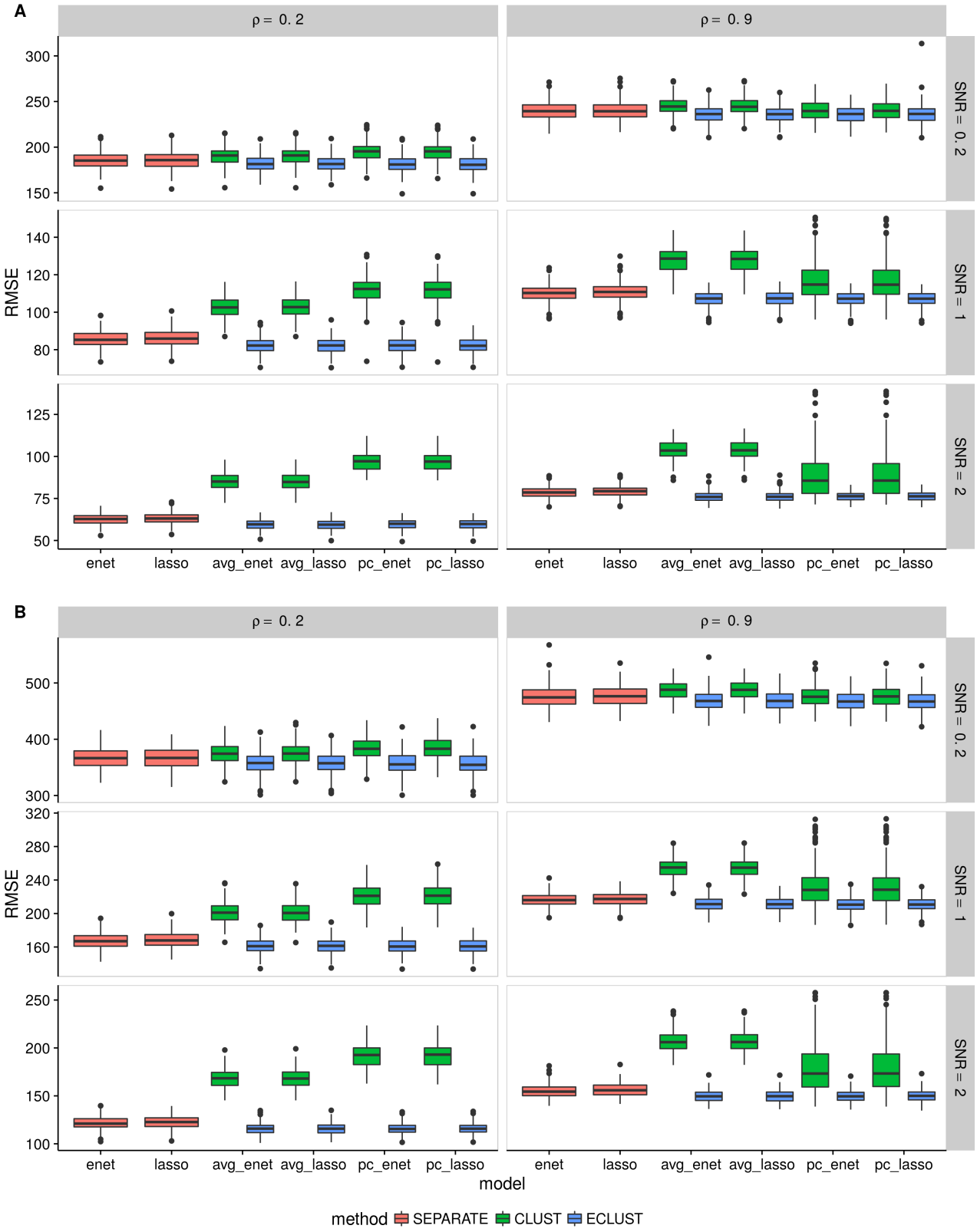


Figure D.7: Simulation 2 – Root mean squared error on an independent test set using the TOM as a measure of similarity from 200 simulation runs. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

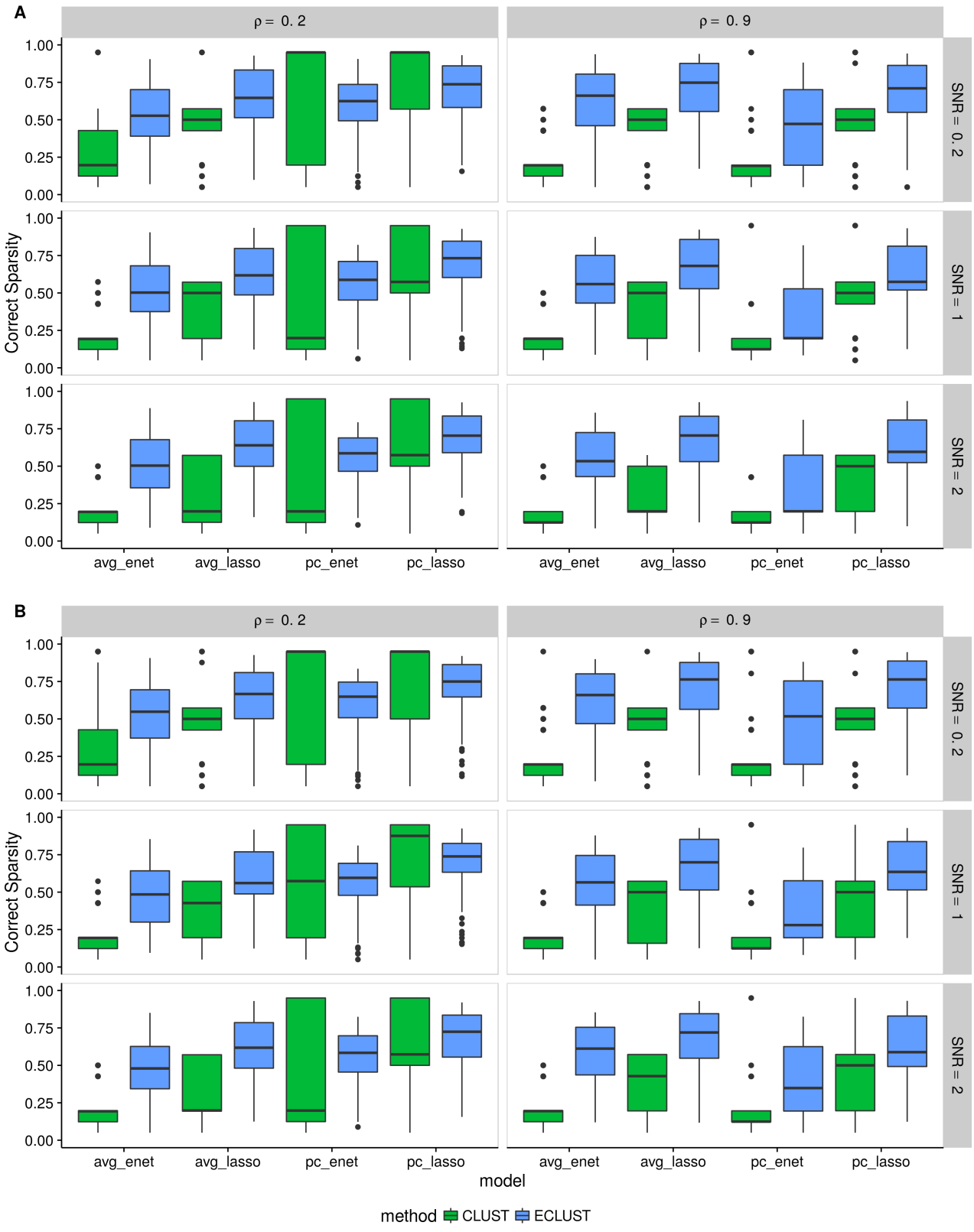


Figure D.8: Simulation 2 – Correct Sparsity based on the training set using the TOM as a measure of similarity from 200 simulation runs. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

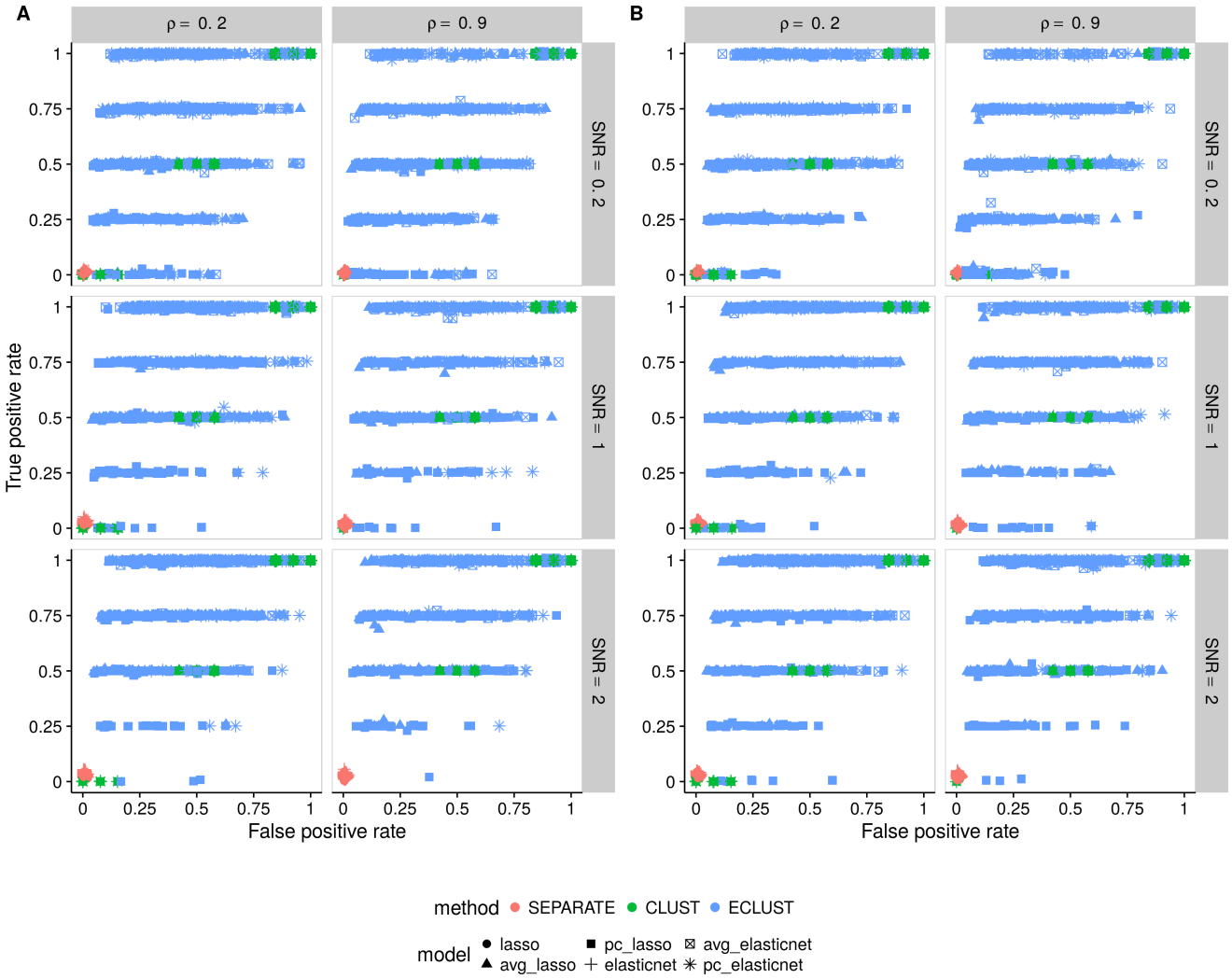


Figure D.9: Simulation 2 – True positive rate vs. false positive rate based on the training set using the TOM as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Each point represents 1 simulation run (there are a total of 200 simulation runs). Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

