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

Sahir Rai Bhatnagar, Yi Yang, Budhachandra Khundrakpam, Alan Evans, Mathieu Blanchette, Luigi Bouchard, Celia MT Greenwood

October 4, 2017

Contents

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 unweighted 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) $\Omega = [\omega_{ij}]$:

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

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 Binary Outcome Simulation Results

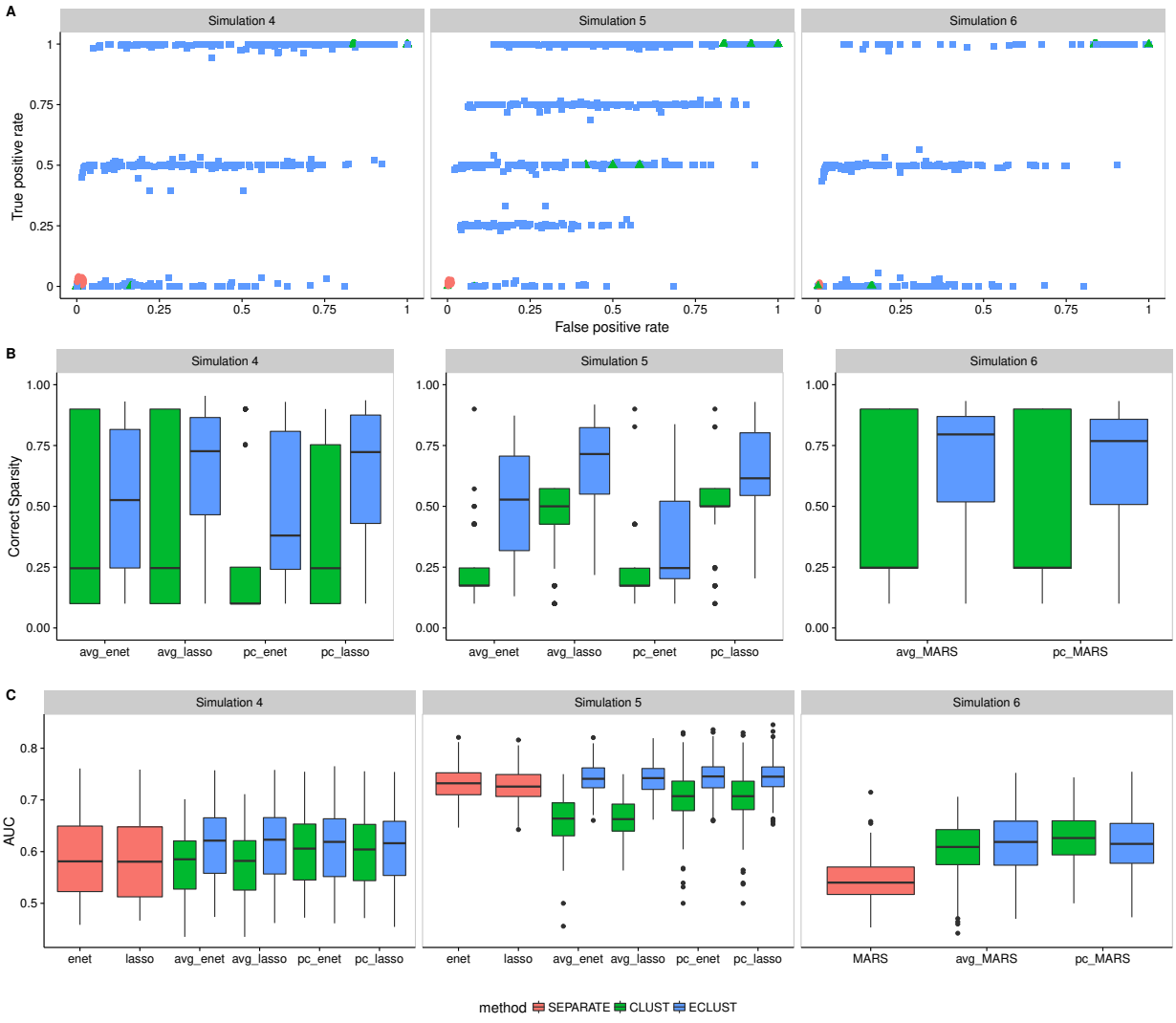


Figure S1: Model fit results from simulations 4, 5 and 6 for $SNR = 1$, $\rho = 0.9$, and $\alpha_j \sim \text{Unif}\left[\log(1.9), \log(2.1)\right]$. SEPARATE results are in pink, CLUST in green and ECLUST in blue.

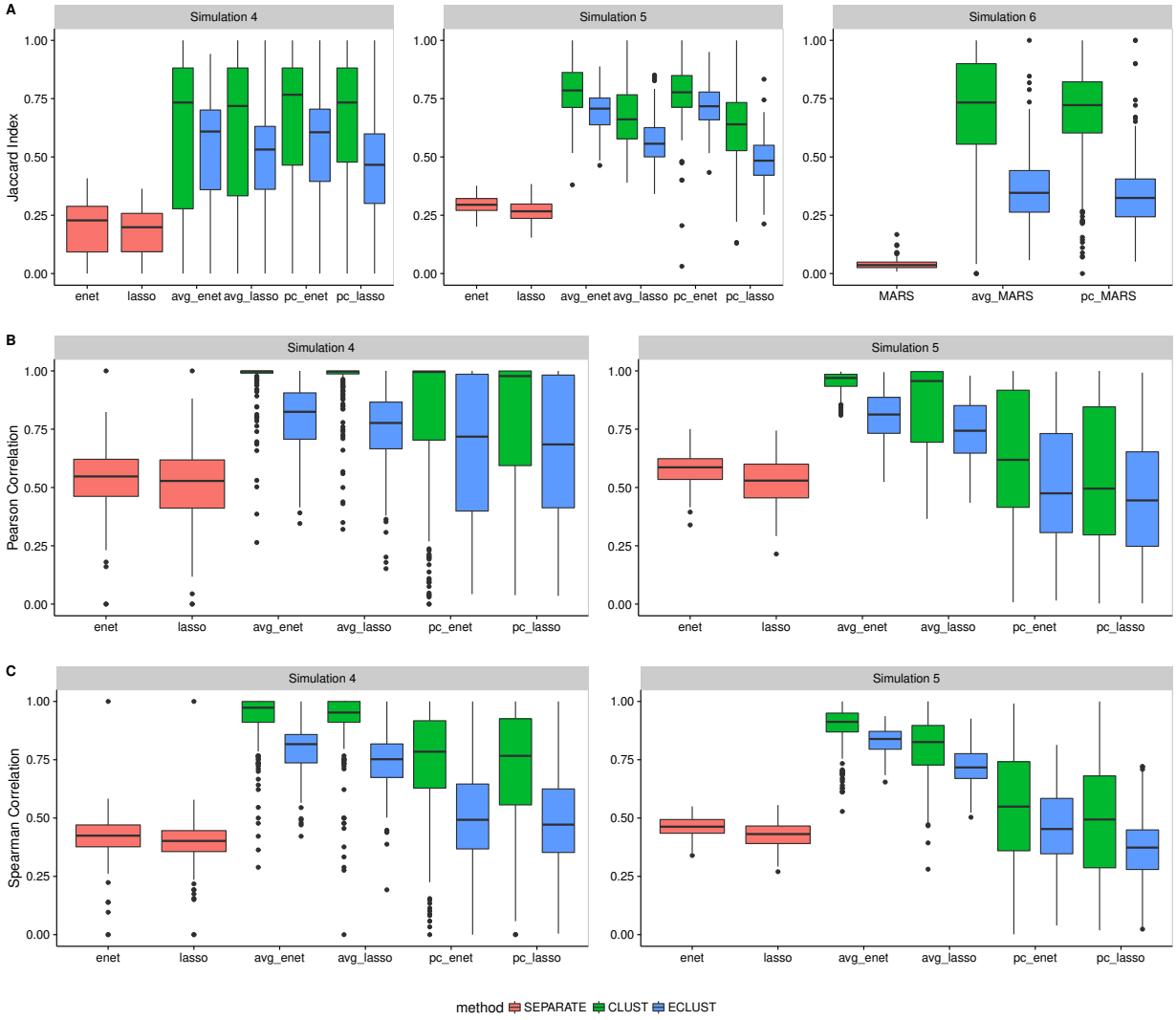


Figure S2: Stability results from simulations 4, 5 and 6 for $SNR = 1$, $\rho = 0.9$, and $\alpha_j \sim \text{Unif}\left[\log(1.9), \log(2.1)\right]$. SEPARATE results are in pink, CLUST in green and ECLUST in blue.

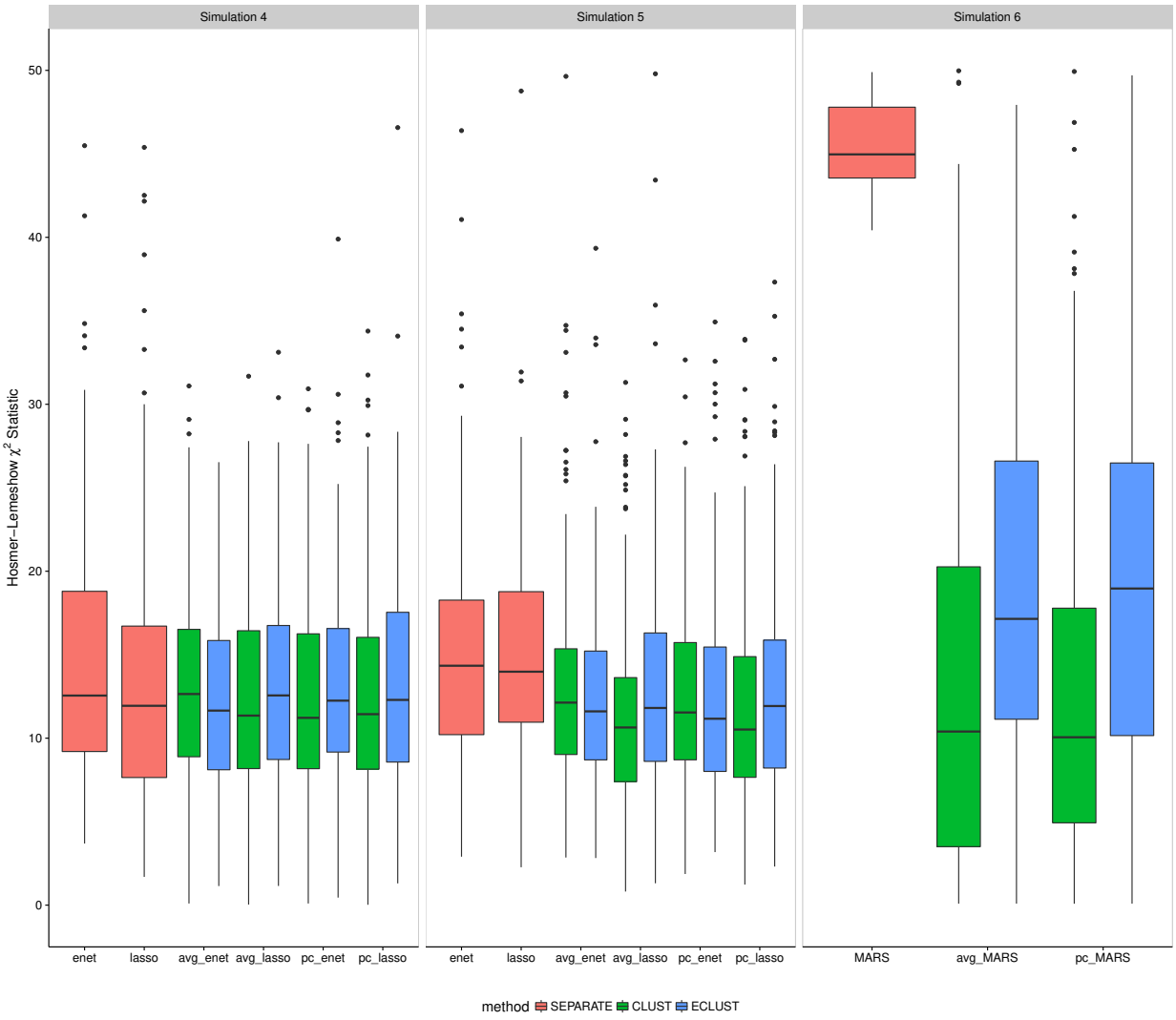


Figure S3: Hosmer-Lemeshow statistics from simulations 4, 5 and 6 for $SNR = 1$, $\rho = 0.9$, and $\alpha_j \sim \text{Unif}\left[\log(1.9), \log(2.1)\right]$. SEPARATE results are in pink, CLUST in green and ECLUST in blue.

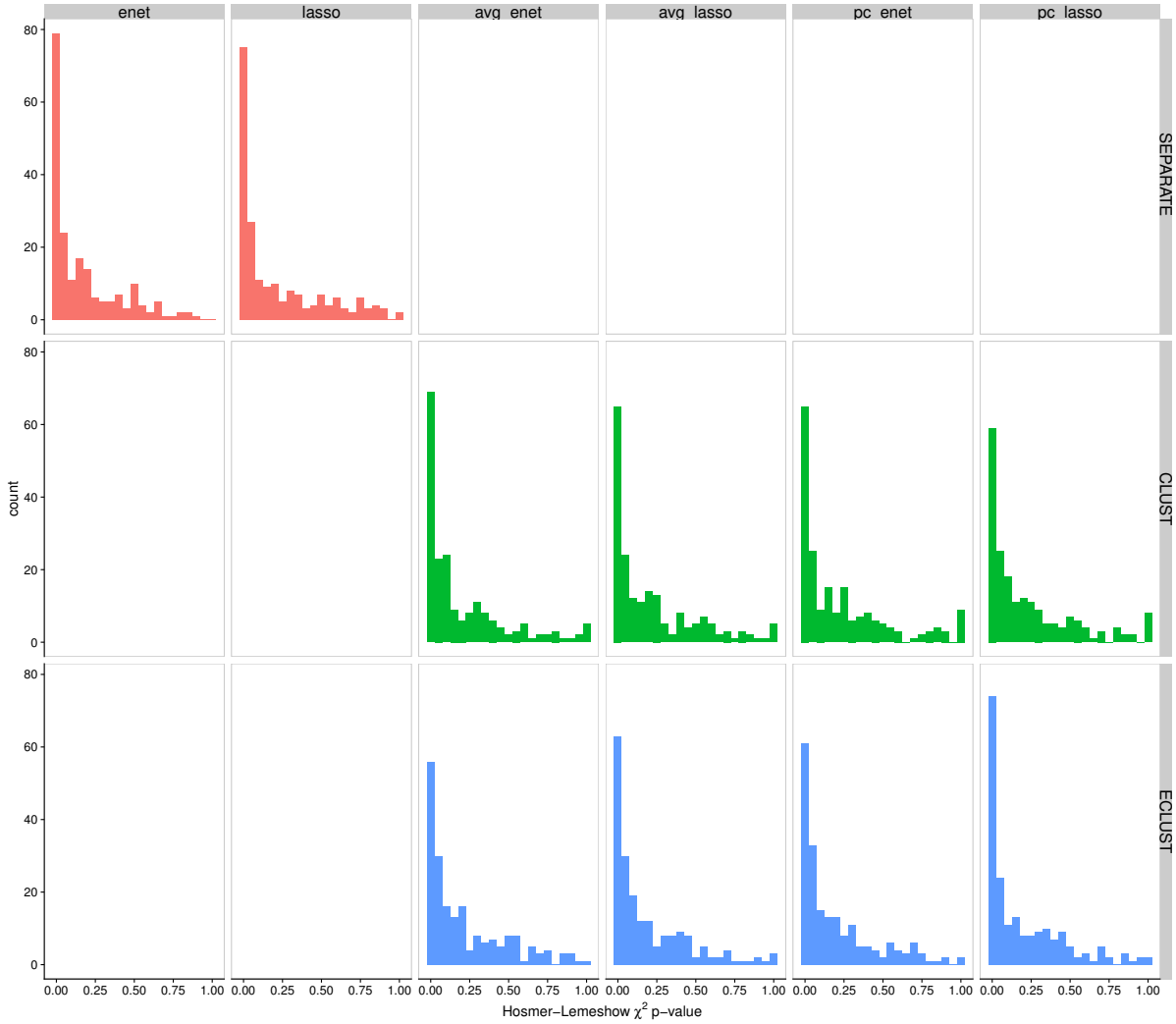


Figure S4: Hosmer-Lemeshow p-values from simulation 4 for $SNR = 1$, $\rho = 0.9$, and $\alpha_j \sim \text{Unif}\left[\log(1.9), \log(2.1)\right]$. SEPARATE results are in pink, CLUST in green and ECLUST in blue.