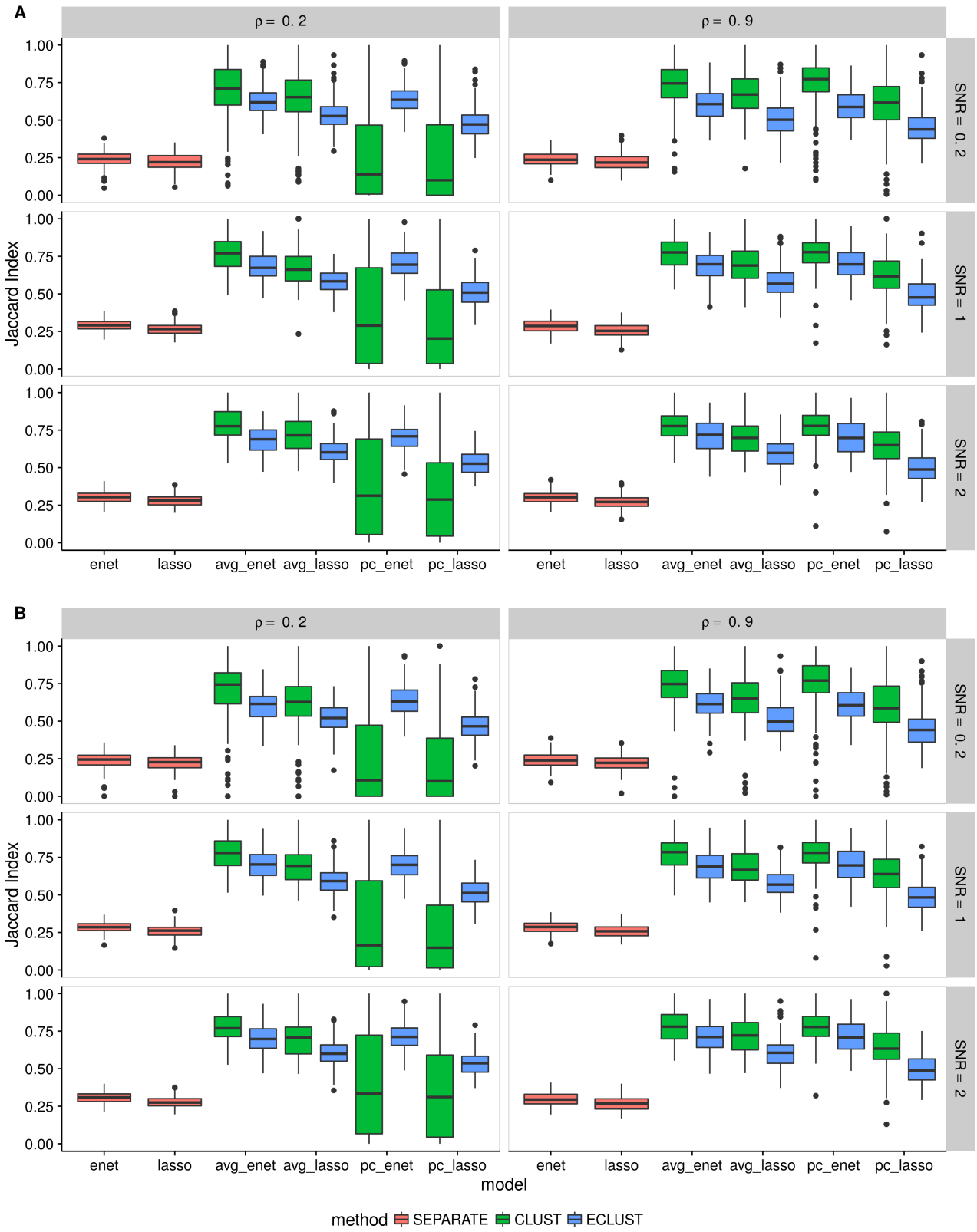


Figure D.10: Simulation 2 – Average Jaccard Index from 10 CV folds of the training set using the TOM as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. We fit the model to each of the 10 CV folds resulting in 10 sets of selected predictors. We then calculate the Jaccard Index between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

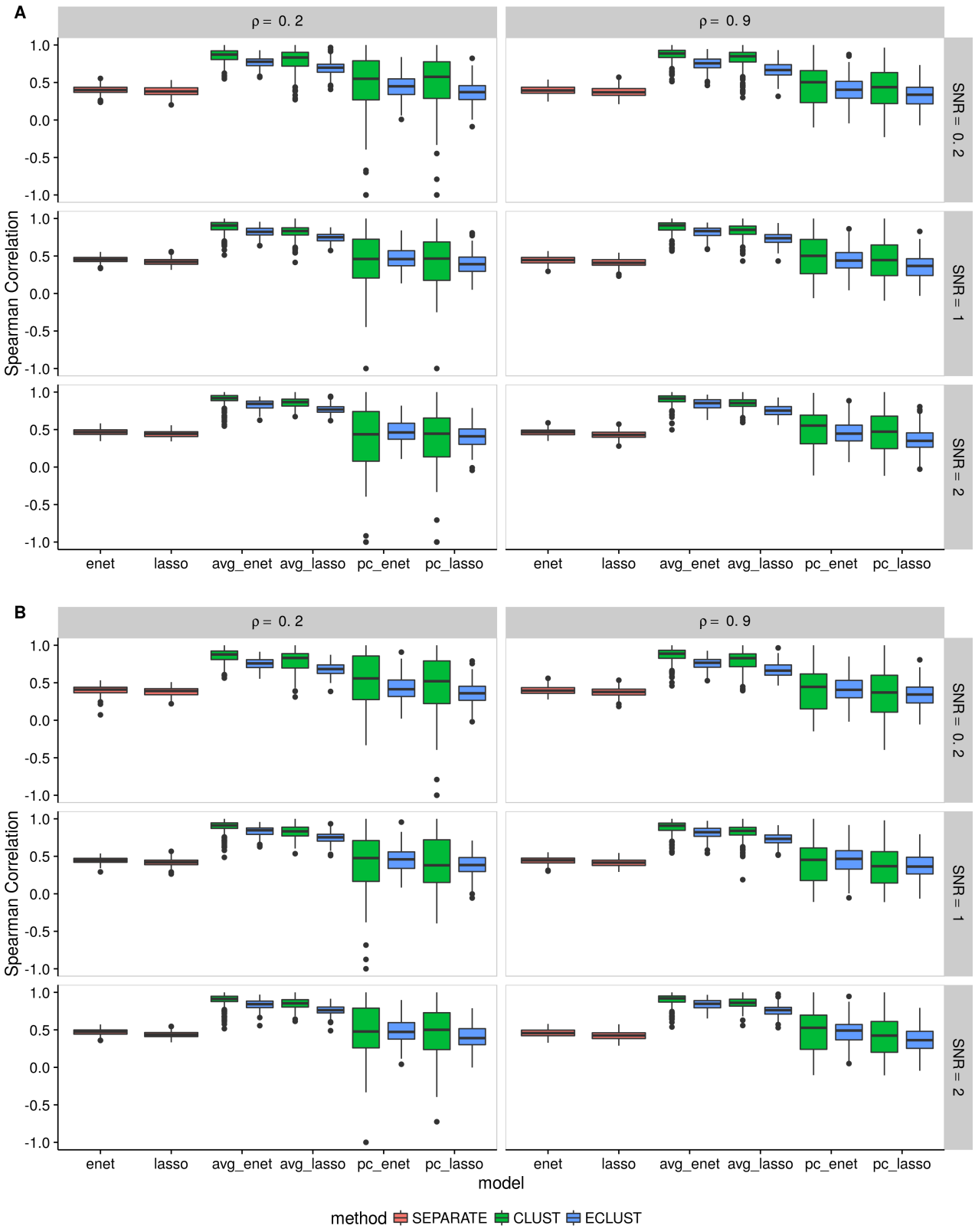


Figure D.11: Simulation 2 – Average Spearman correlation from 10 CV folds of the training set using the TOM as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated regression coefficients. We then calculate the Spearman correlation between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

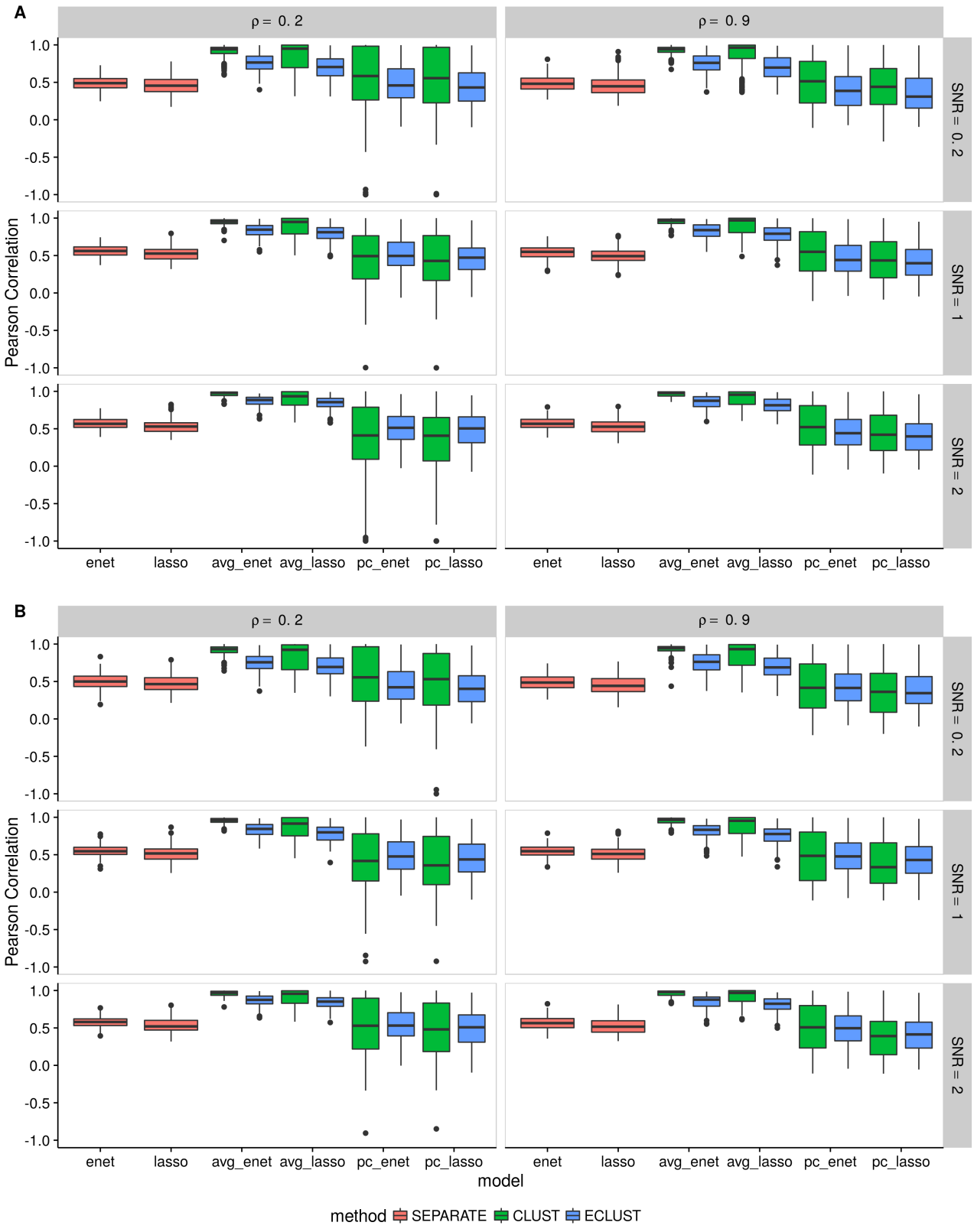


Figure D.12: Simulation 3 – Average Pearson correlation from 10 CV folds of the training set using the TOM as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated regression coefficients. We then calculate the Pearson correlation between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

D.3 Simulation 3

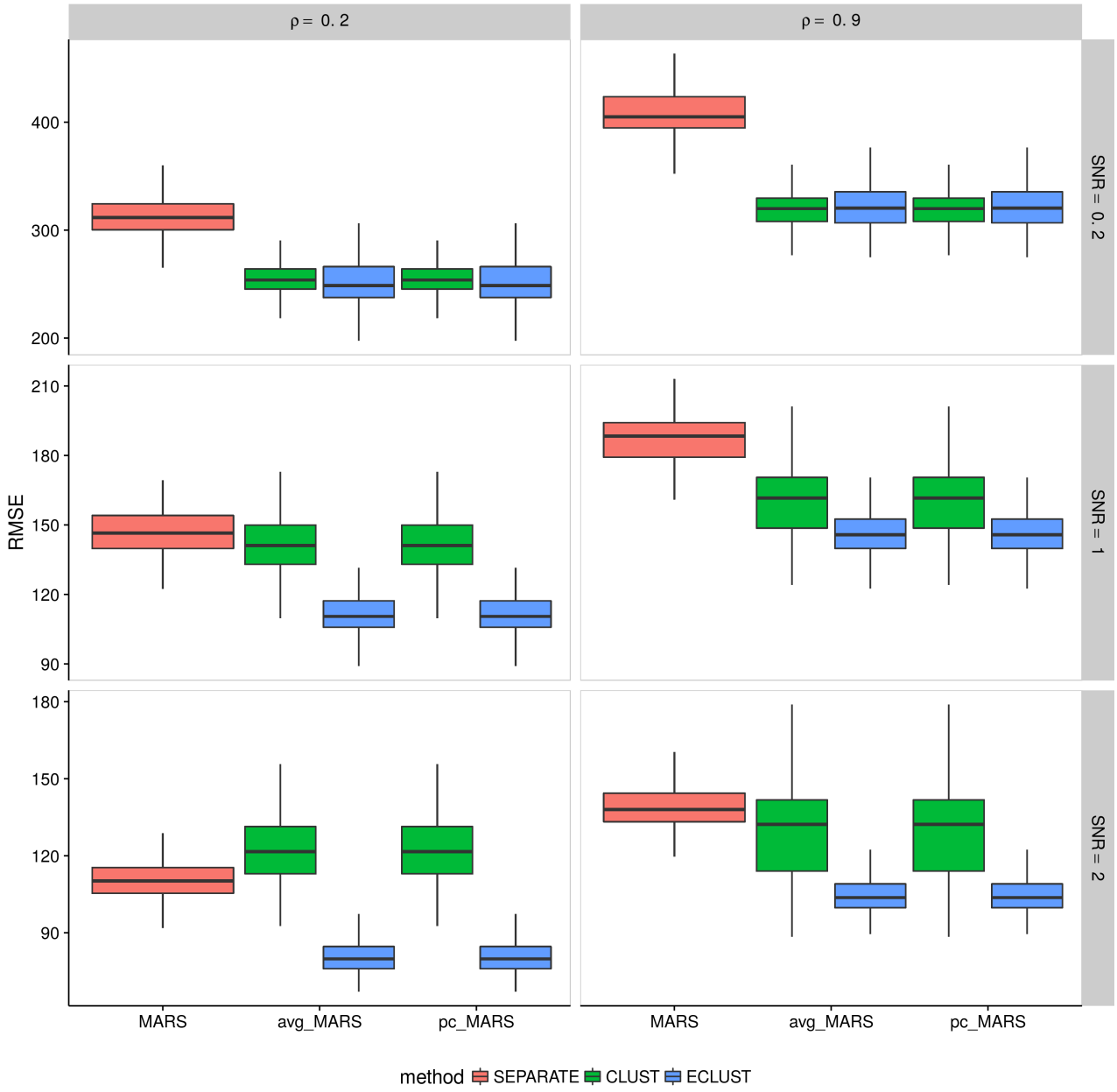


Figure D.13: Simulation 3 – Root mean squared error on an independent test set using the TOM as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

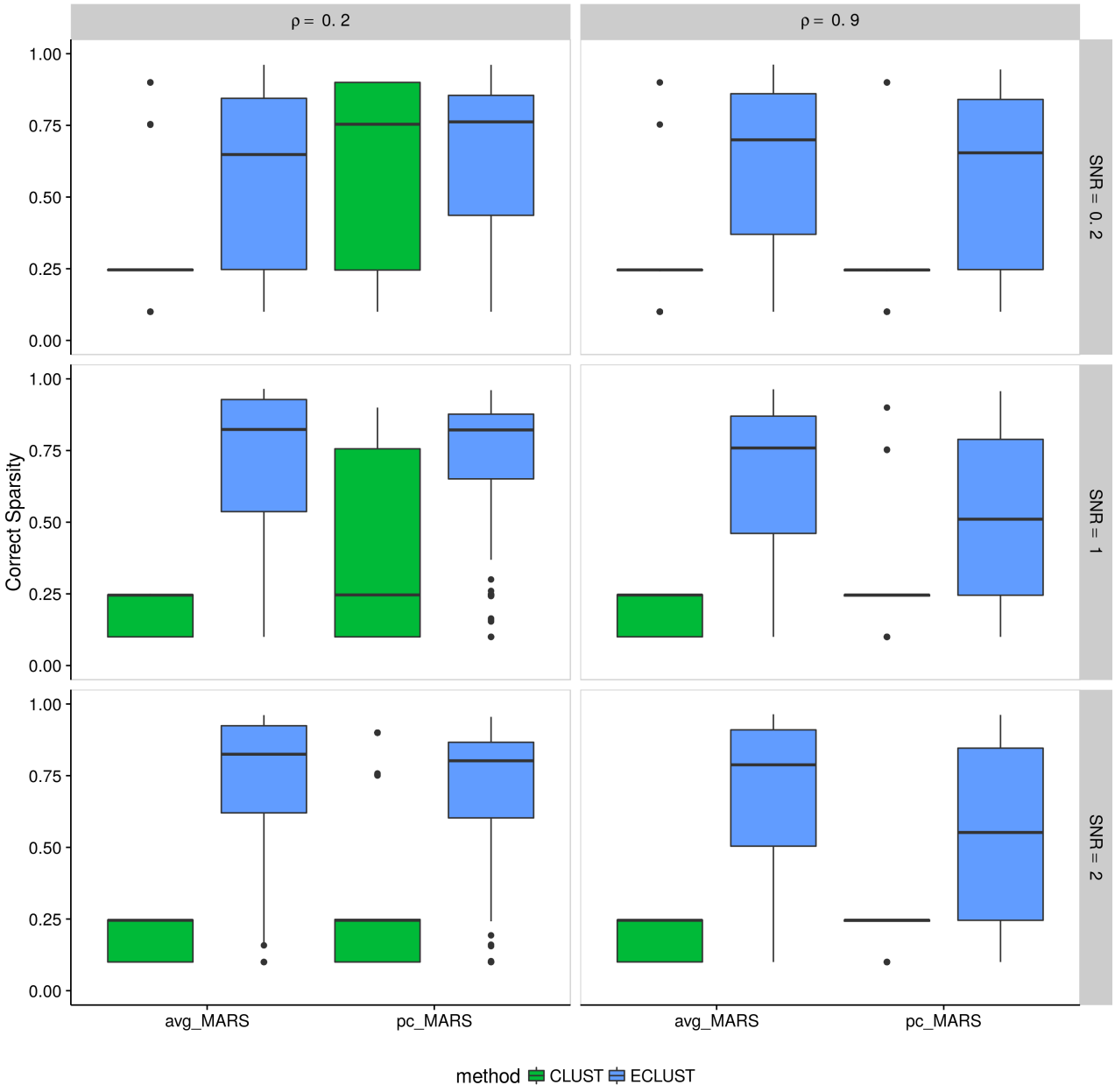


Figure D.14: Simulation 3 – Correct Sparsity based on the training set using the TOM as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

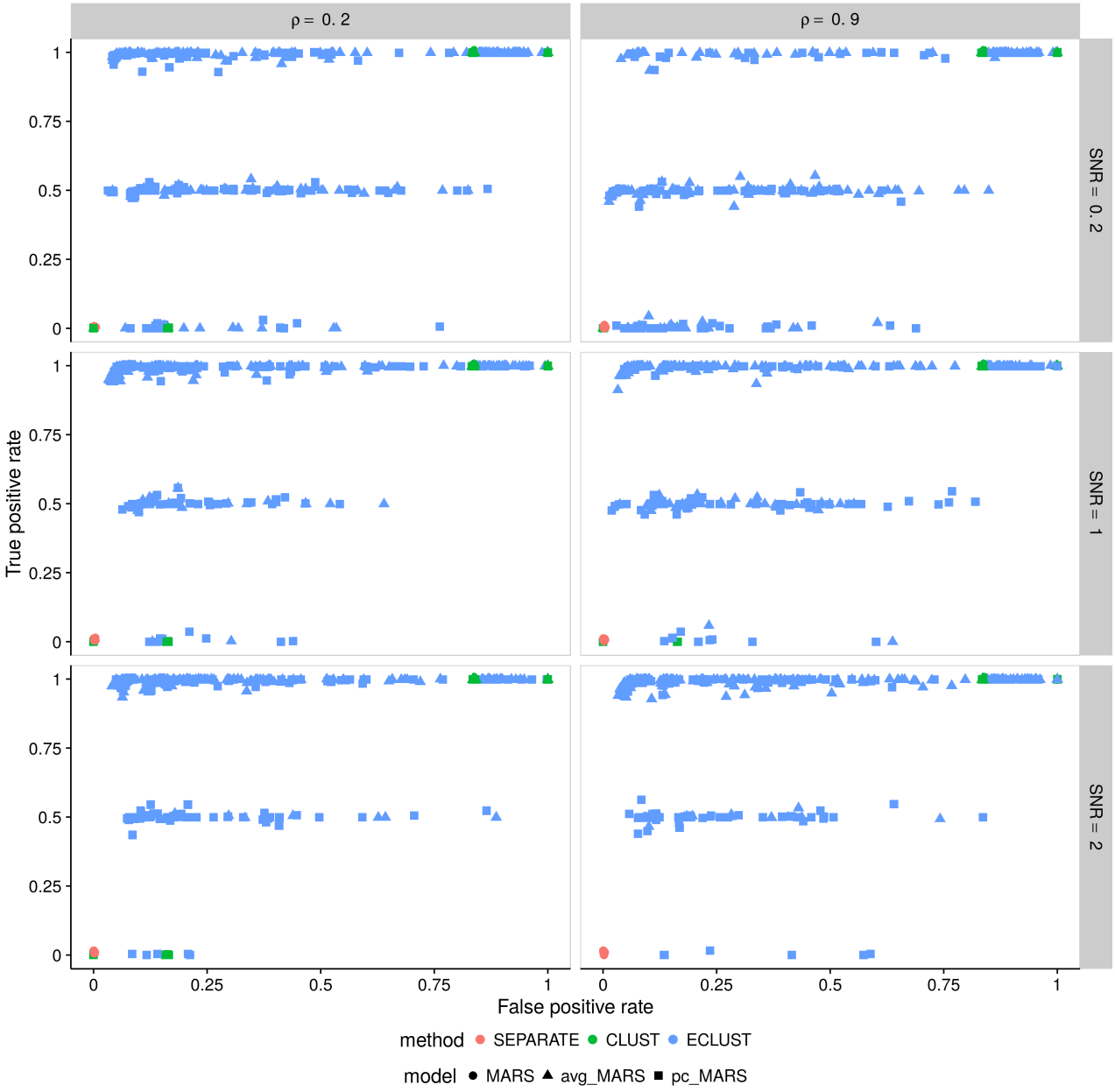


Figure D.15: Simulation 3 – True positive rate vs. false positive rate based on the training set using the TOM as a measure of similarity. Each point represents 1 simulation run (there are a total of 200 simulation runs). Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

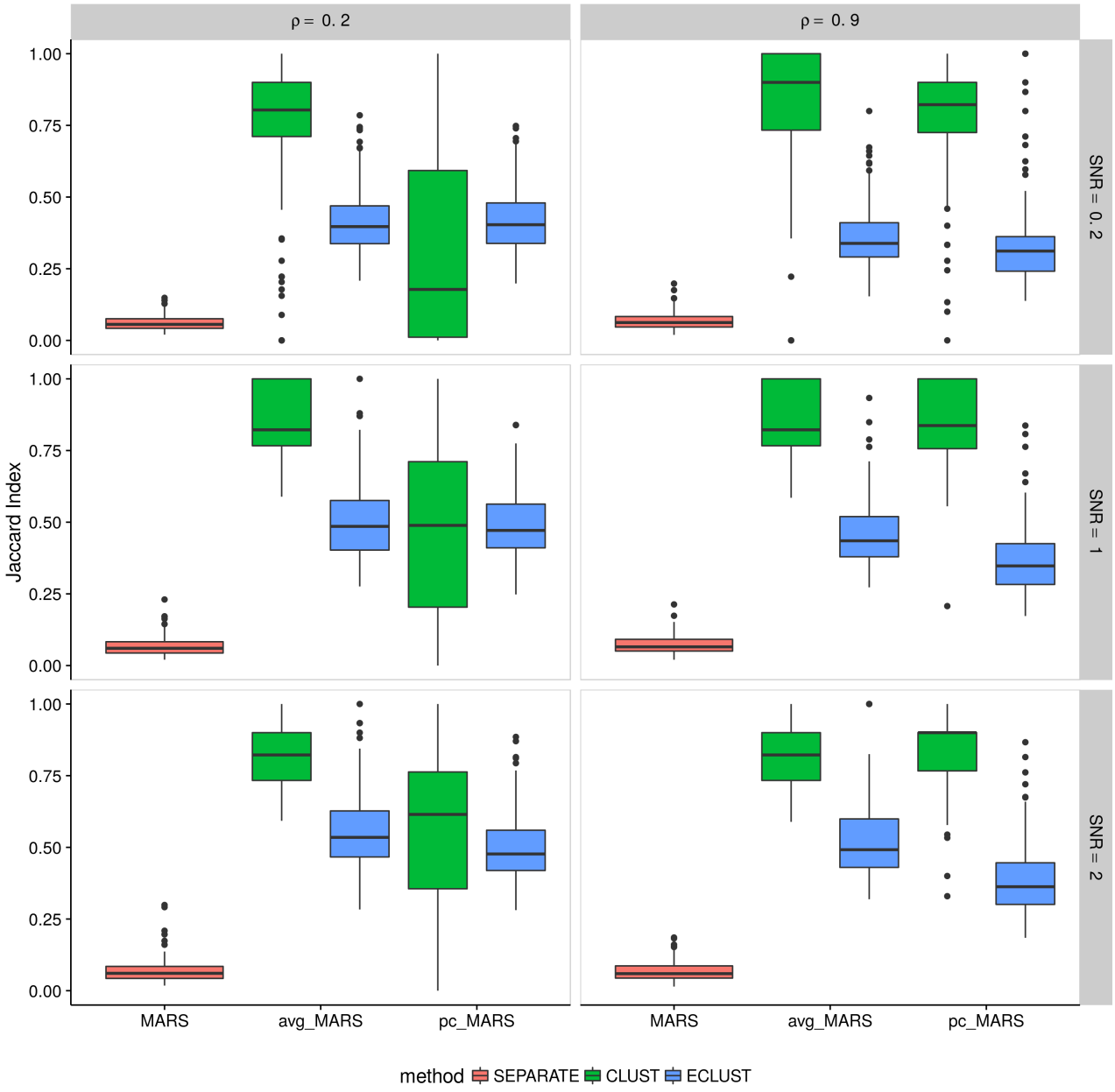


Figure D.16: Simulation 3 – Average Jaccard Index from 10 CV folds of the training set using the TOM as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of selected predictors. We then calculate the Jaccard Index between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