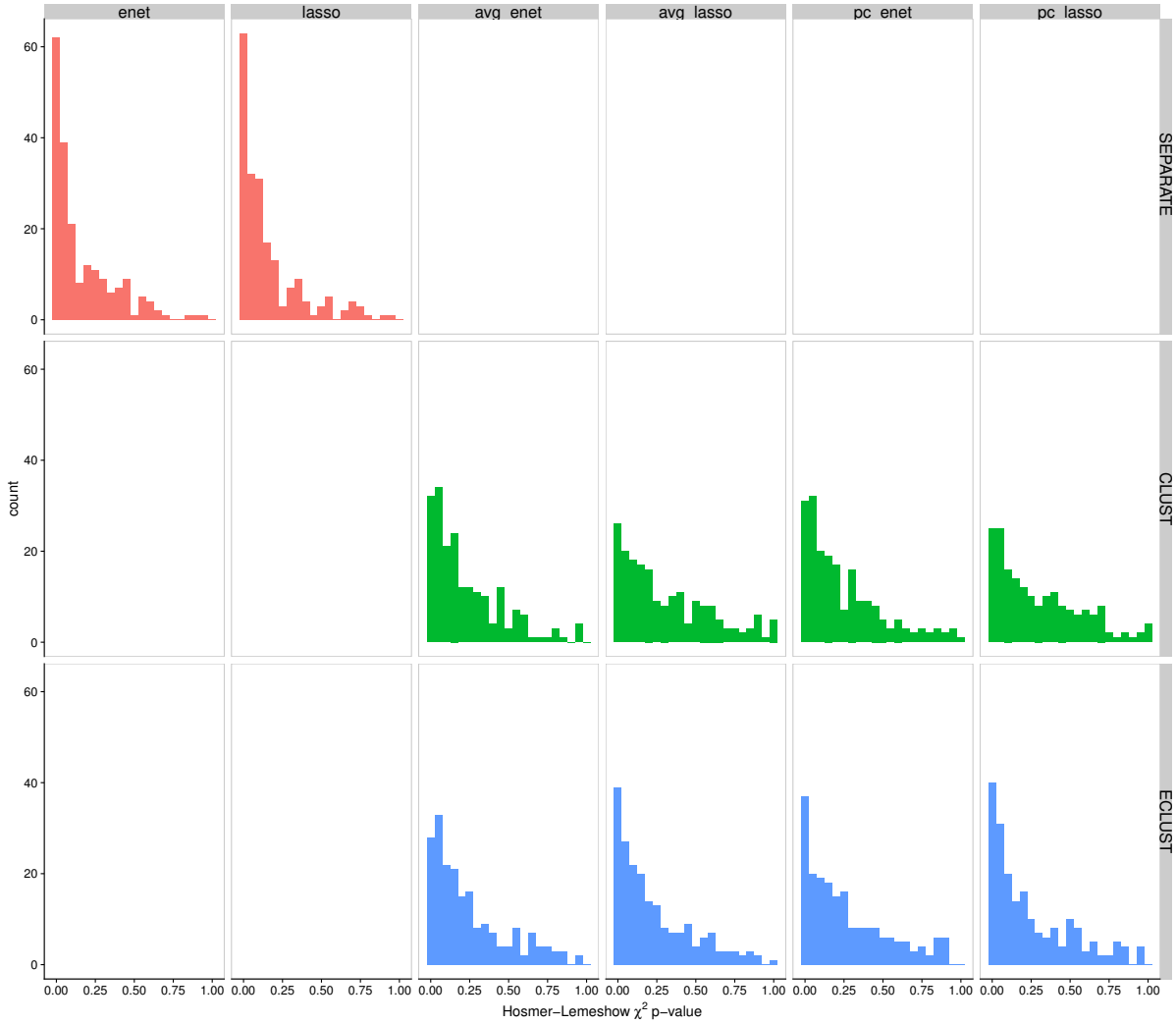


Figure S5: Hosmer-Lemeshow p-values from simulation 5 for $SNR = 1$, $\rho = 0.9$, and $\alpha_j \sim \text{Unif}\left[\log(1.9), \log(2.1)\right]$. SEPARATE results are in pink, CLUST in green and ECLUST in blue.

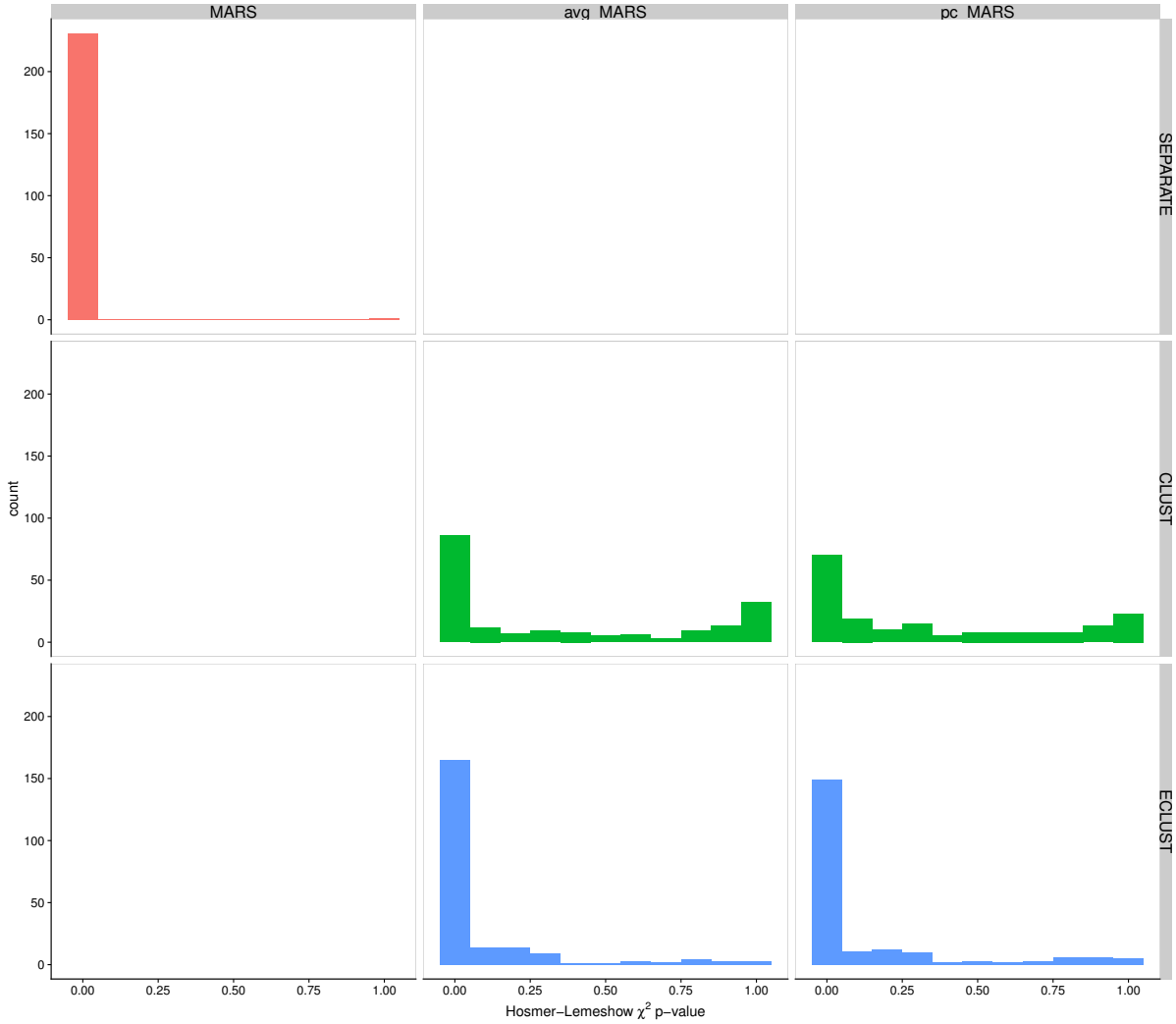


Figure S6: Hosmer-Lemeshow p-values from simulation 6 for $SNR = 1$, $\rho = 0.9$, and $\alpha_j \sim \text{Unif}\left[\log(1.9), \log(2.1)\right]$. SEPARATE results are in pink, CLUST in green and ECLUST in blue.

C Analysis of Clusters

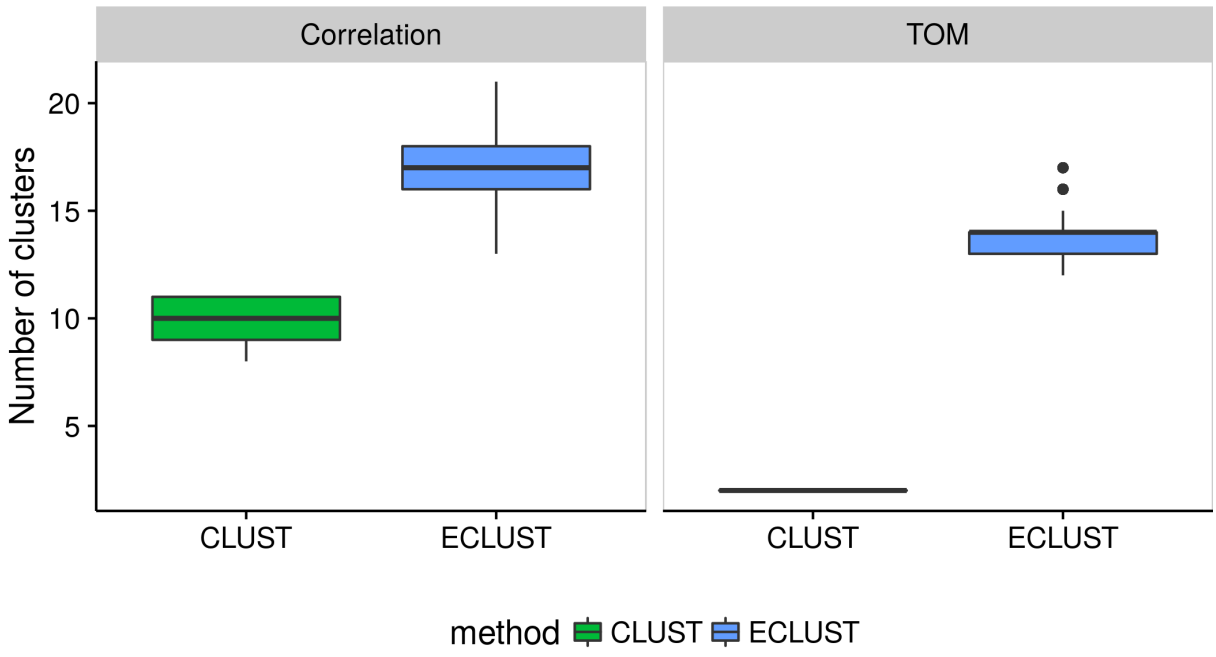


Figure S7: Number of estimated clusters from applying the `dynamicTreeCut` algorithm to hierarchical clustering of the dissimilarity matrix with average linkage. Left panel: CLUST uses $1 - Cor(X_{all})$ and ECLUST uses the euclidean distance of $Cor(X_{\text{diff}})$ as measures of dissimilarity. Right 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.

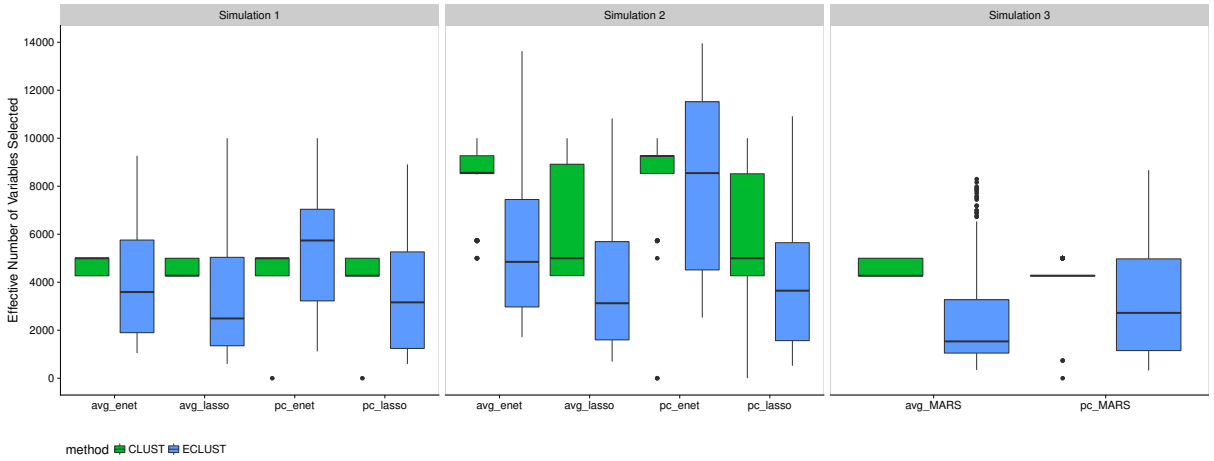


Figure S8: Effective number of selected variables for simulations 1-3 for $SNR = 1, \rho = 0.9$. A variable was considered "selected" if its corresponding cluster representative was selected.

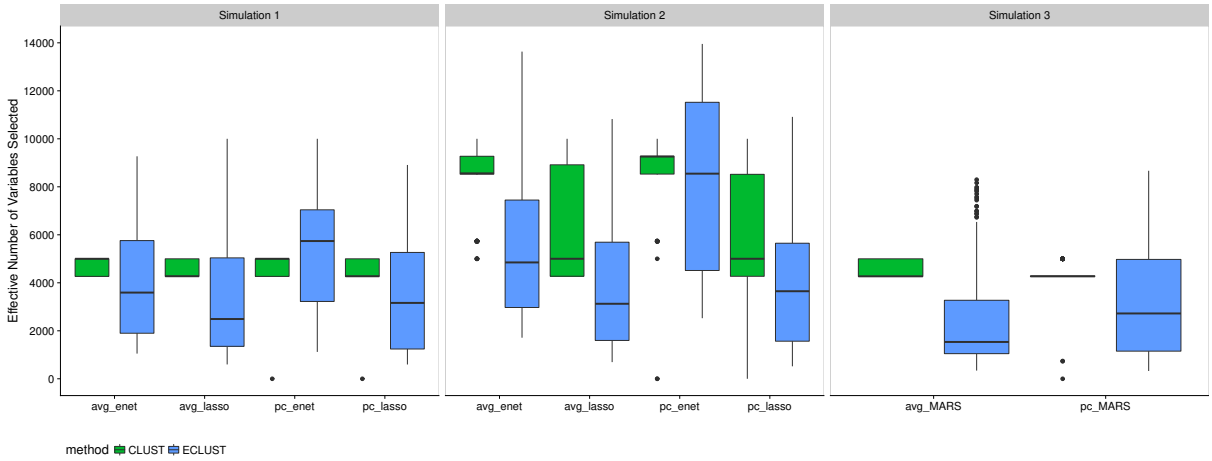


Figure S9: Effective number of selected variables for simulations 4-6 for $SNR = 1, \rho = 0.9$ and $\alpha_j \sim \text{Unif}\left[\log(1.9), \log(2.1)\right]$. A variable was considered "selected" if its corresponding cluster representative was selected.