E Simulation Results Using Pearson Correlations as a Measure of Similarity

E.1 Simulation 1

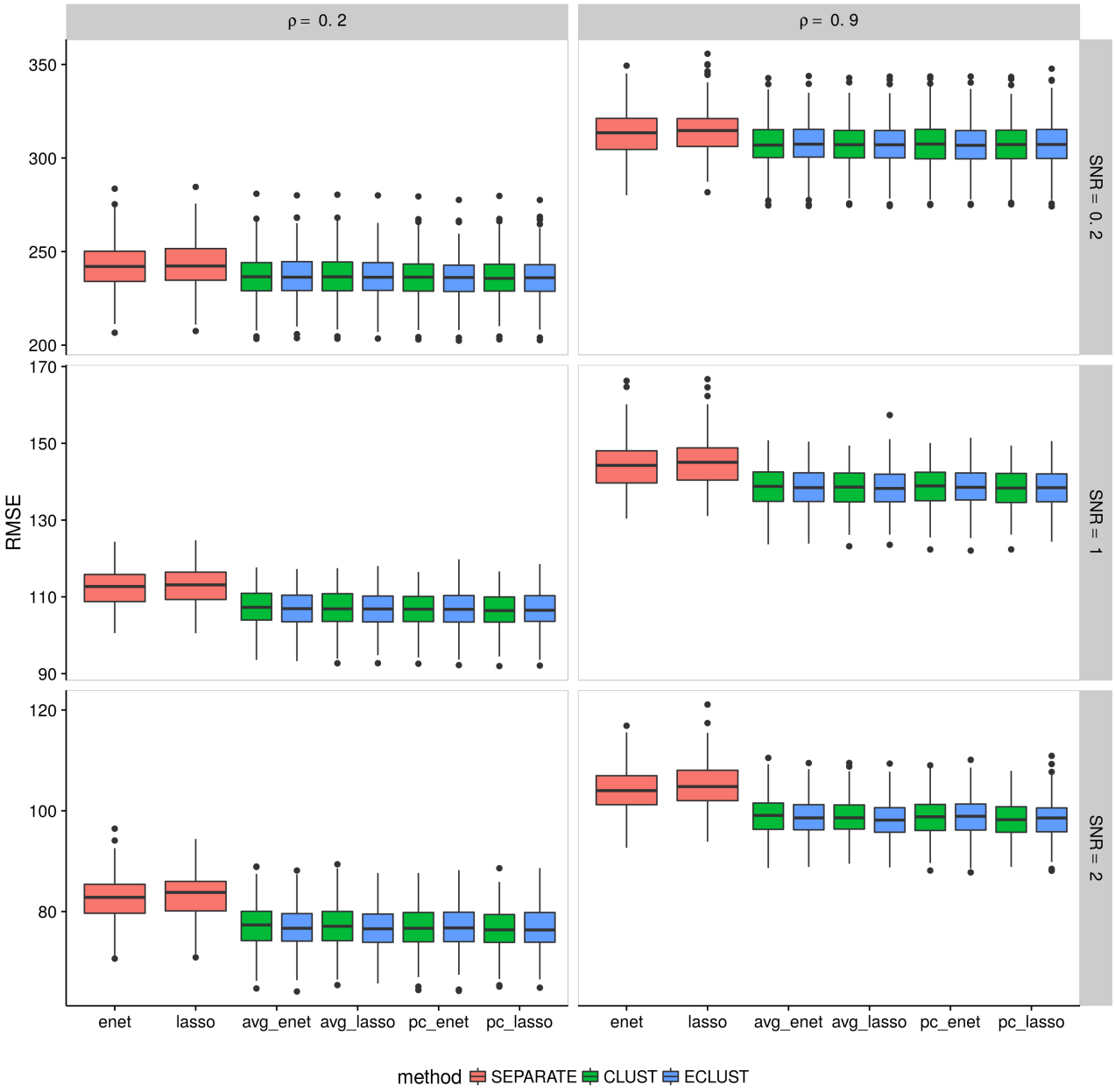


Figure E.1: Simulation 1 – Root mean squared error on an independent test set using the Correlation as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

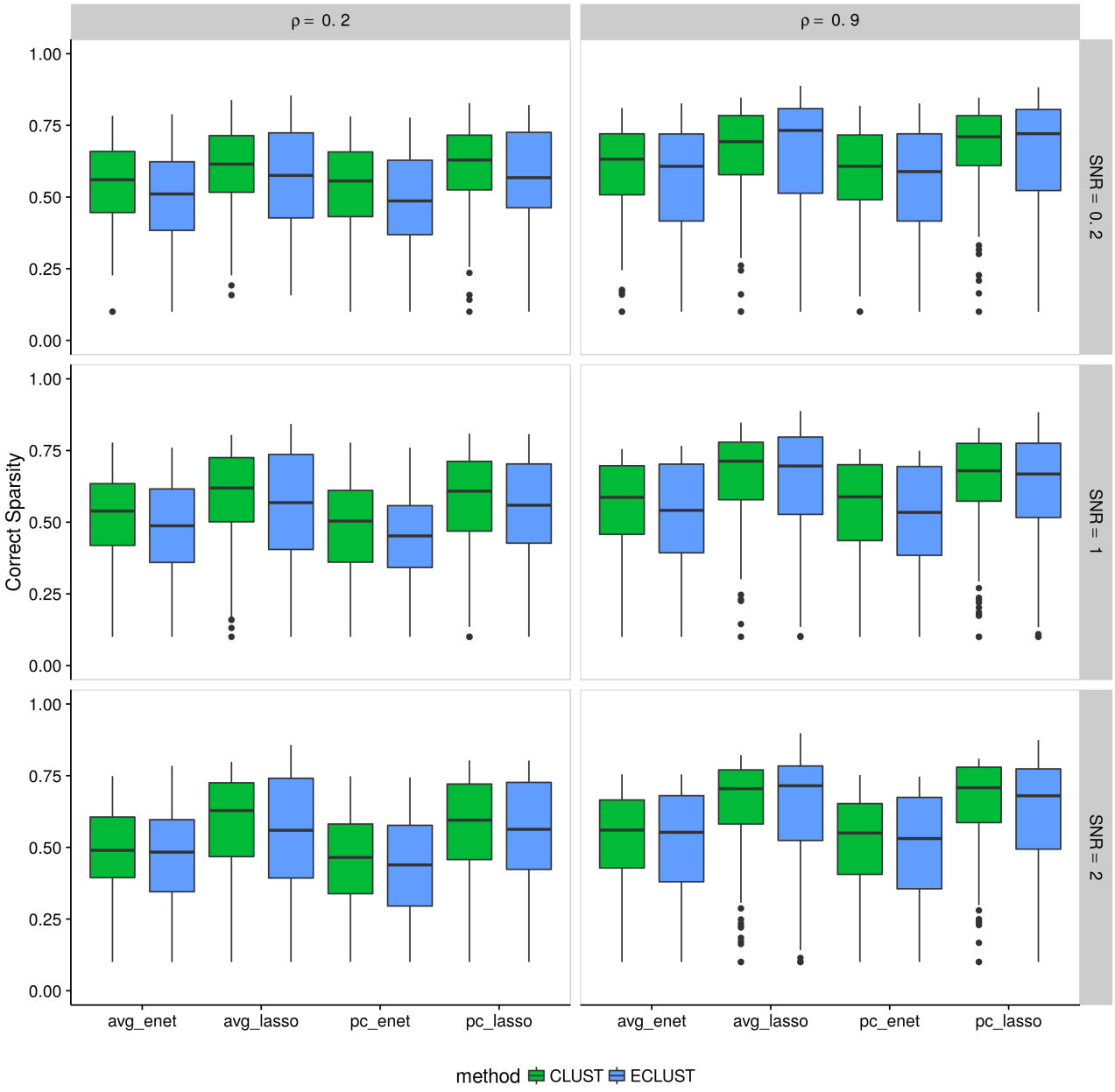


Figure E.2: Simulation 1 – Correct Sparsity based on the training set using the Pearson correlation as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

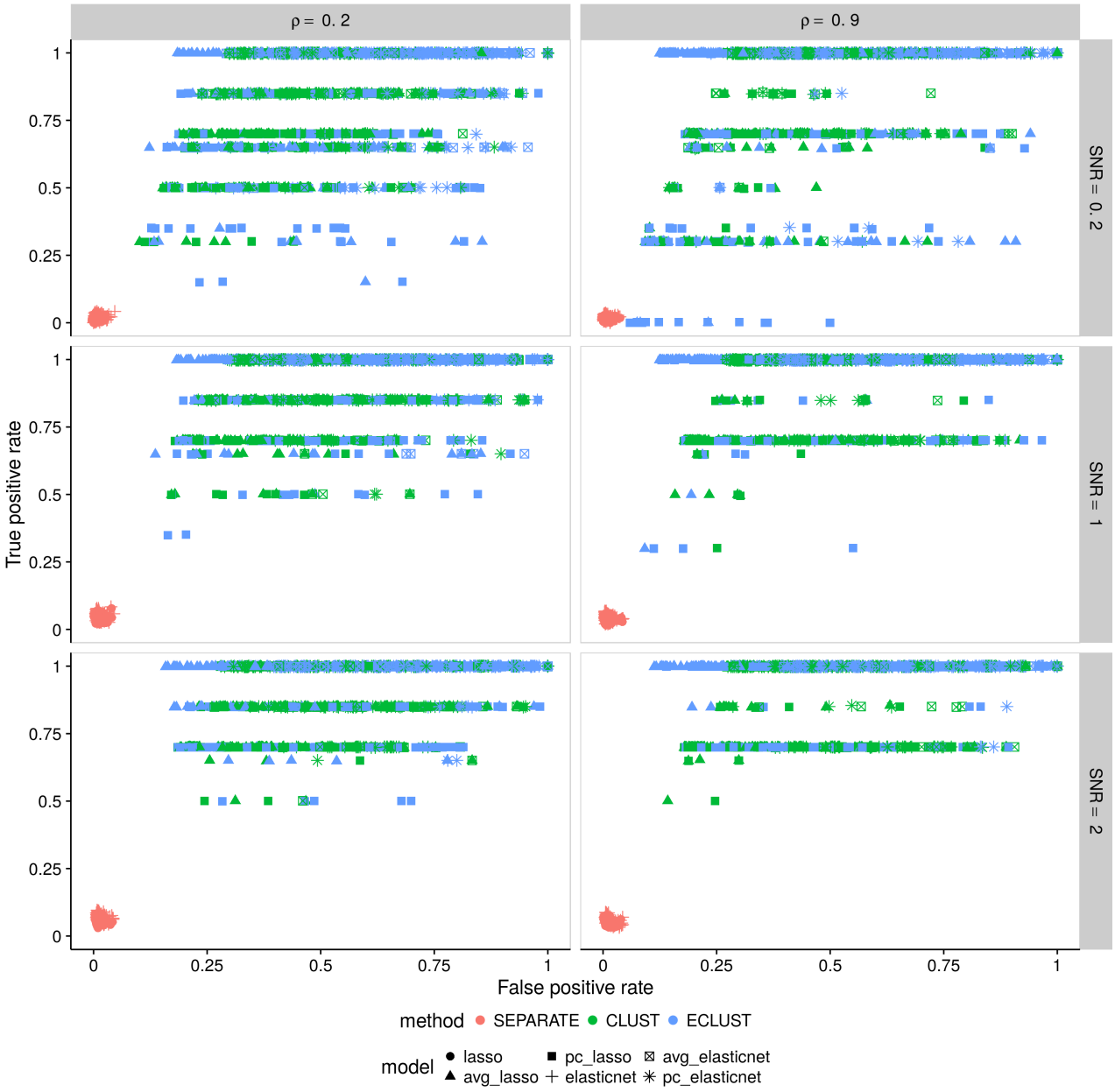


Figure E.3: Simulation 1 – True positive rate vs. false positive rate based on the training set using the Pearson correlation as a measure of similarity. Each point represents 1 simulation run (there are a total of 200 simulation runs). Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

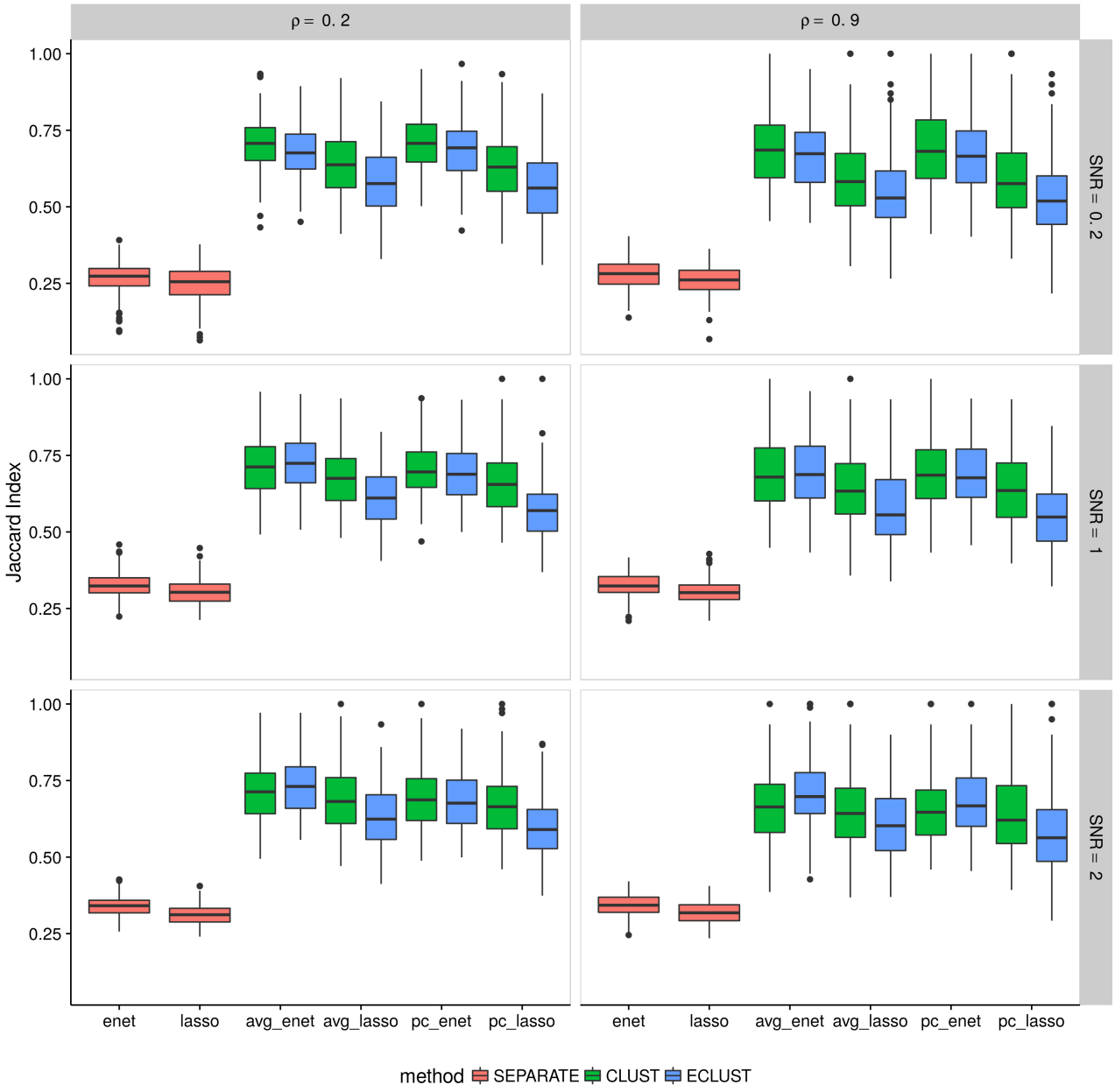


Figure E.4: Simulation 1 – Average Jaccard Index from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of selected predictors. We then calculate the Jaccard Index between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

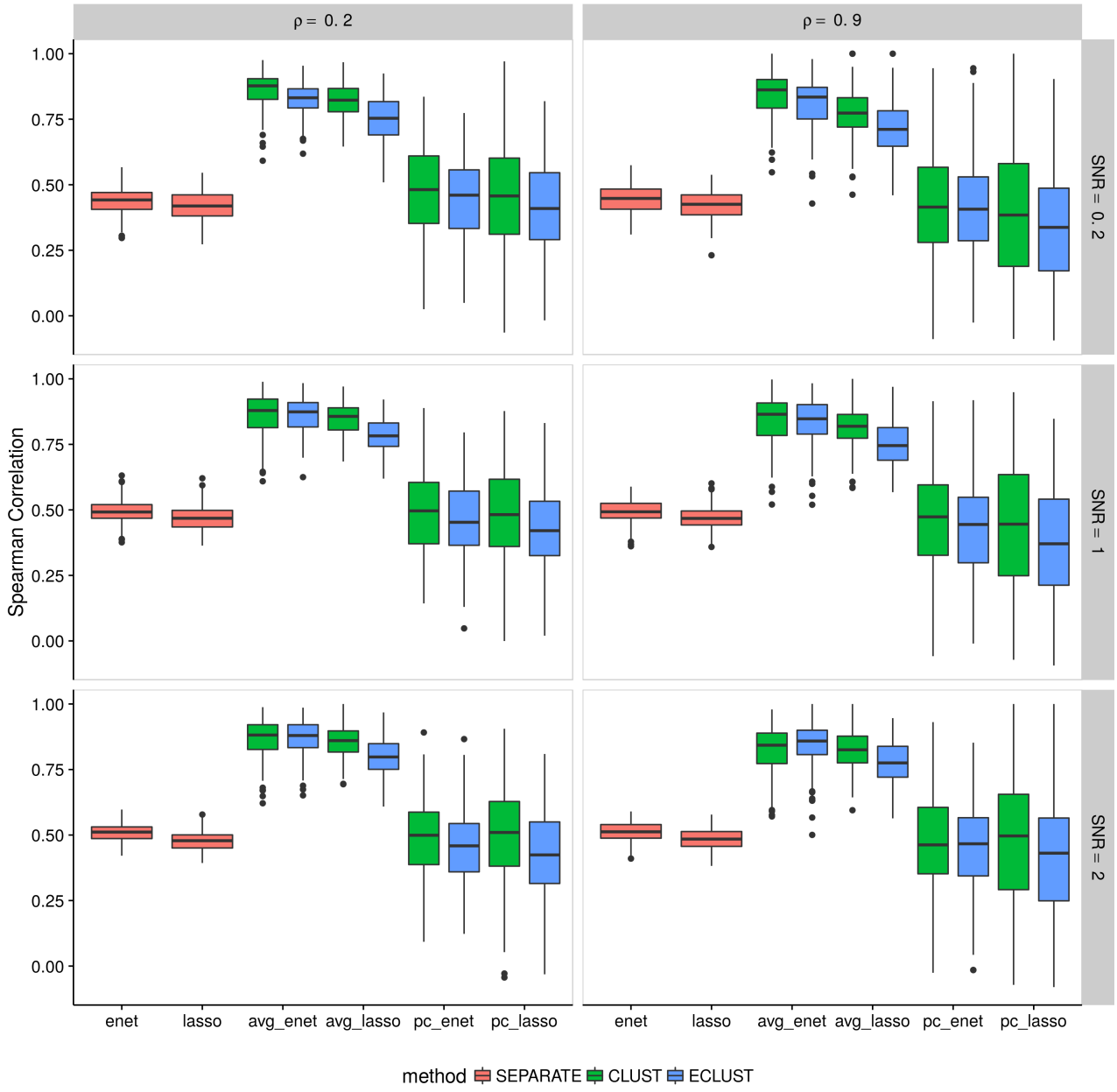


Figure E.5: Simulation 1 – Average Spearman correlation from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated regression coefficients. We then calculate the Spearman correlation between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

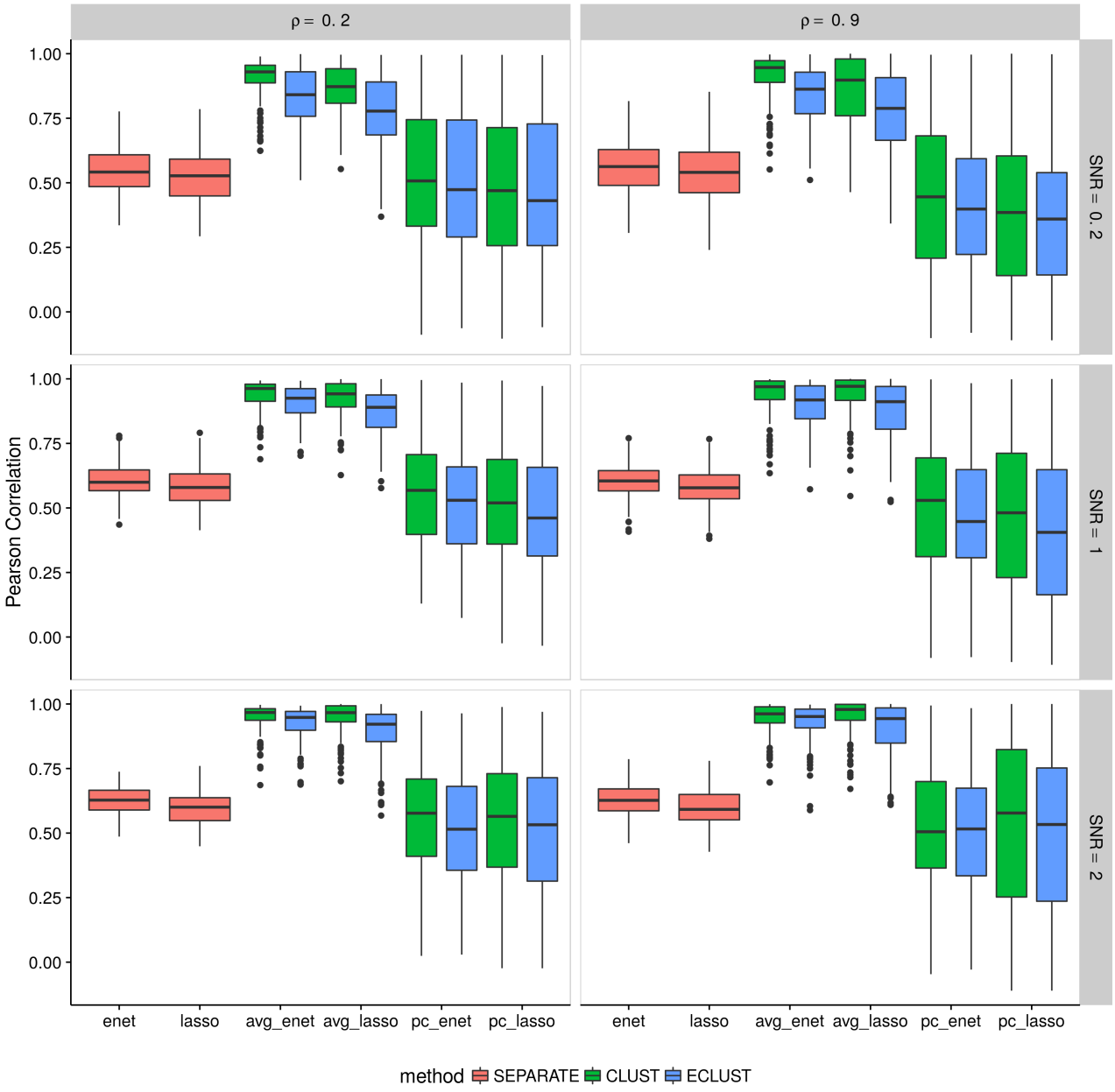


Figure E.6: Simulation 1 – Average Pearson correlation from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated regression coefficients. We then calculate the Pearson correlation between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