D Simulation Results Using TOM as a Measure of Similarity

D.1 Simulation 1

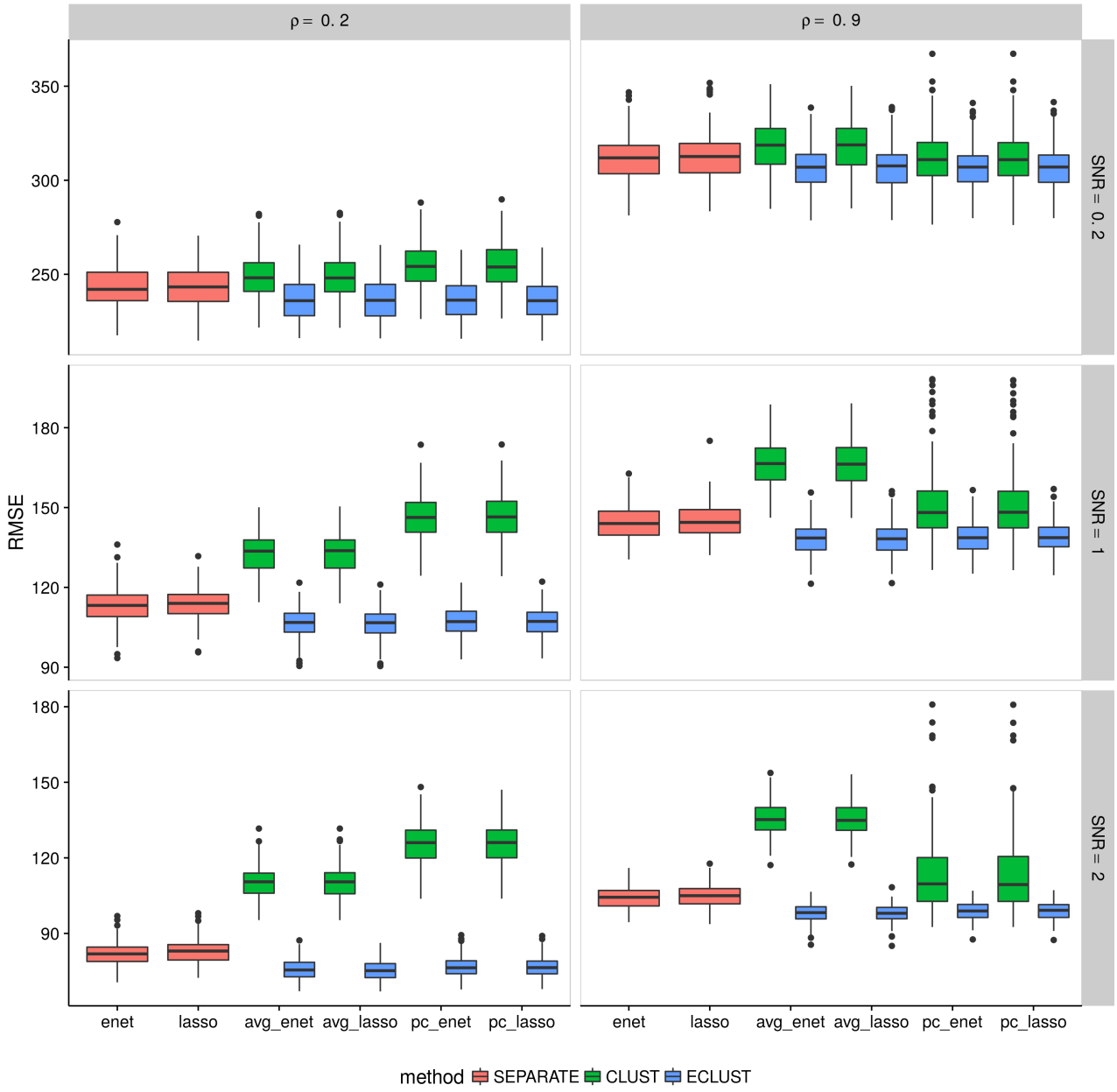


Figure S10: Simulation 1 – Root mean squared error on an independent test set using the TOM as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

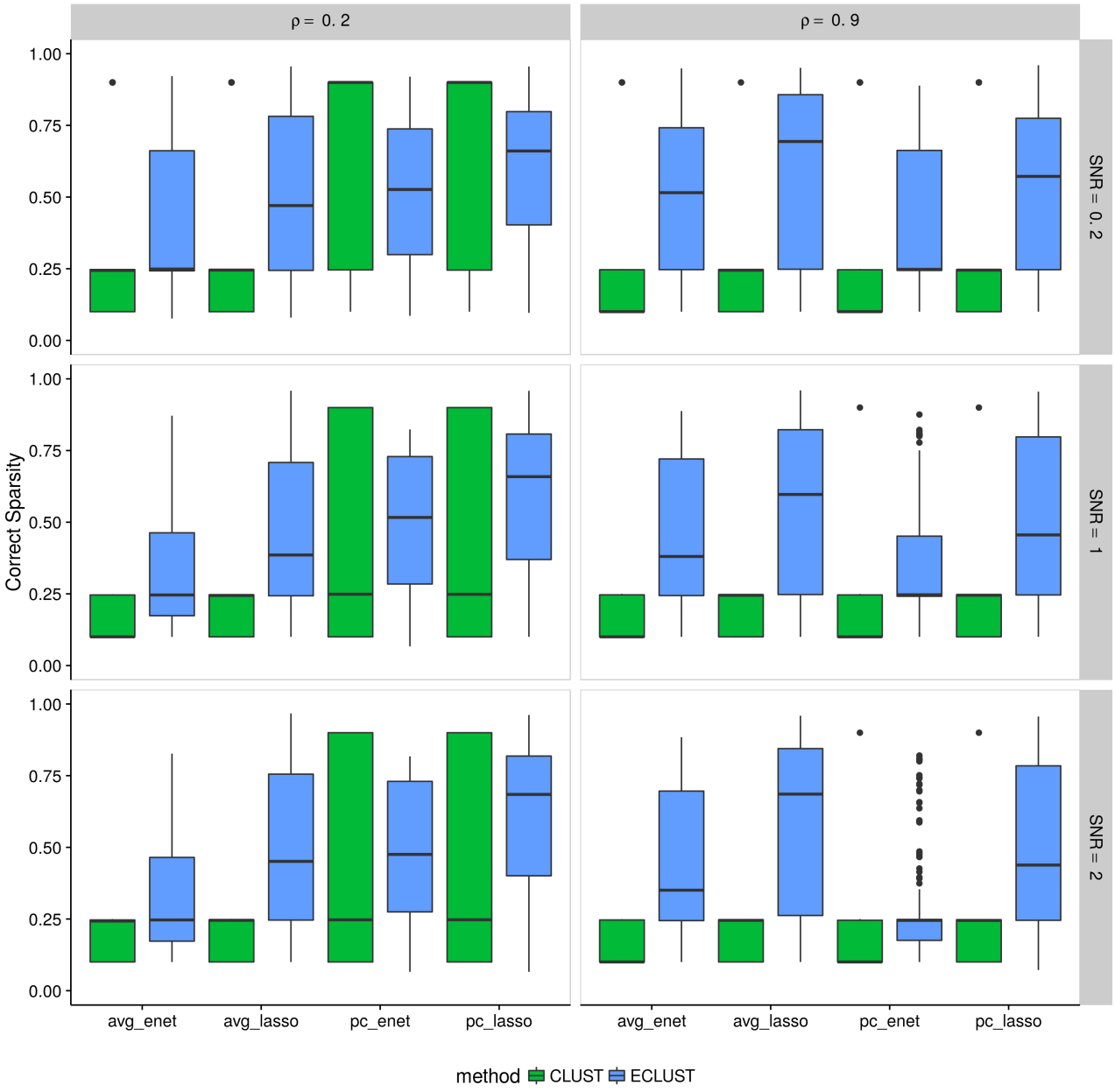


Figure S11: Simulation 1 – Correct Sparsity based on the training set using the TOM as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

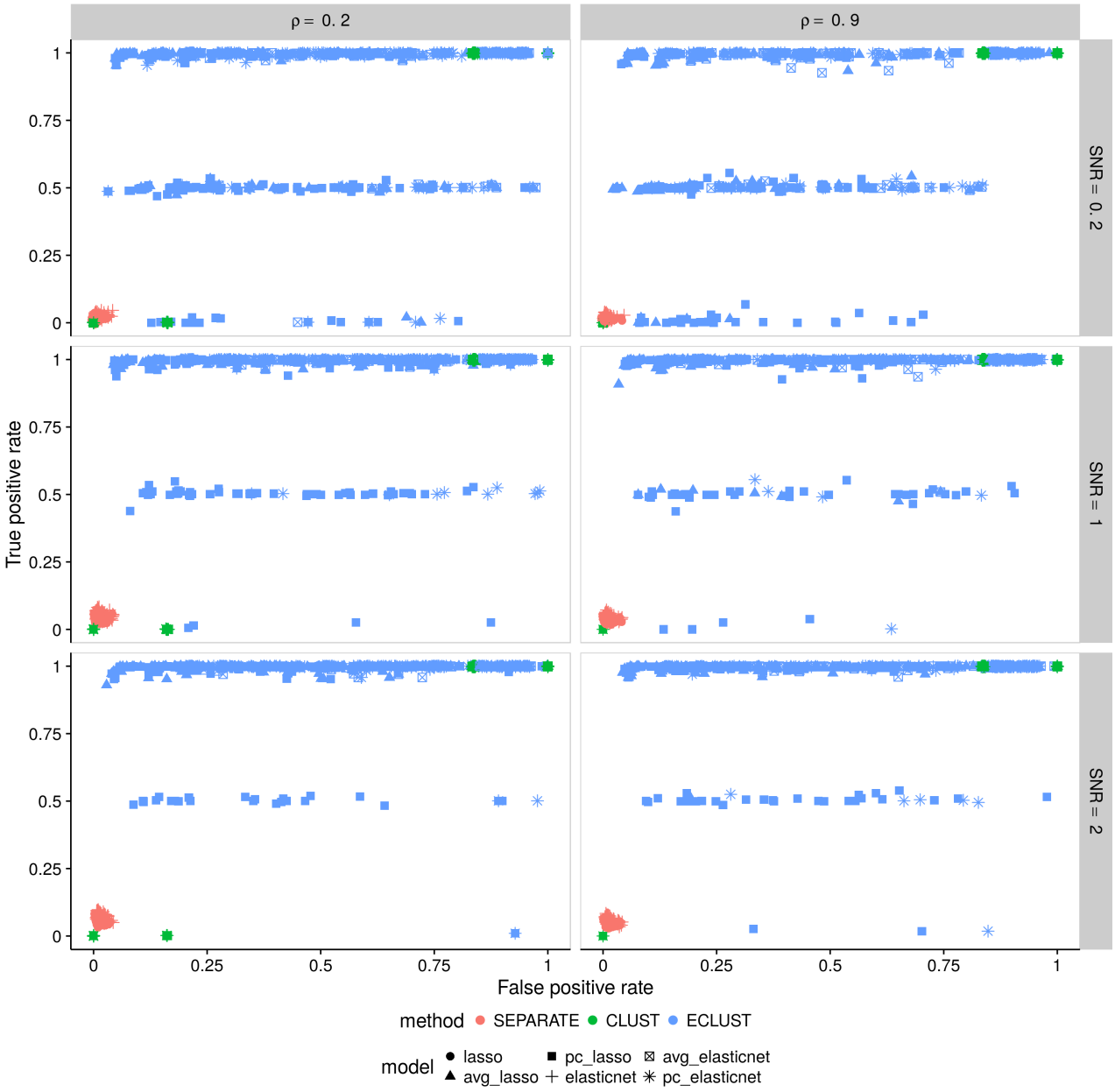


Figure S12: Simulation 1 – True positive rate vs. false positive rate based on the training set using the TOM as a measure of similarity. Each point represents 1 simulation run (there are a total of 200 simulation runs). Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

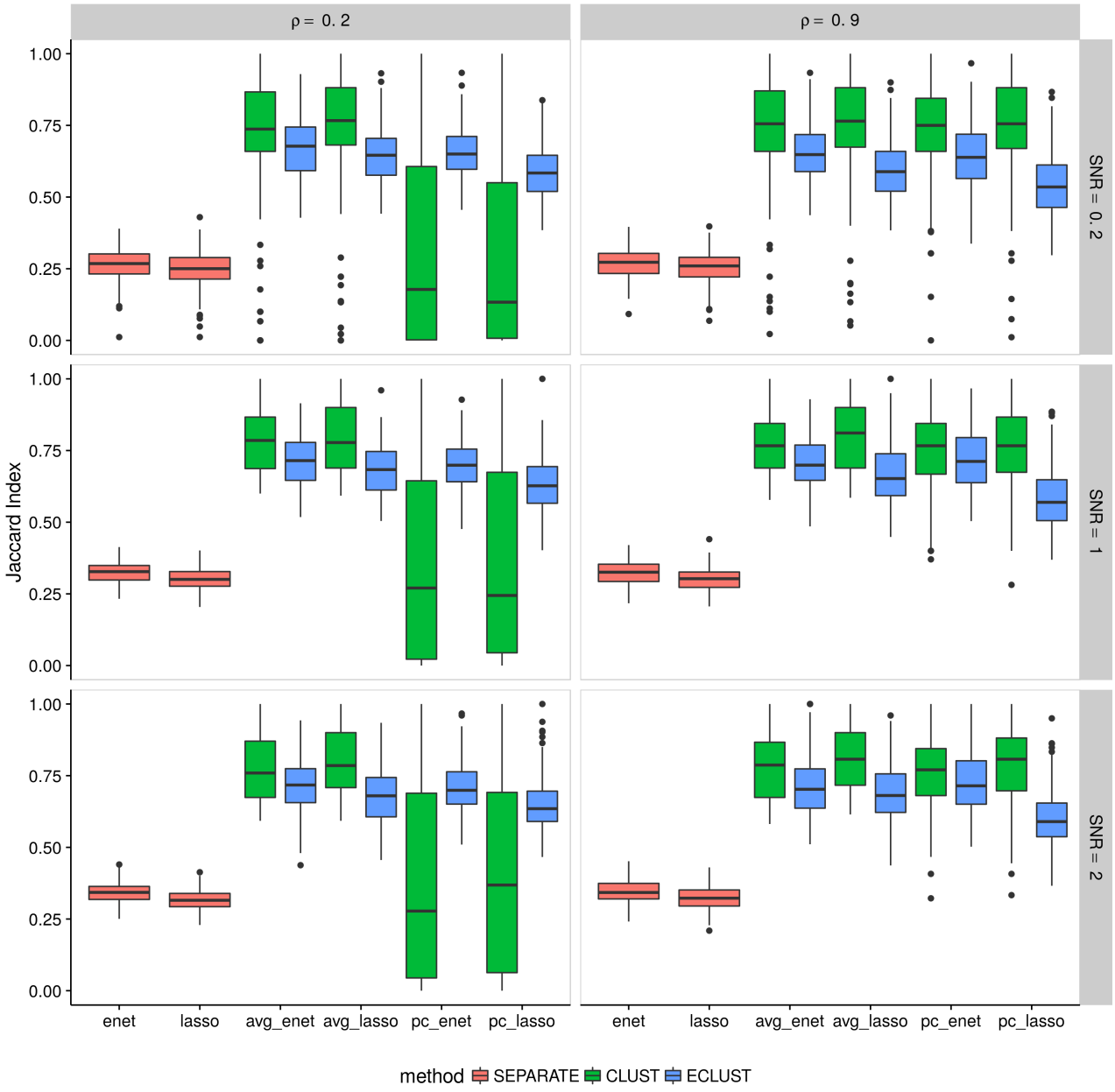


Figure S13: Simulation 1 – Average Jaccard Index from 10 CV folds of the training set using the TOM as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of selected predictors. We then calculate the Jaccard Index between all $\binom{10}{2}$ possible combinations 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.

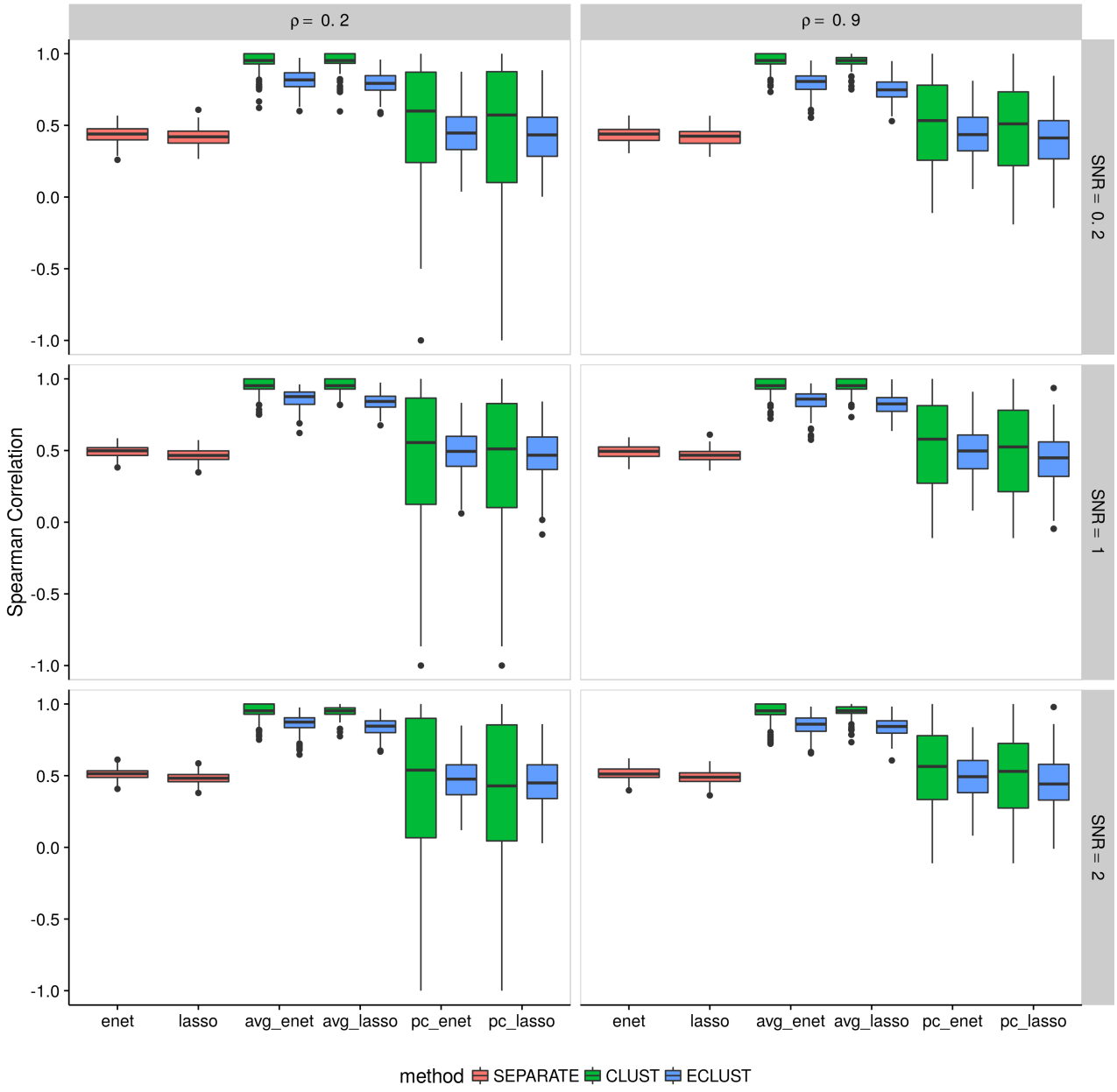


Figure S14: Simulation 1 – Average Spearman correlation from 10 CV folds of the training set using the TOM as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated regression coefficients. We then calculate the Spearman correlation between all $\binom{10}{2}$ 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.

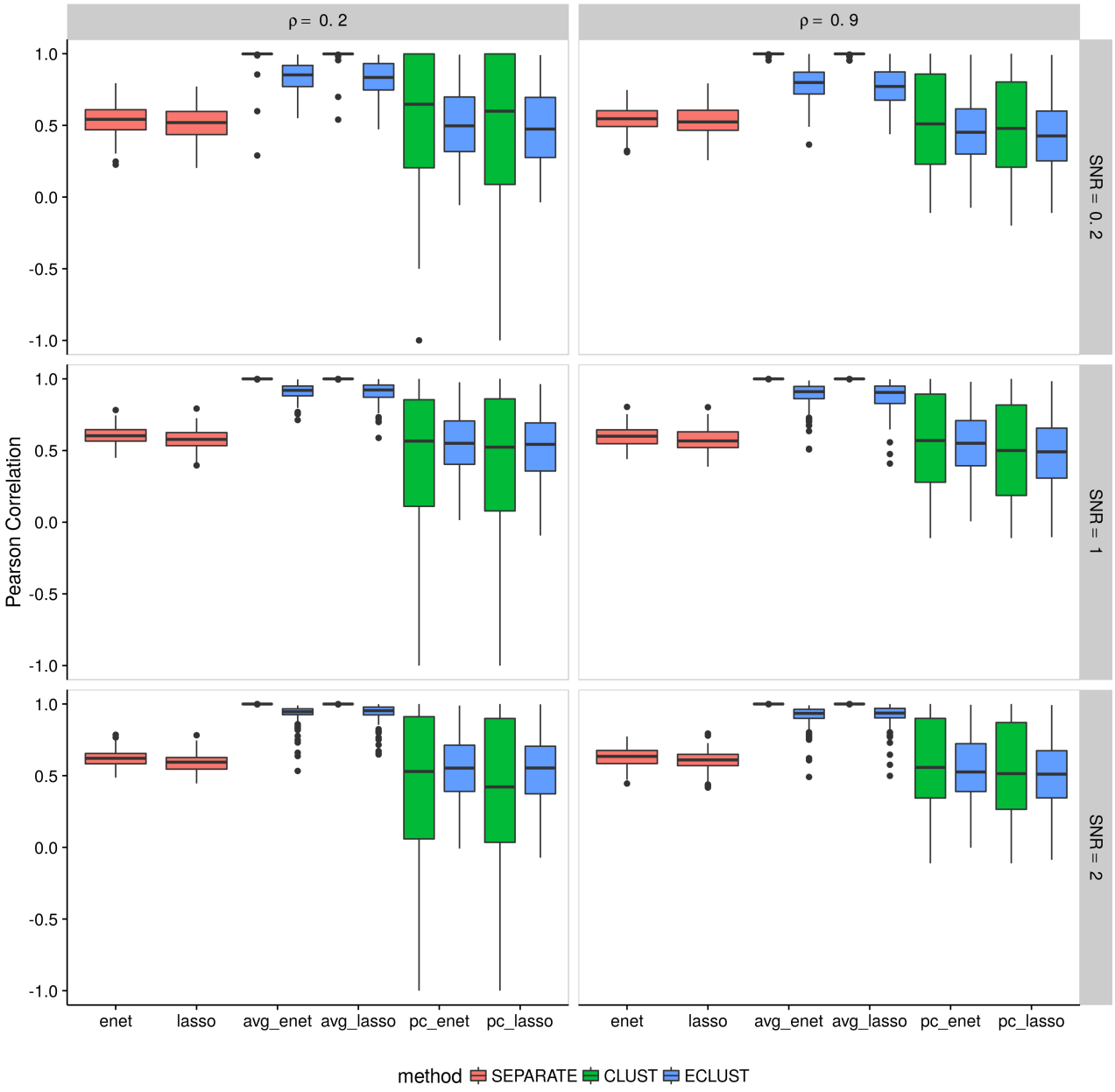


Figure S15: Simulation 1 – Average Pearson correlation from 10 CV folds of the training set using the TOM as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated regression coefficients. We then calculate the Pearson correlation between all $\binom{10}{2}$ 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

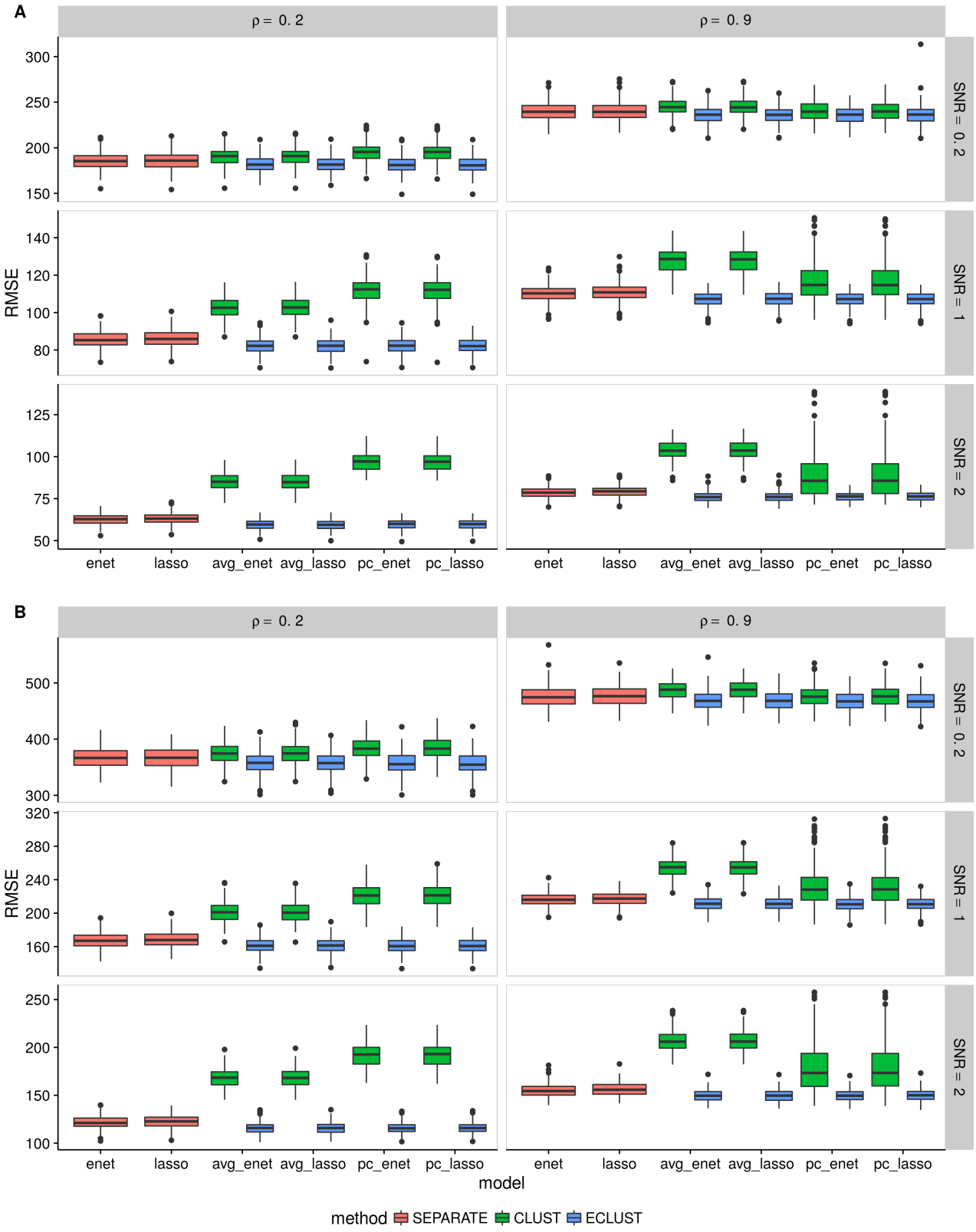


Figure S16: Simulation 2 – Root mean squared error on an independent test set using the TOM as a measure of similarity from 200 simulation runs. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

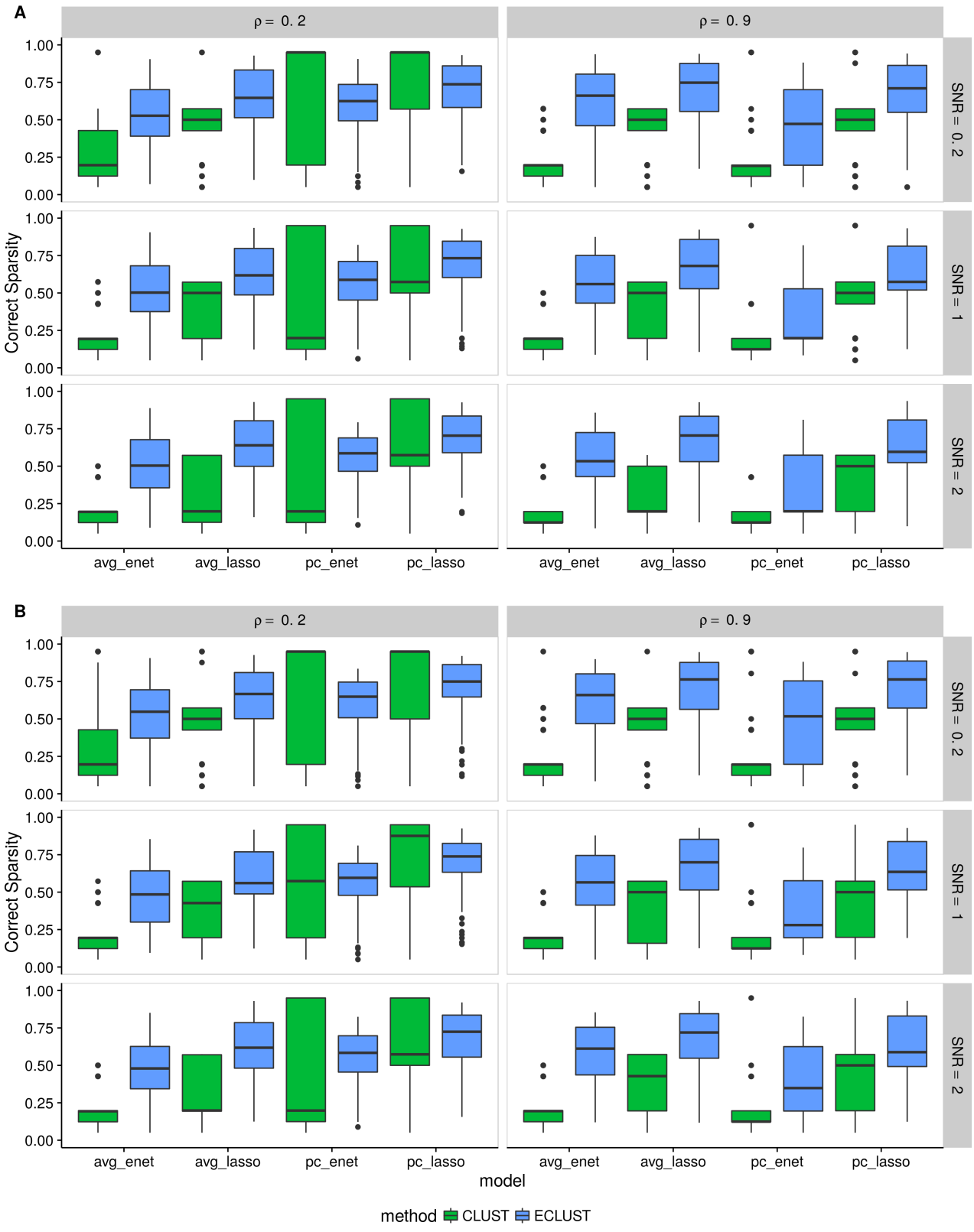


Figure S17: Simulation 2 – Correct Sparsity based on the training set using the TOM as a measure of similarity from 200 simulation runs. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