E.2 Simulation 2

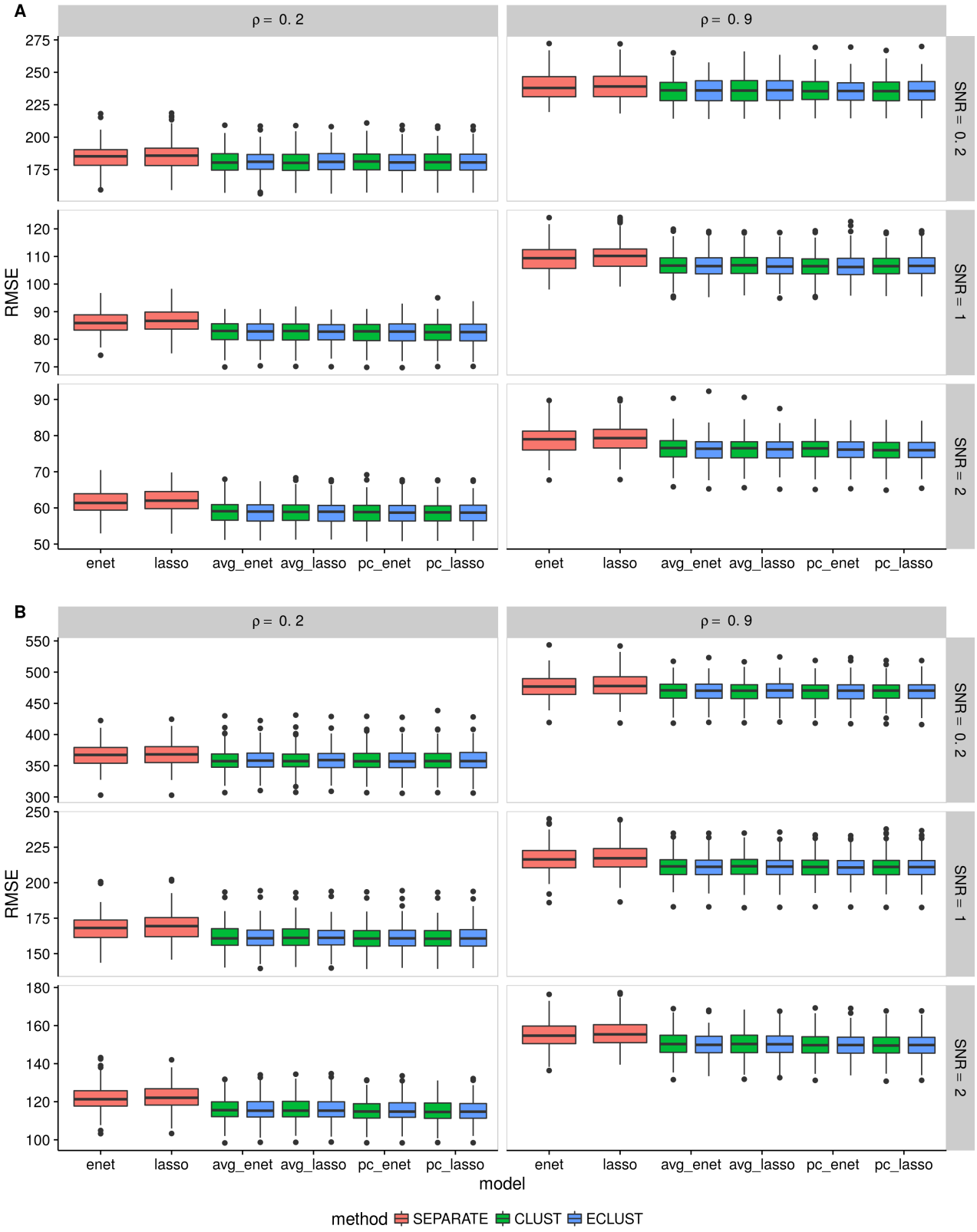


Figure E.7: Simulation 2 – Root mean squared error on an independent test set using the Pearson correlation as a measure of similarity from 200 simulation runs. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

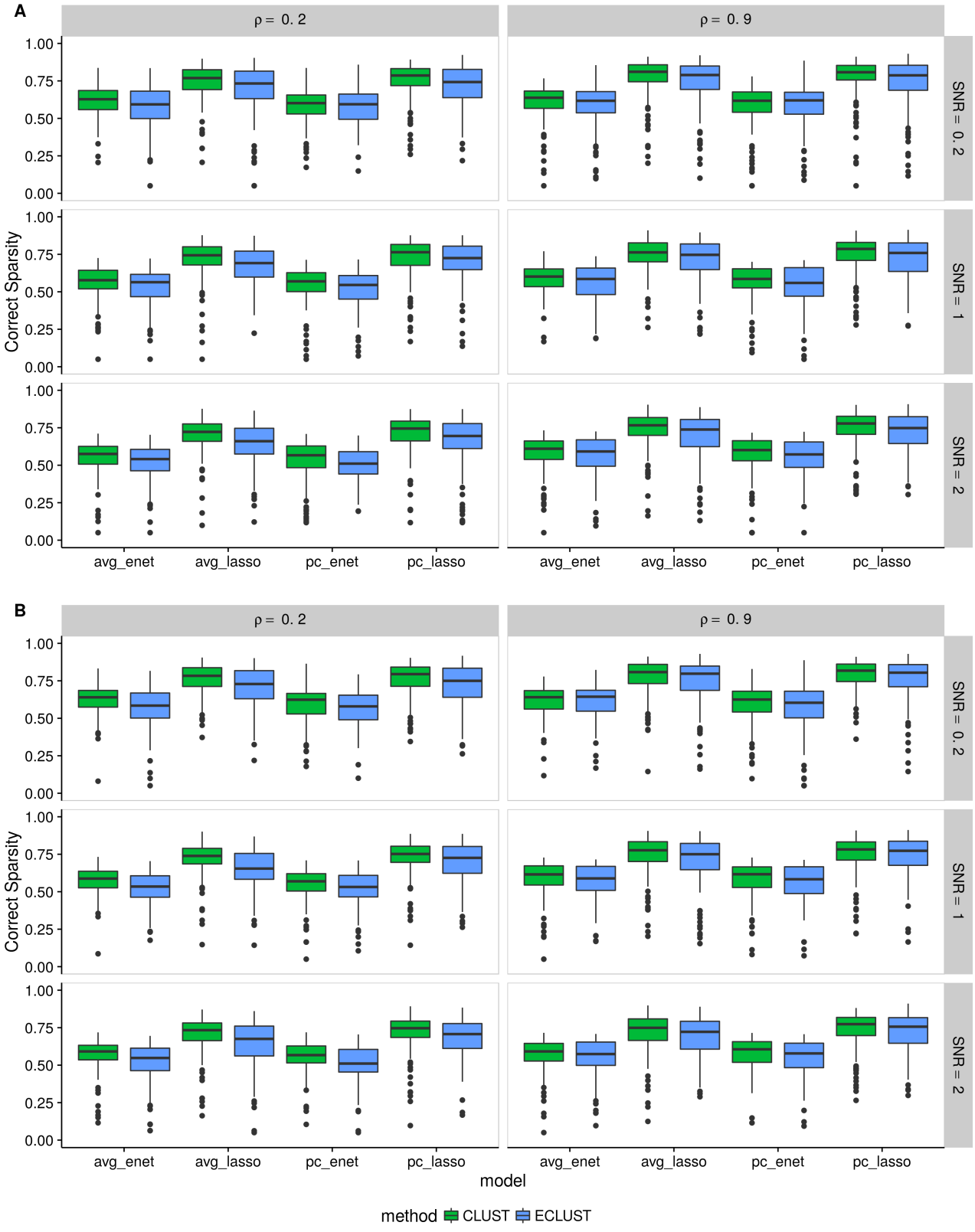


Figure E.8: Simulation 2 – Correct Sparsity based on the training set using the Pearson correlation as a measure of similarity from 200 simulation runs. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

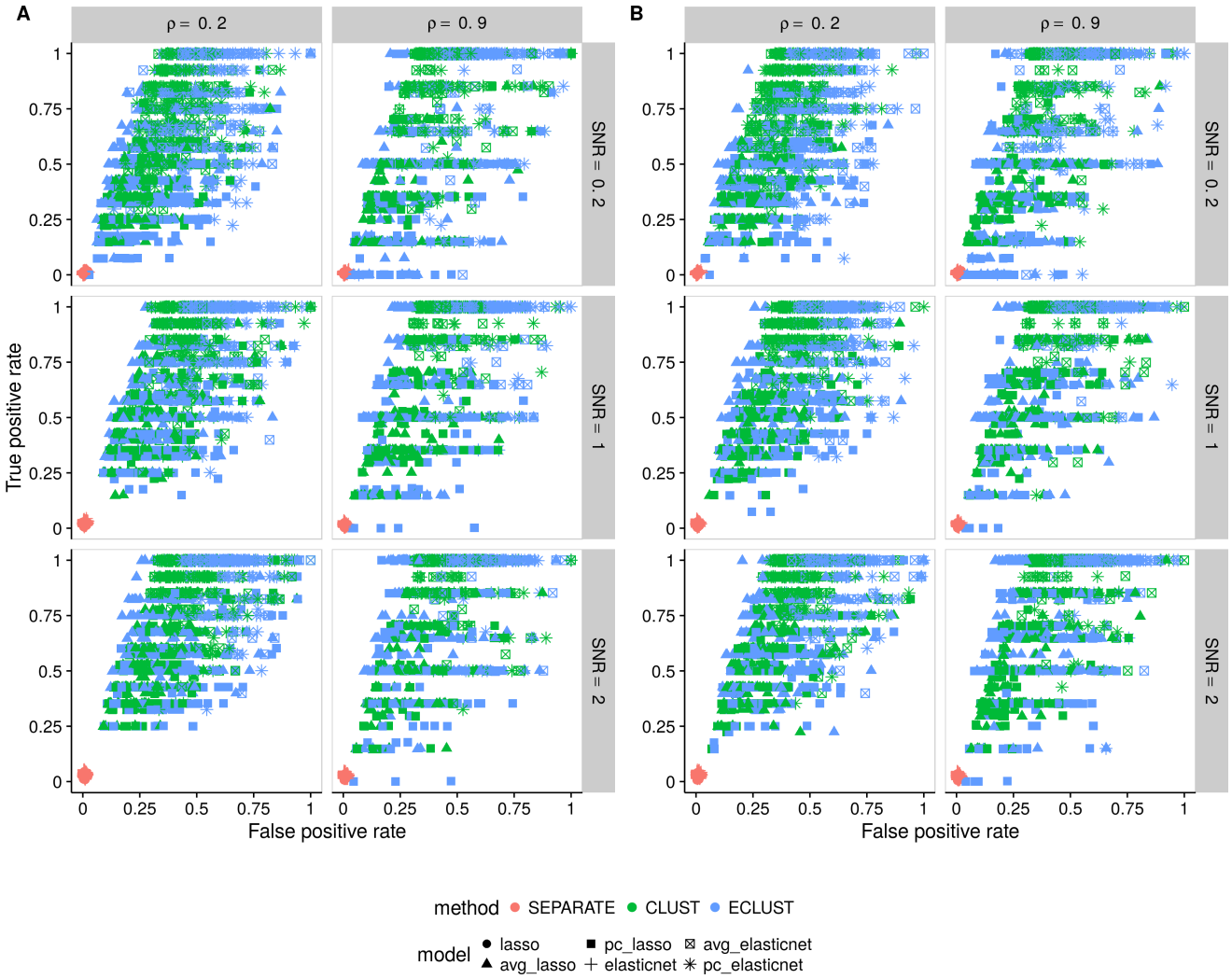


Figure E.9: Simulation 2 – True positive rate vs. false positive rate based on the training set using the Pearson correlation as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Each point represents 1 simulation run (there are a total of 200 simulation runs). Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