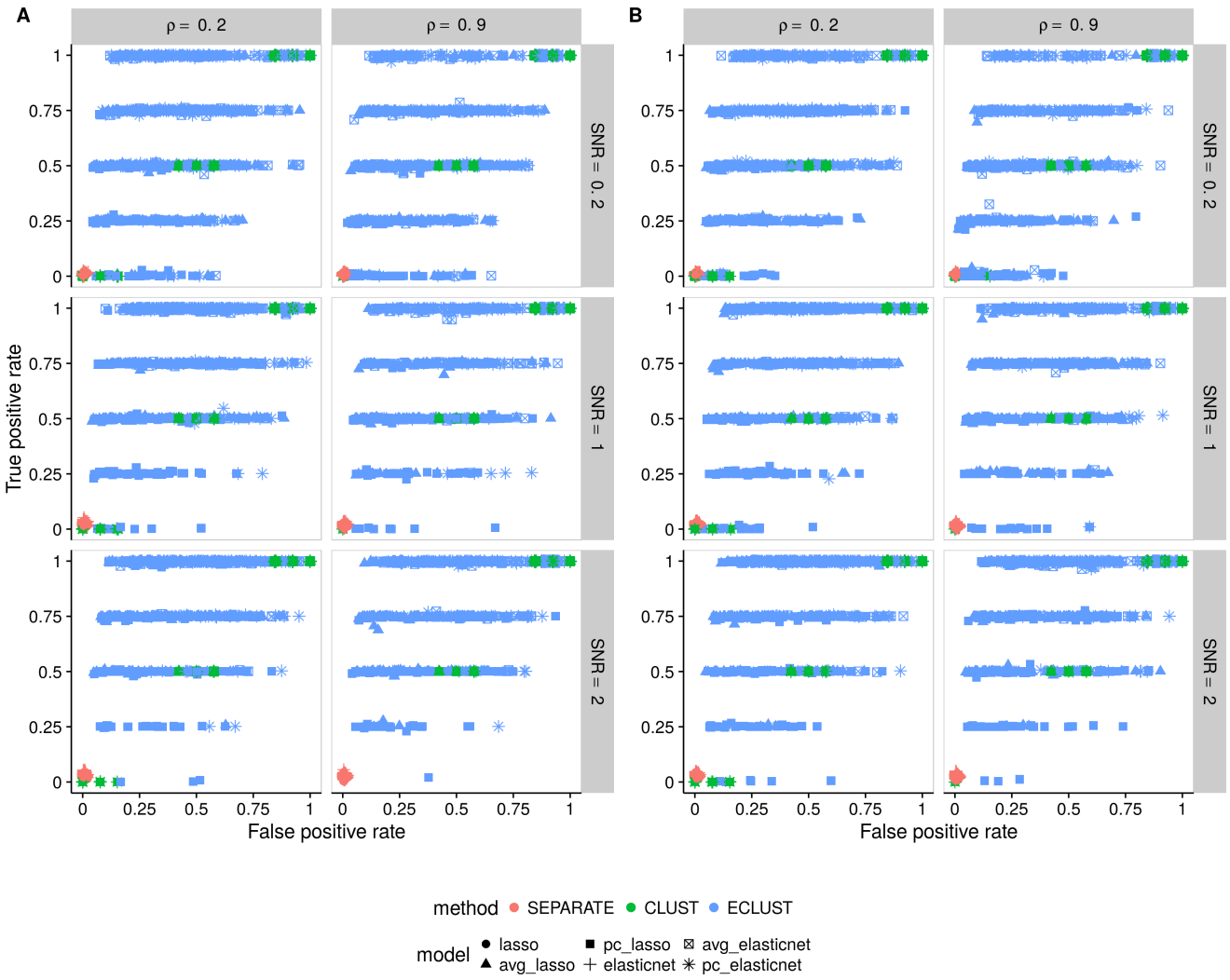


Figure S18: Simulation 2 – True positive rate vs. false positive rate based on the training set using the TOM as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Each point represents 1 simulation run (there are a total of 200 simulation runs). Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

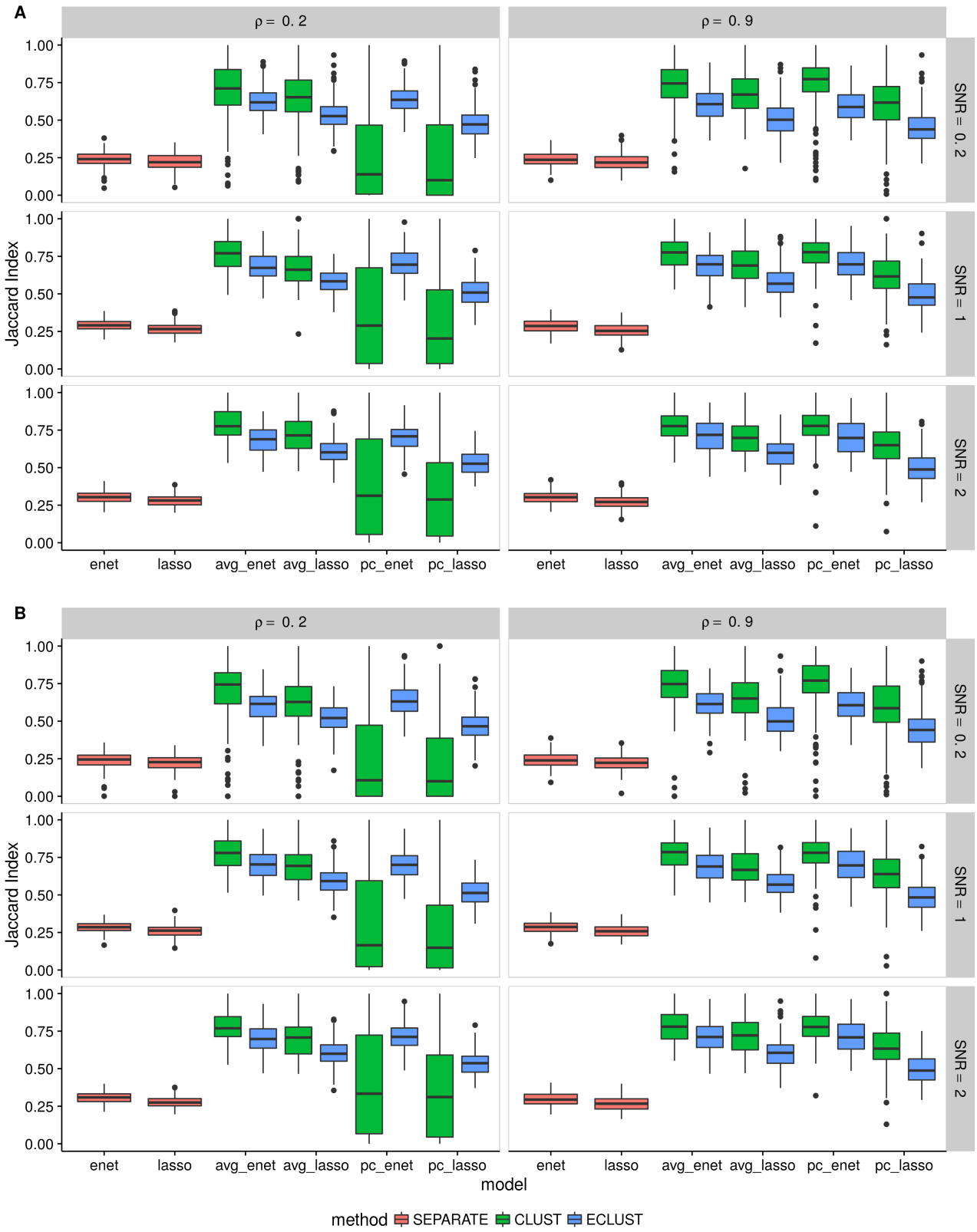


Figure S19: Simulation 2 – Average Jaccard Index from 10 CV folds of the training set using the TOM as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. We fit the model to each of the 10 CV folds resulting in 10 sets of selected predictors. 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.

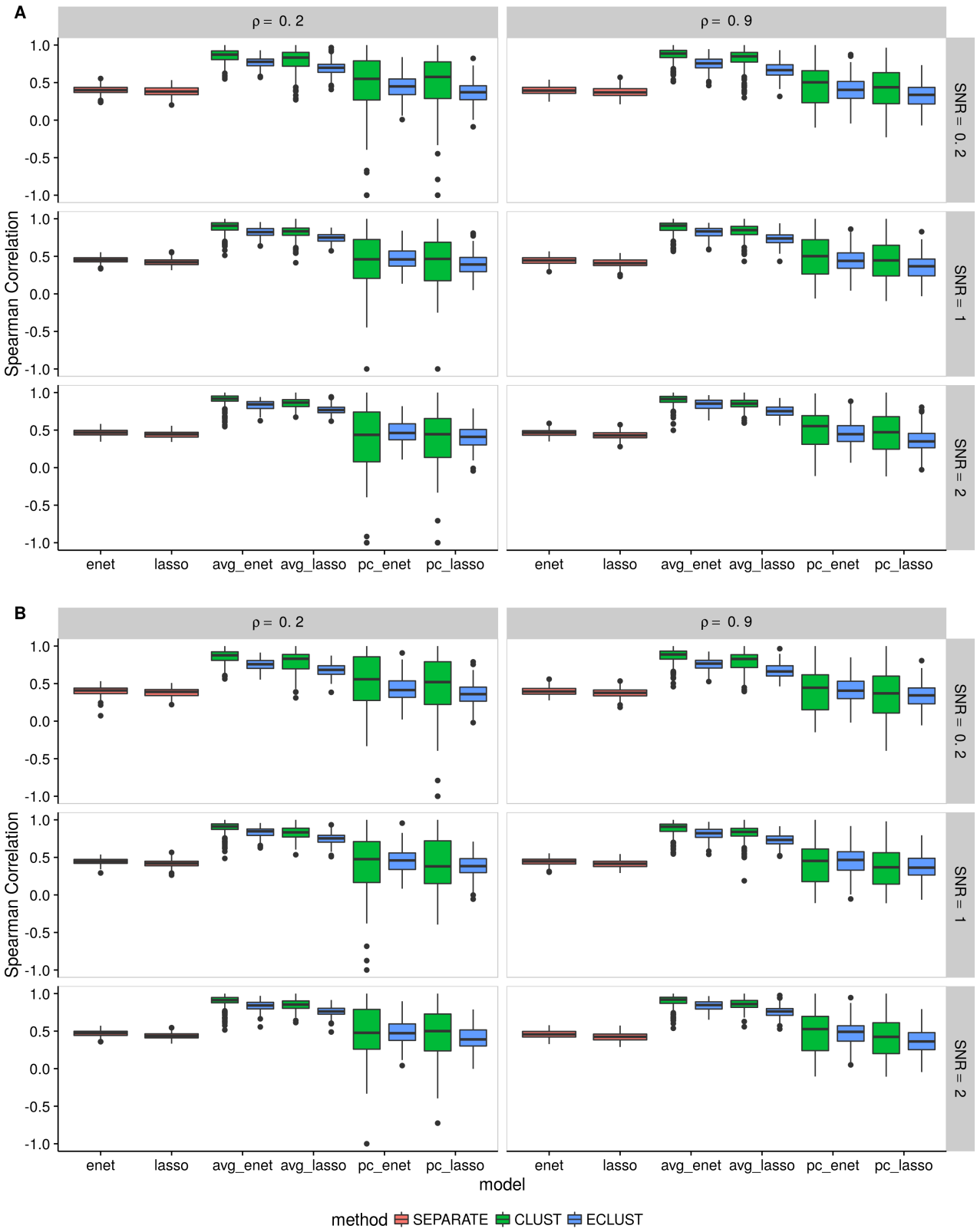


Figure S20: Simulation 2 – Average Spearman correlation from 10 CV folds of the training set using the TOM as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated 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.

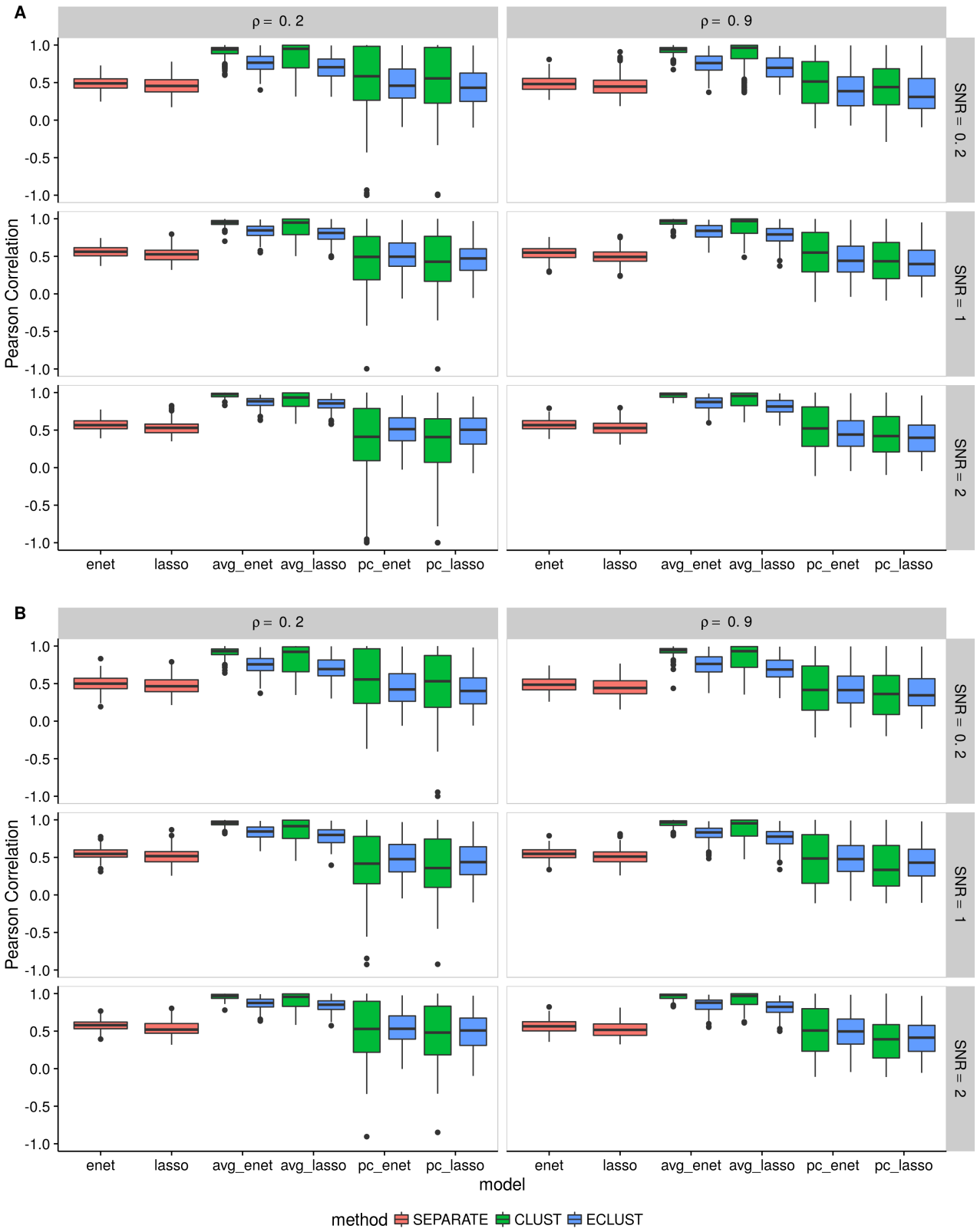


Figure S21: Simulation 2 – Average Pearson correlation from 10 CV folds of the training set using the TOM as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated 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

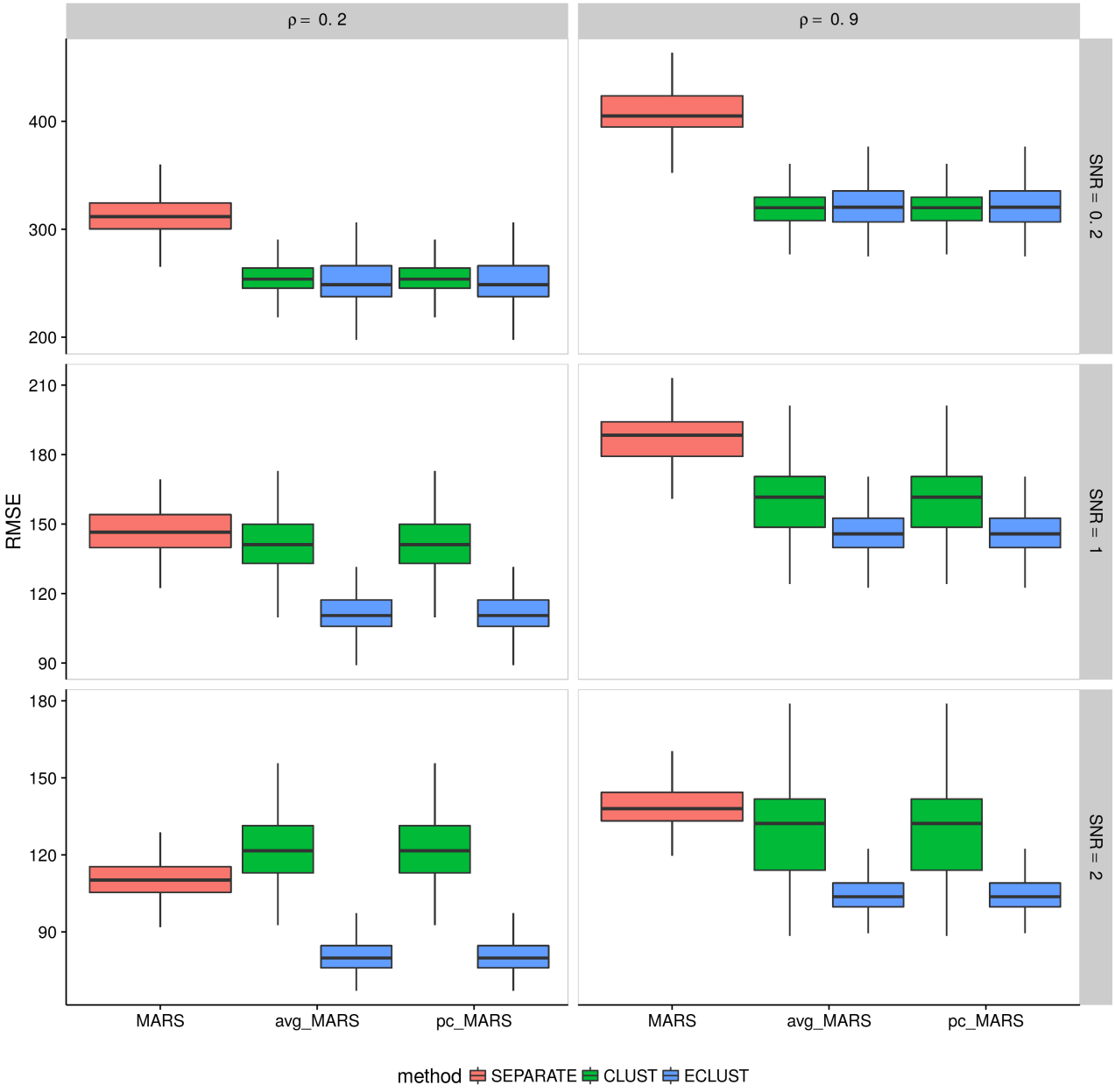


Figure S22: Simulation 3 – Root mean squared error on an independent test set using the TOM as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

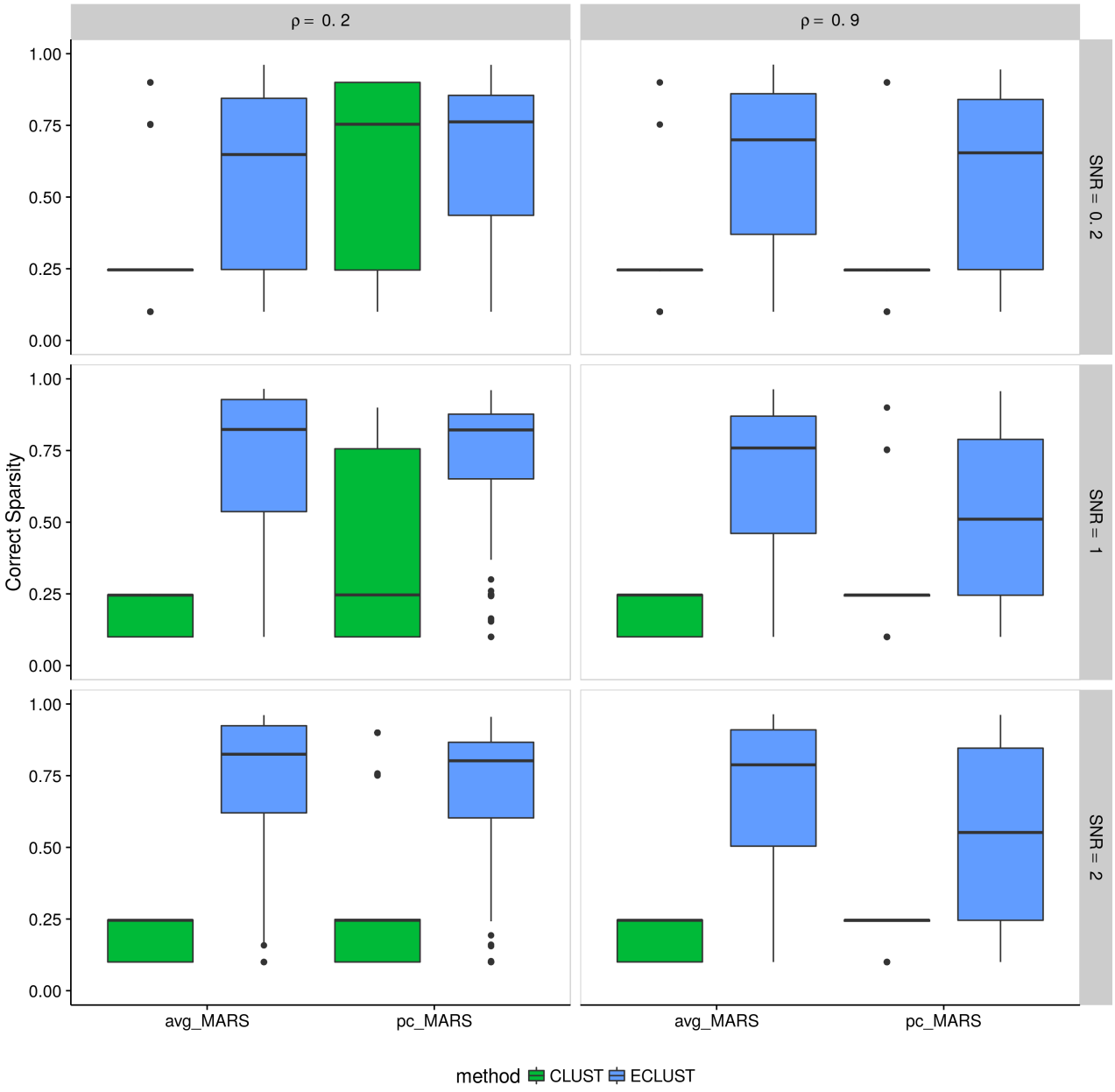


Figure S23: Simulation 3 – Correct Sparsity based on the training set using the TOM as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

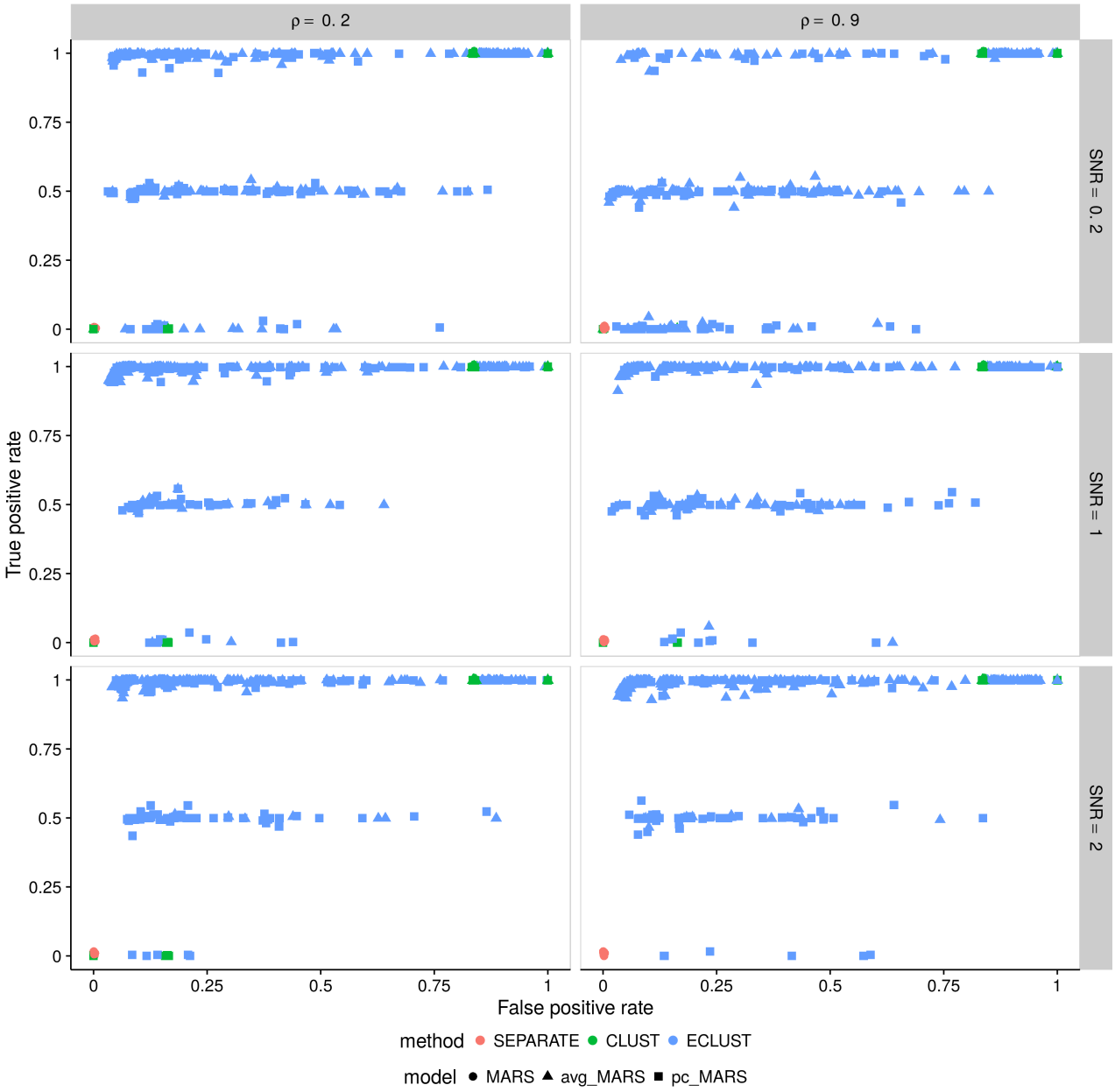


Figure S24: Simulation 3 – True positive rate vs. false positive rate based on the training set using the TOM as a measure of similarity. Each point represents 1 simulation run (there are a total of 200 simulation runs). Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

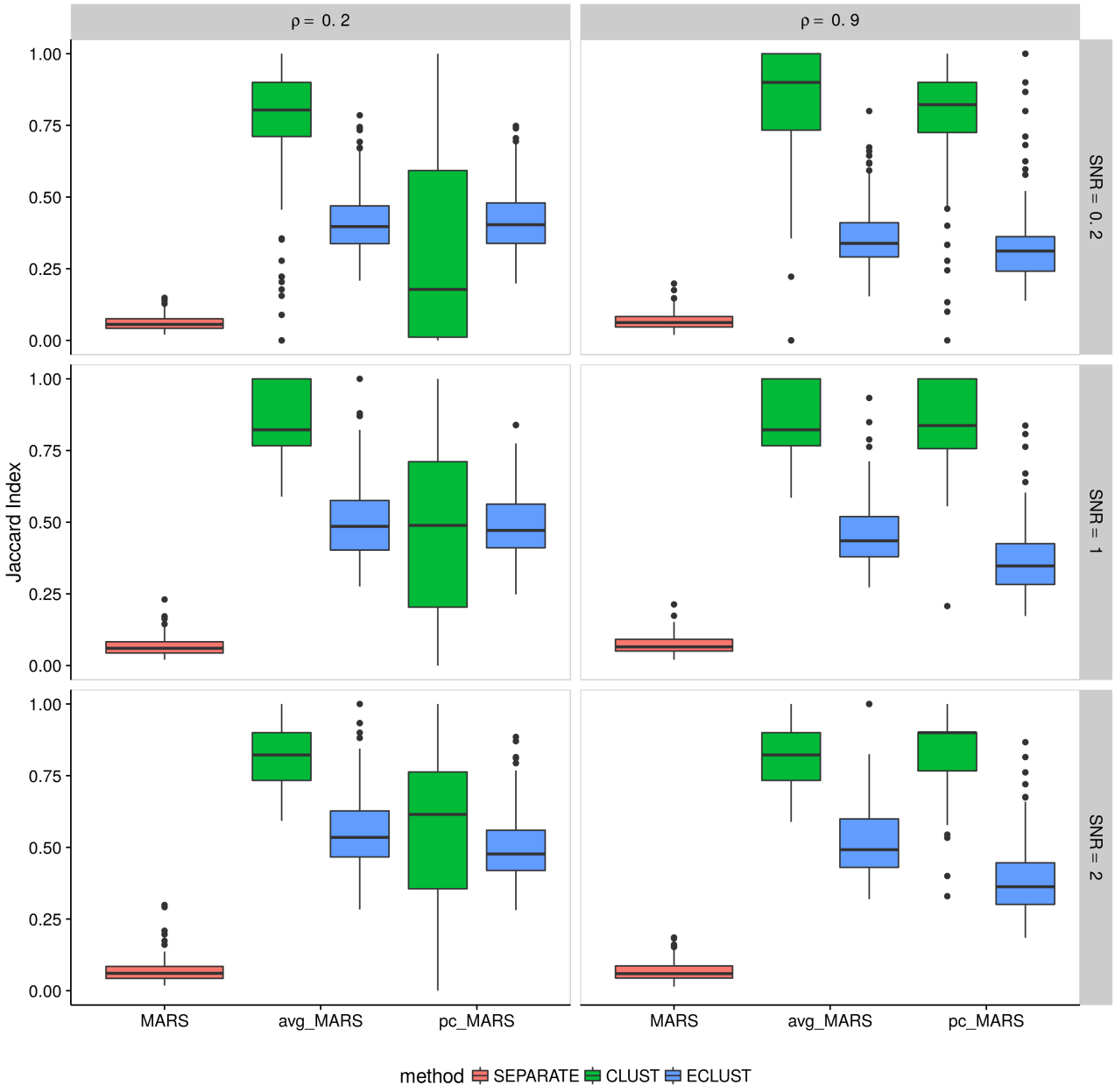


Figure S25: Simulation 3 – Average Jaccard Index from 10 CV folds of the training set using the TOM as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of selected predictors. We then calculate the Jaccard Index between all $\binom{10}{2}$ possible combinations 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