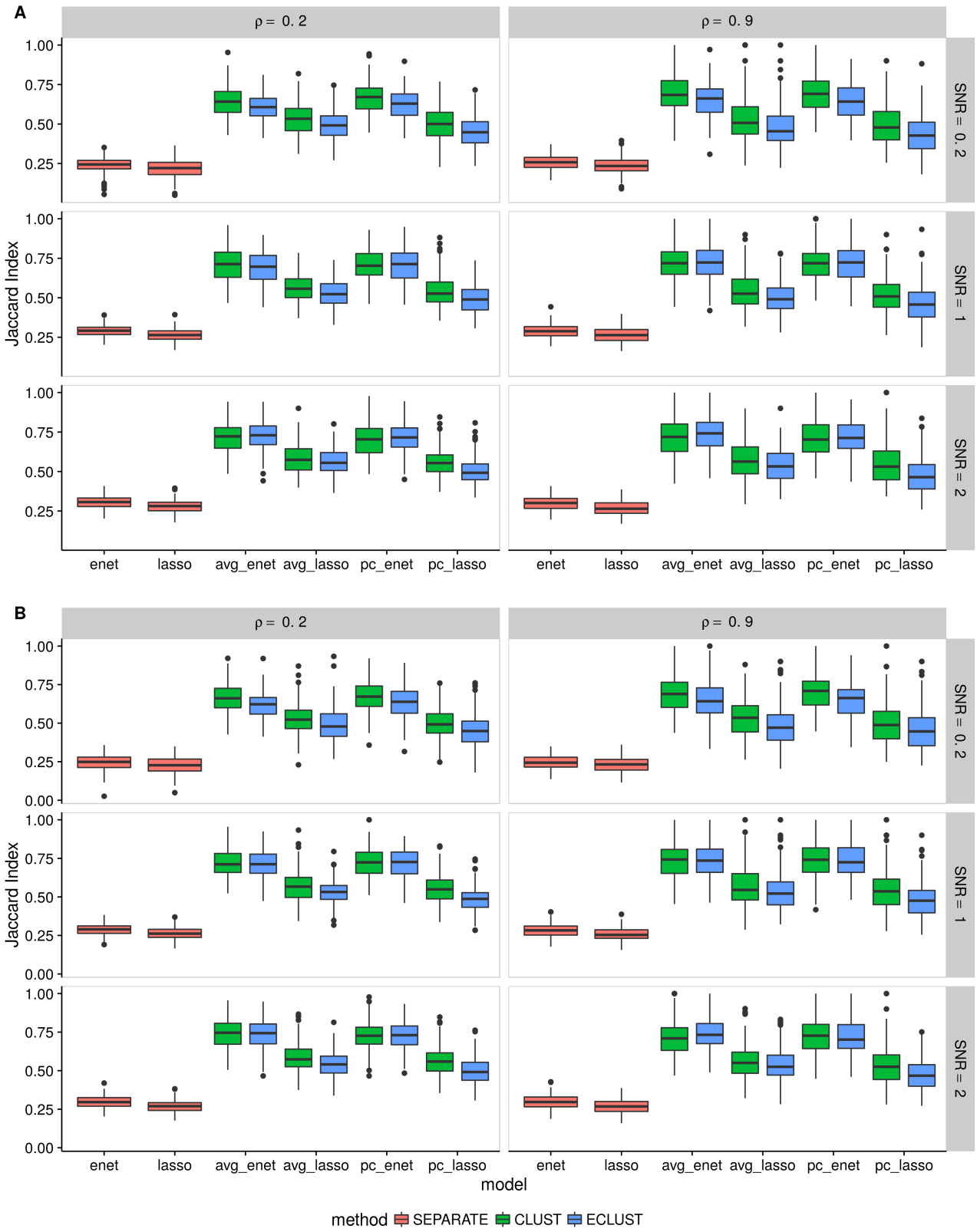


Figure E.10: Simulation 2 – Average Jaccard Index from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. We fit the model to each of the 10 CV folds resulting in 10 sets of selected predictors. We then calculate the Jaccard Index between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

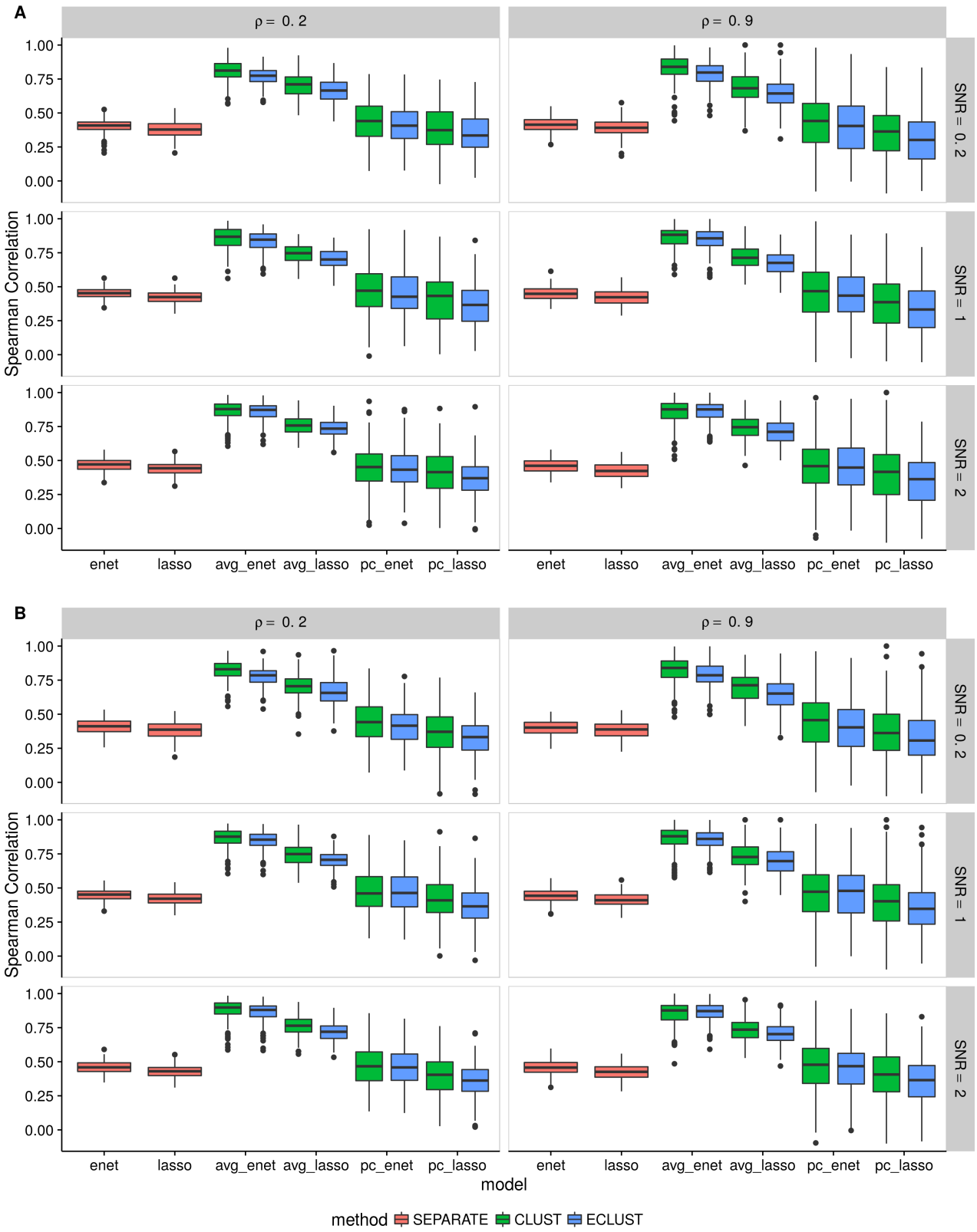


Figure E.11: Simulation 2 – Average Spearman correlation from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated regression coefficients. We then calculate the Spearman correlation between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

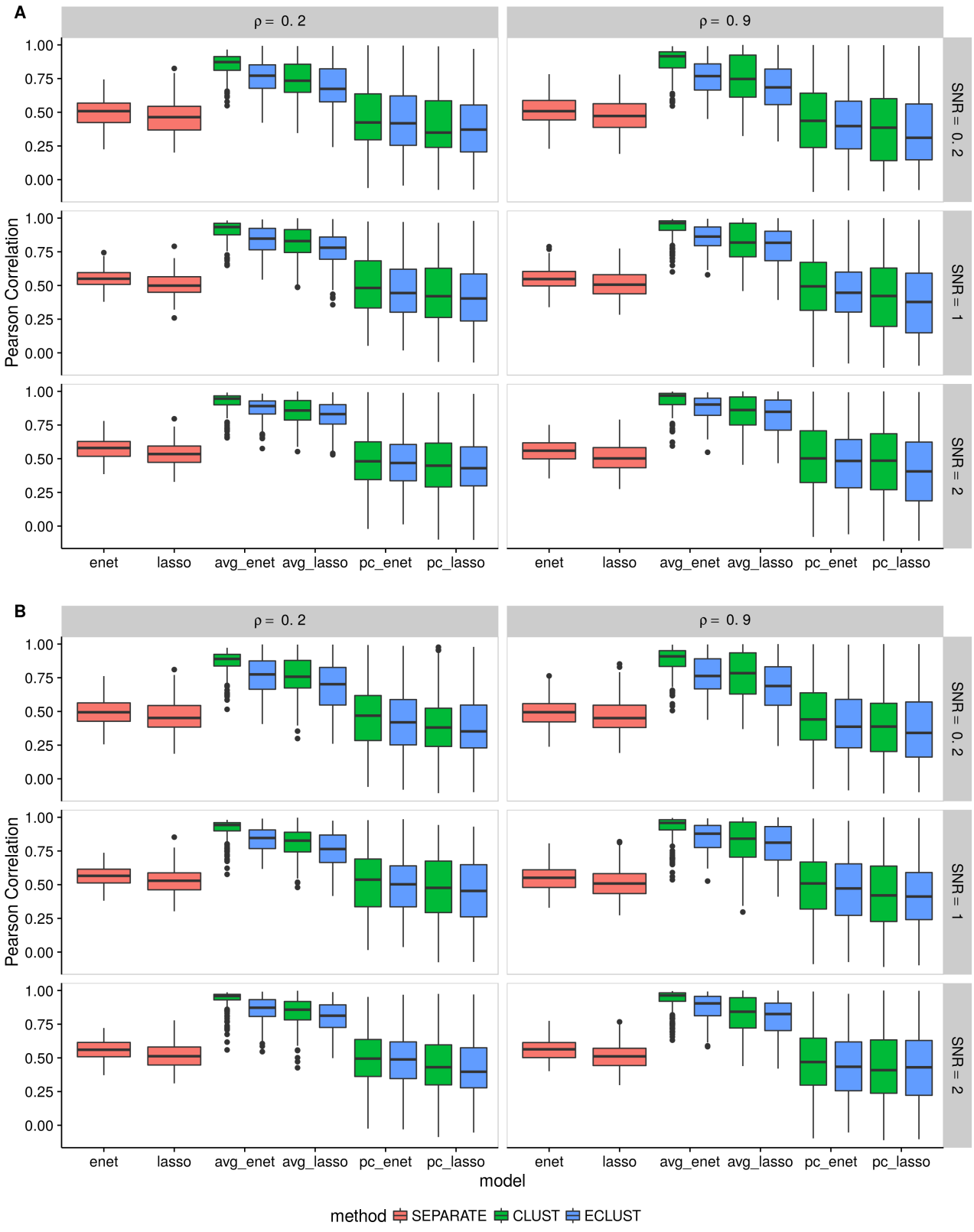


Figure E.12: Simulation 3 – Average Pearson correlation from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated regression coefficients. We then calculate the Pearson correlation between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

E.3 Simulation 3

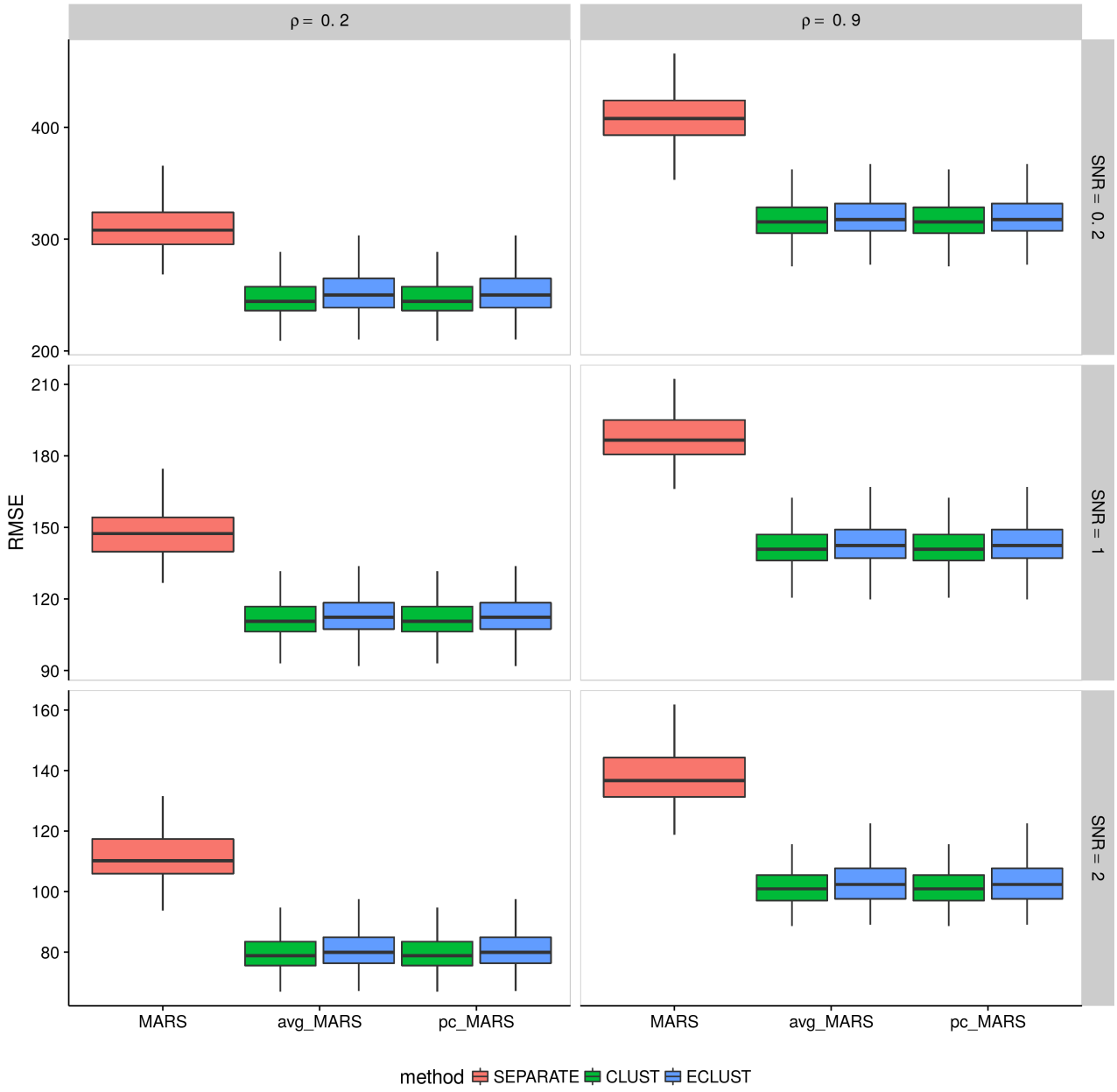


Figure E.13: Simulation 3 – Root mean squared error on an independent test set using the Pearson correlation as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

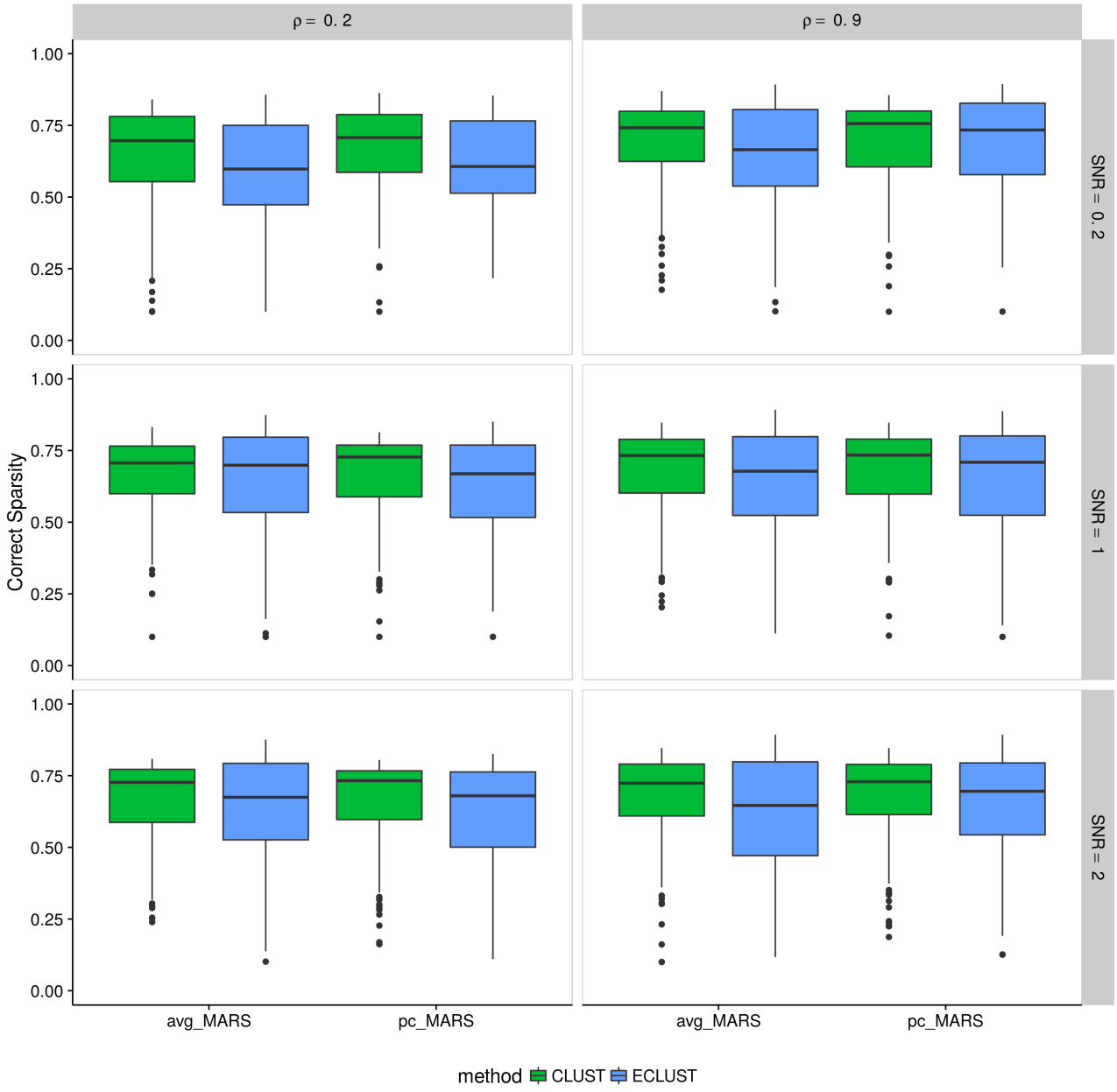


Figure E.14: Simulation 3 – Correct Sparsity based on the training set using the Pearson correlation as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

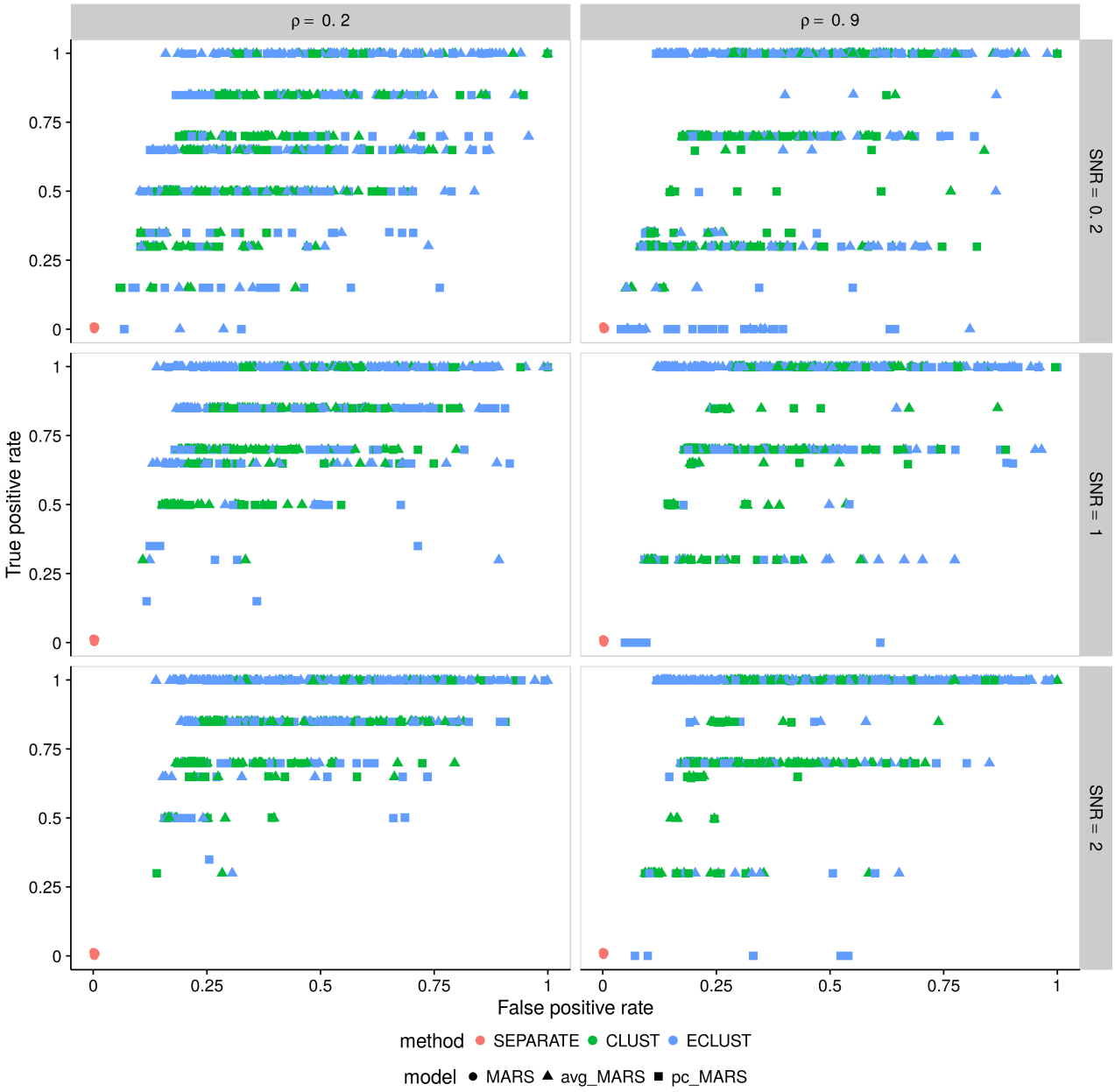


Figure E.15: Simulation 3 – True positive rate vs. false positive rate based on the training set using the Pearson correlation as a measure of similarity. Each point represents 1 simulation run (there are a total of 200 simulation runs). Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

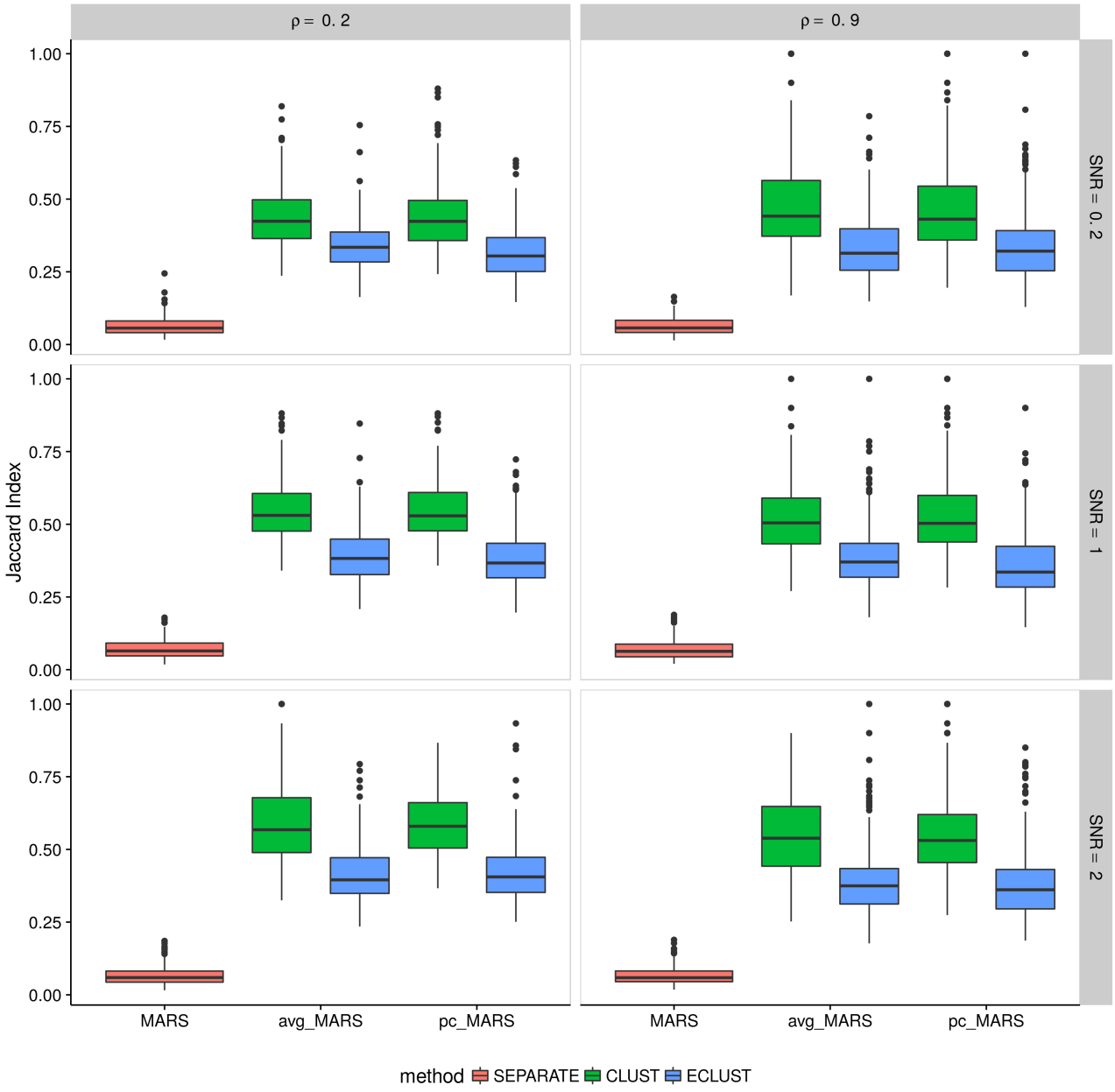


Figure E.16: Simulation 3 – Average Jaccard Index from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of selected predictors. We then calculate the Jaccard Index between all $\binom{10}{2}$ possible combinations of these sets and take the average. This process is repeated for each of the 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.