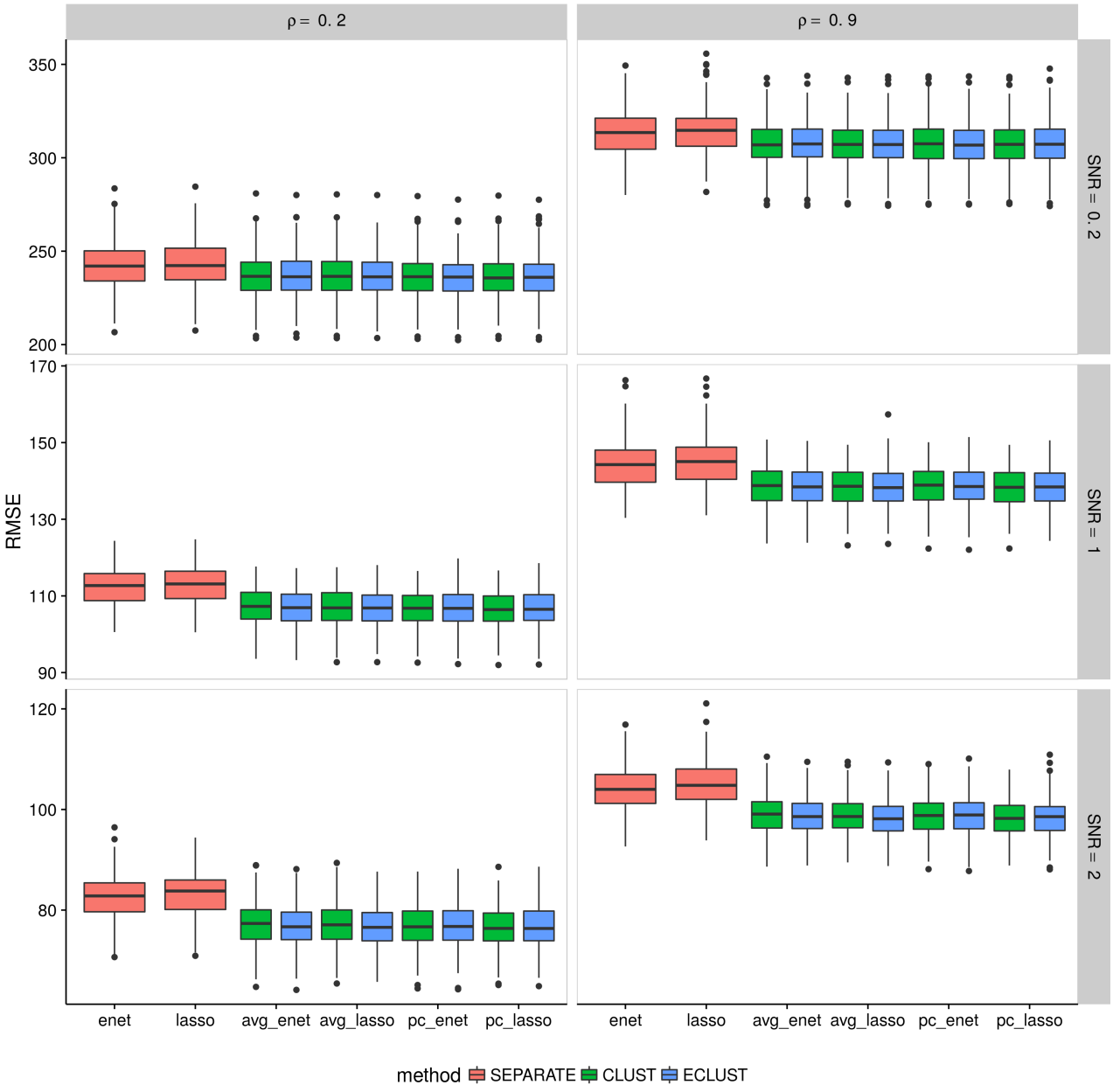


Figure S26: Simulation 1 – Root mean squared error on an independent test set using the Correlation as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

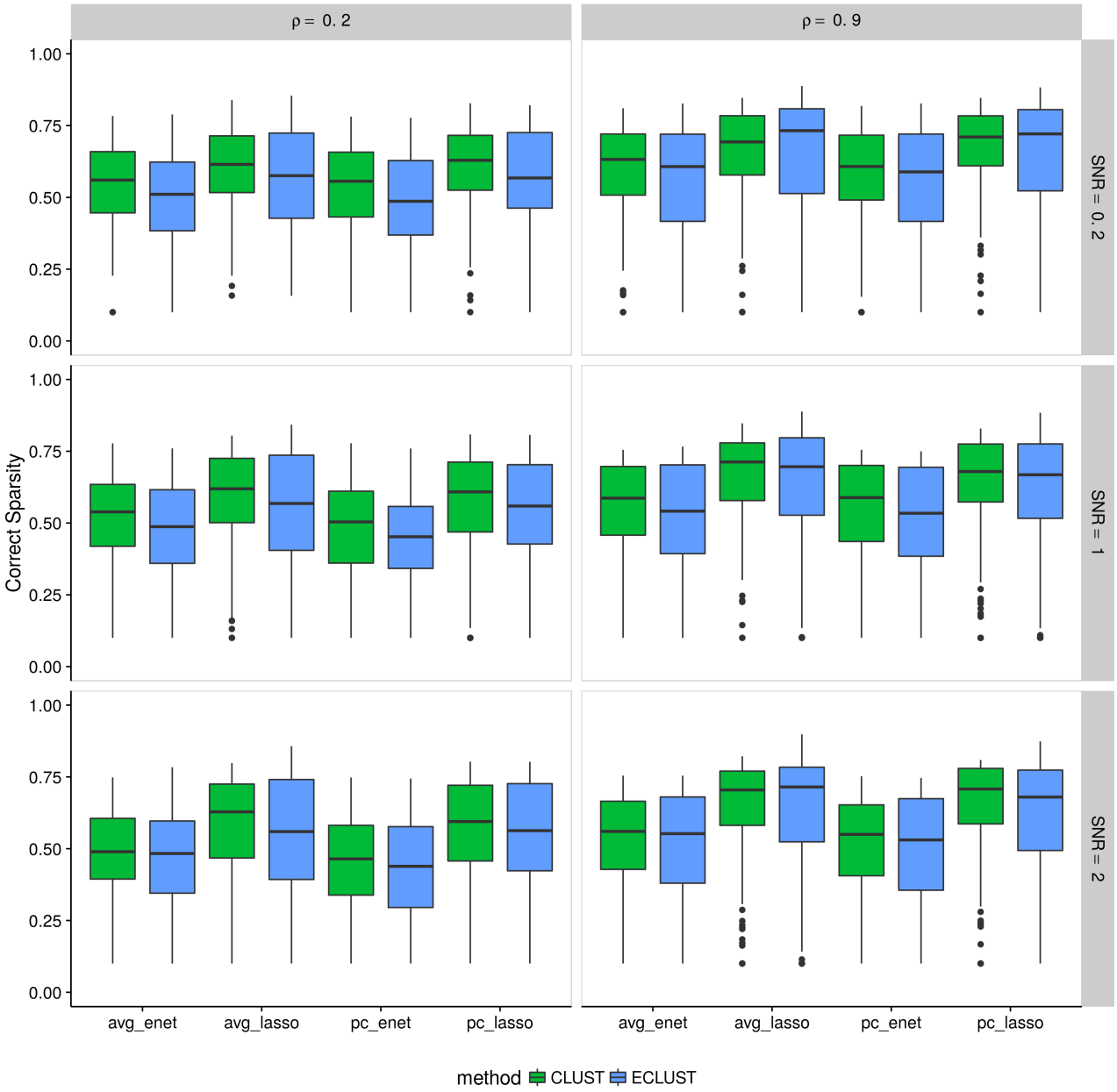


Figure S27: Simulation 1 – Correct Sparsity based on the training set using the Pearson correlation as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

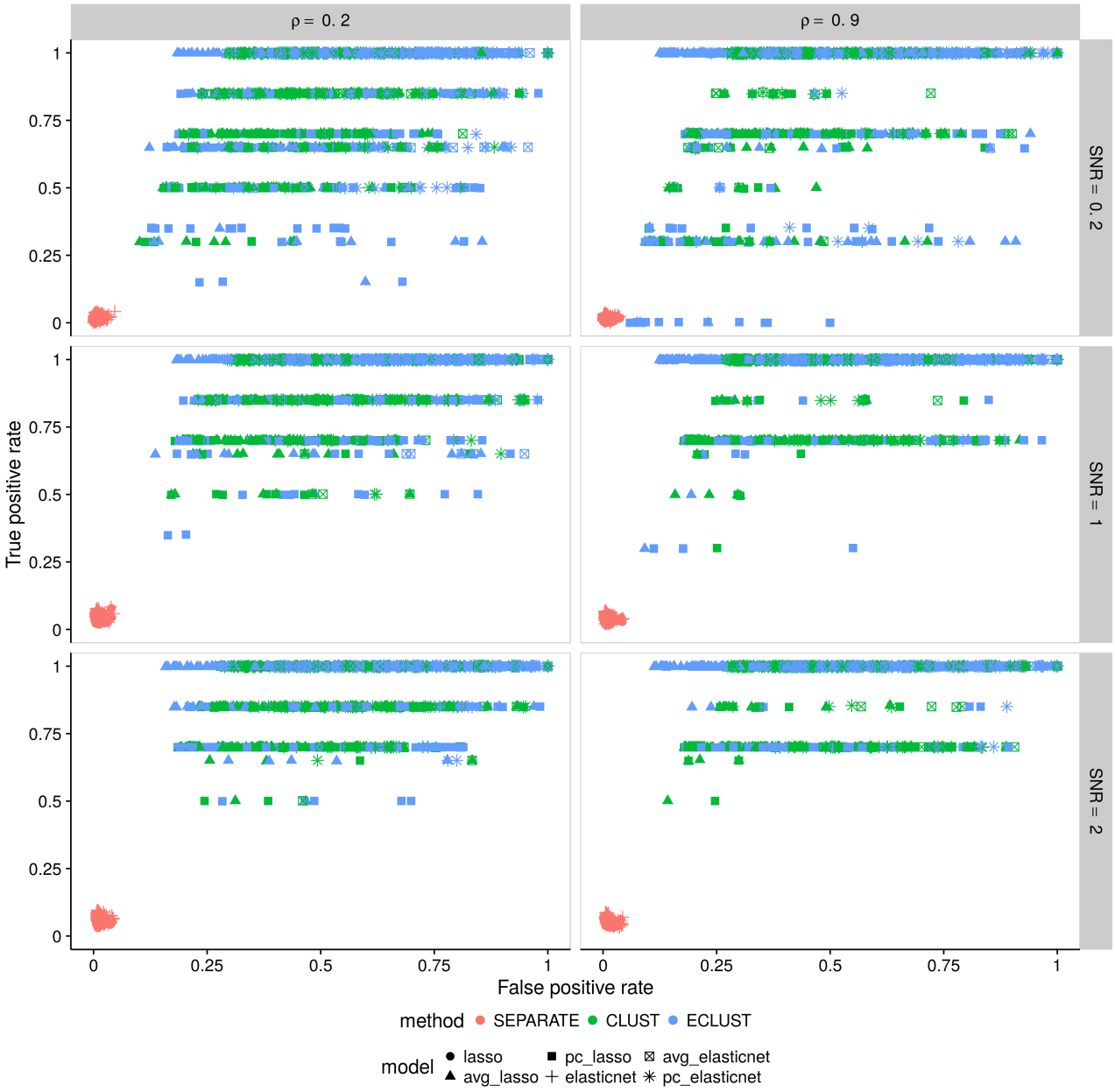


Figure S28: Simulation 1 – True positive rate vs. false positive rate based on the training set using the Pearson correlation as a measure of similarity. Each point represents 1 simulation run (there are a total of 200 simulation runs). Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

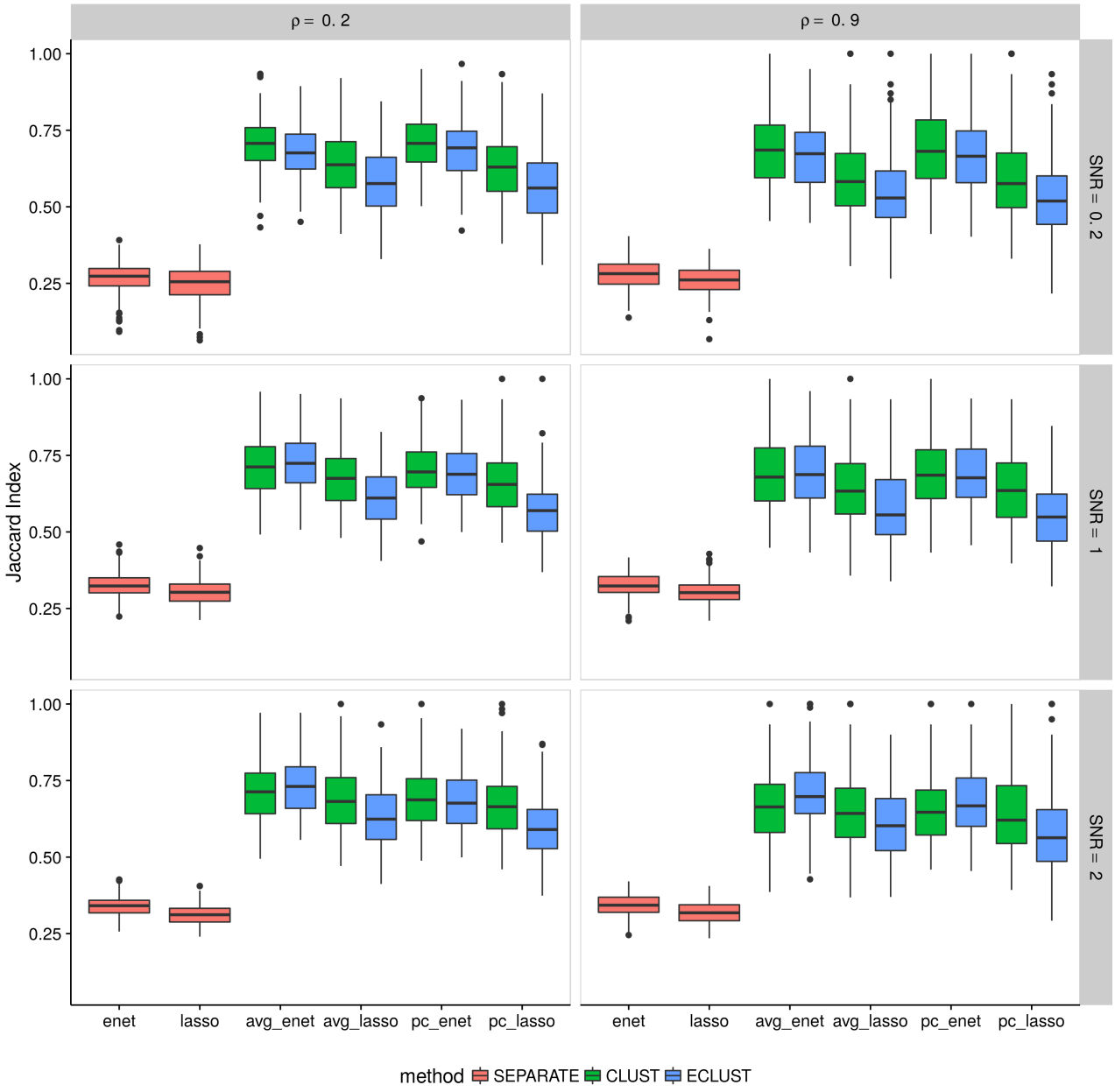


Figure S29: Simulation 1 – Average Jaccard Index from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of selected predictors. We then calculate the Jaccard Index between all $\binom{10}{2}$ possible 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.

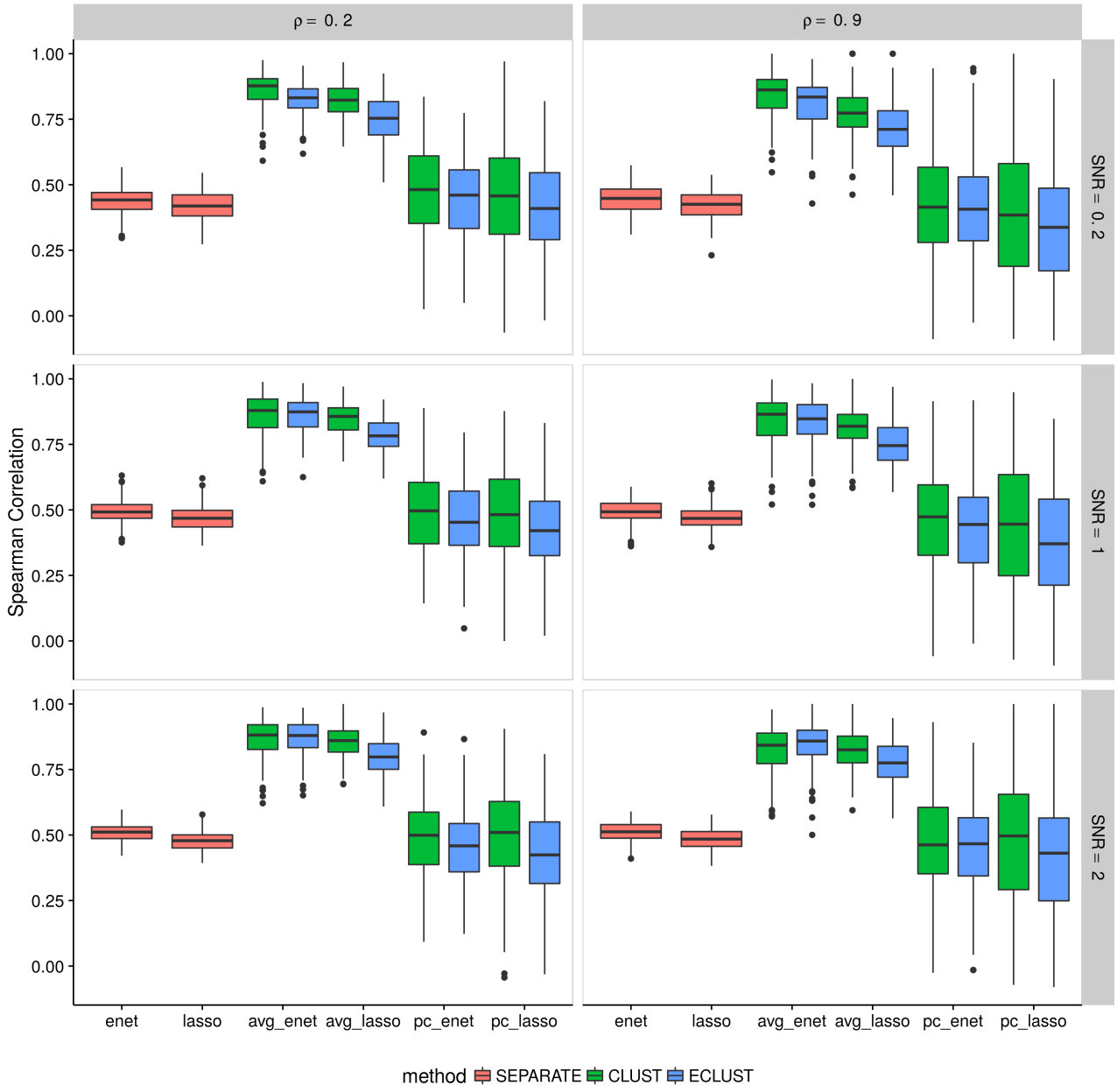


Figure S30: Simulation 1 – Average Spearman correlation from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated regression coefficients. We then calculate the Spearman correlation between 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.

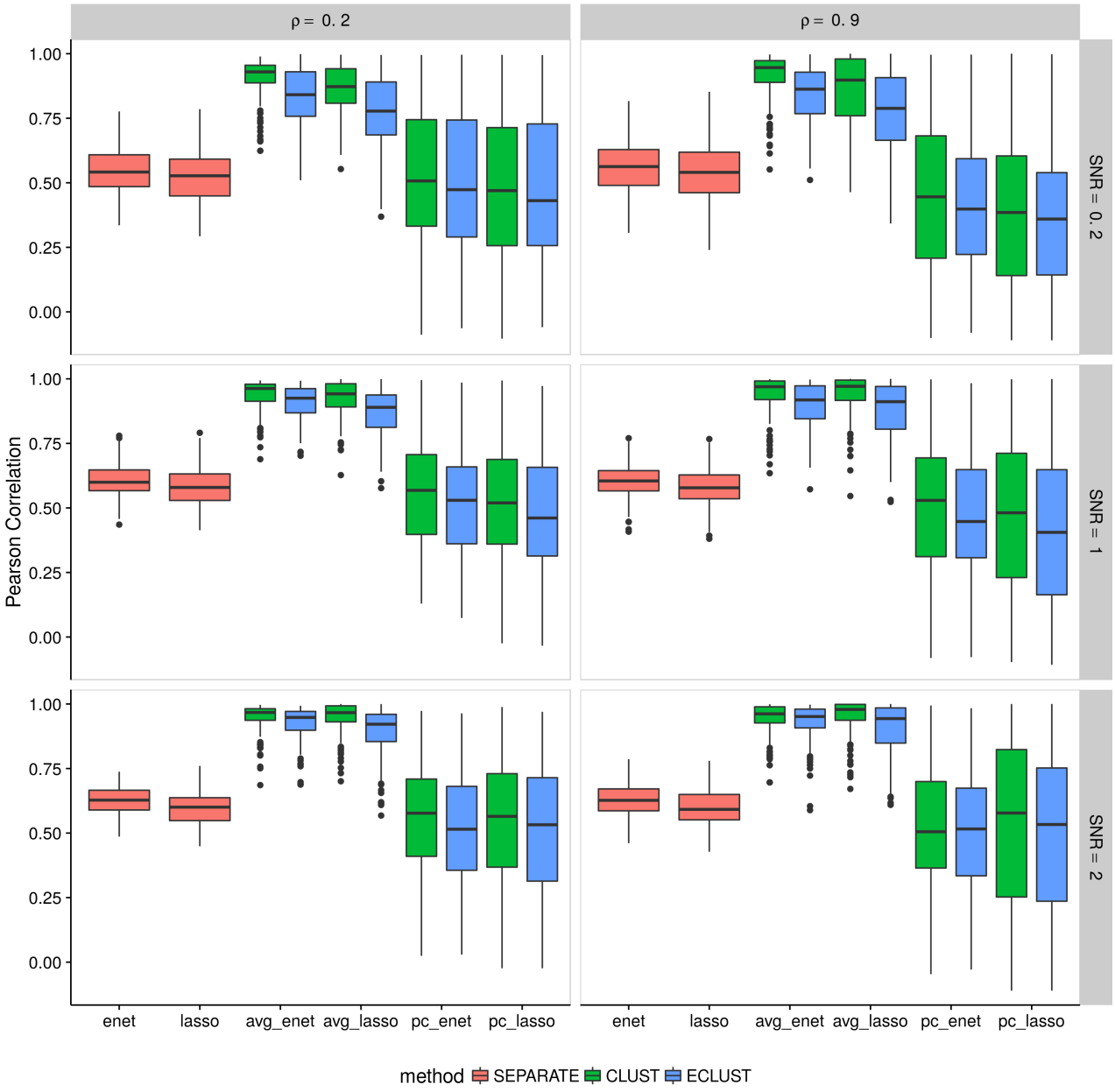


Figure S31: Simulation 1 – Average Pearson correlation from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of estimated regression coefficients. We then calculate the Pearson correlation between 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

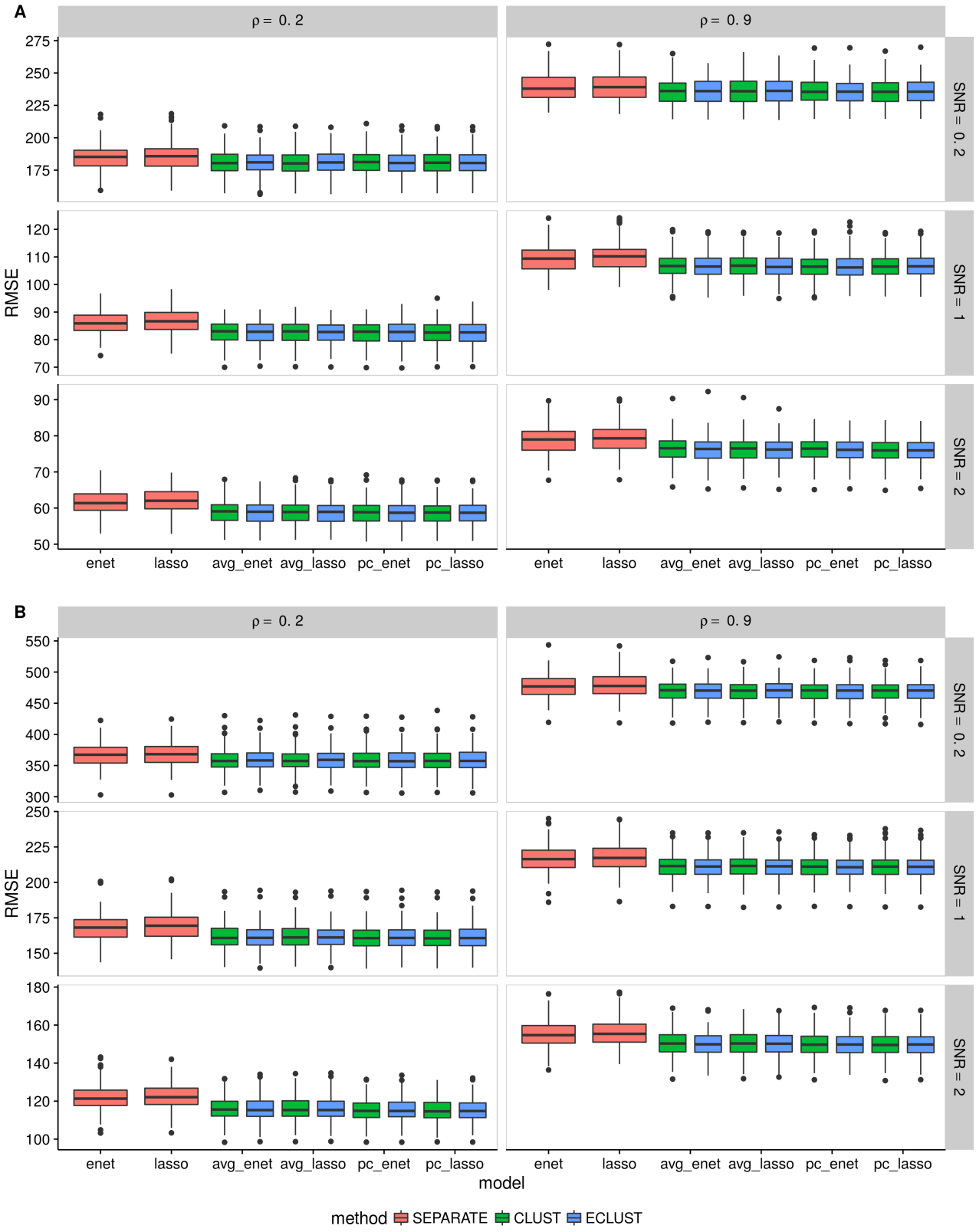


Figure S32: Simulation 2 – Root mean squared error on an independent test set using the Pearson correlation as a measure of similarity from 200 simulation runs. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

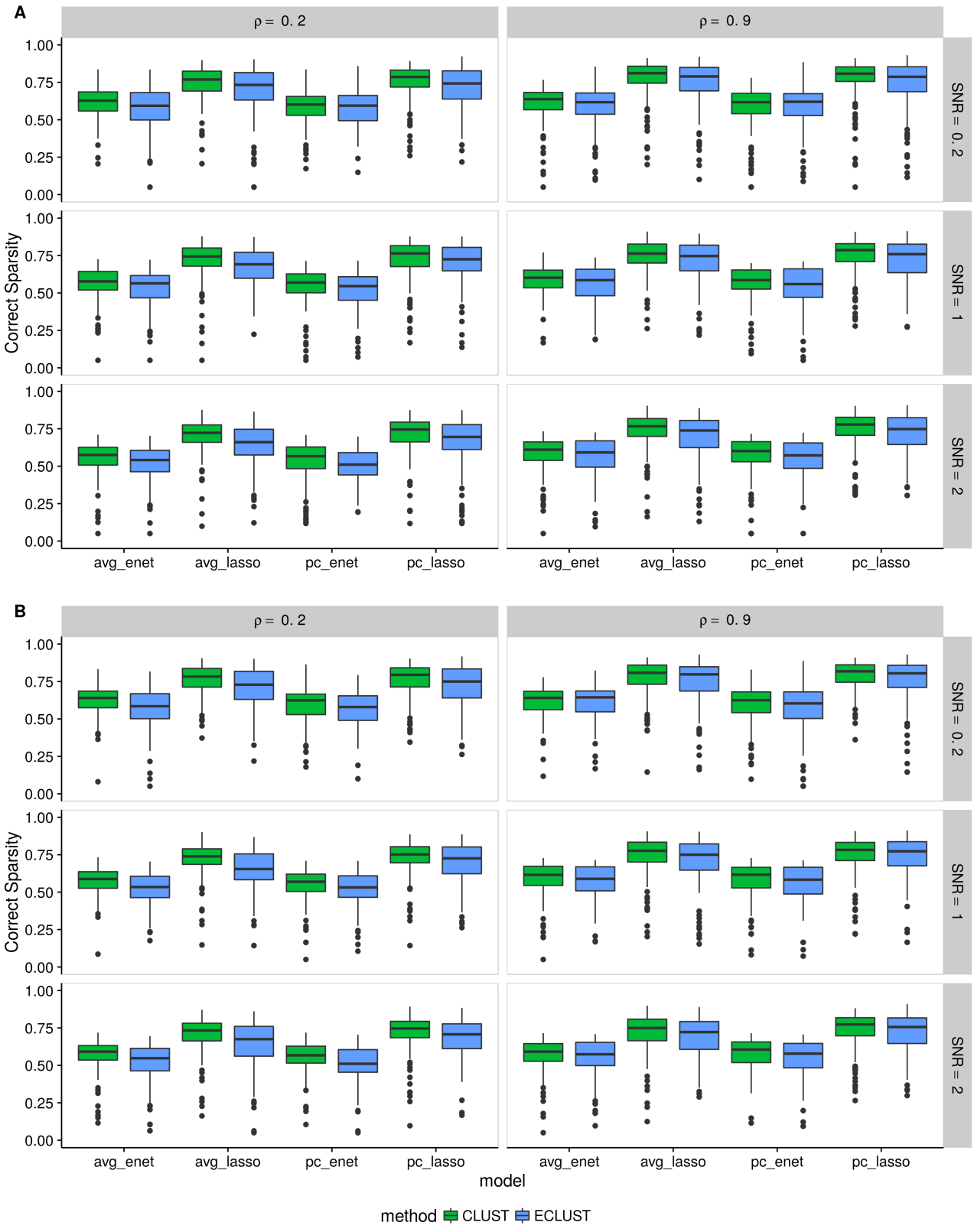


Figure S33: Simulation 2 – Correct Sparsity based on the training set using the Pearson correlation as a measure of similarity from 200 simulation runs. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

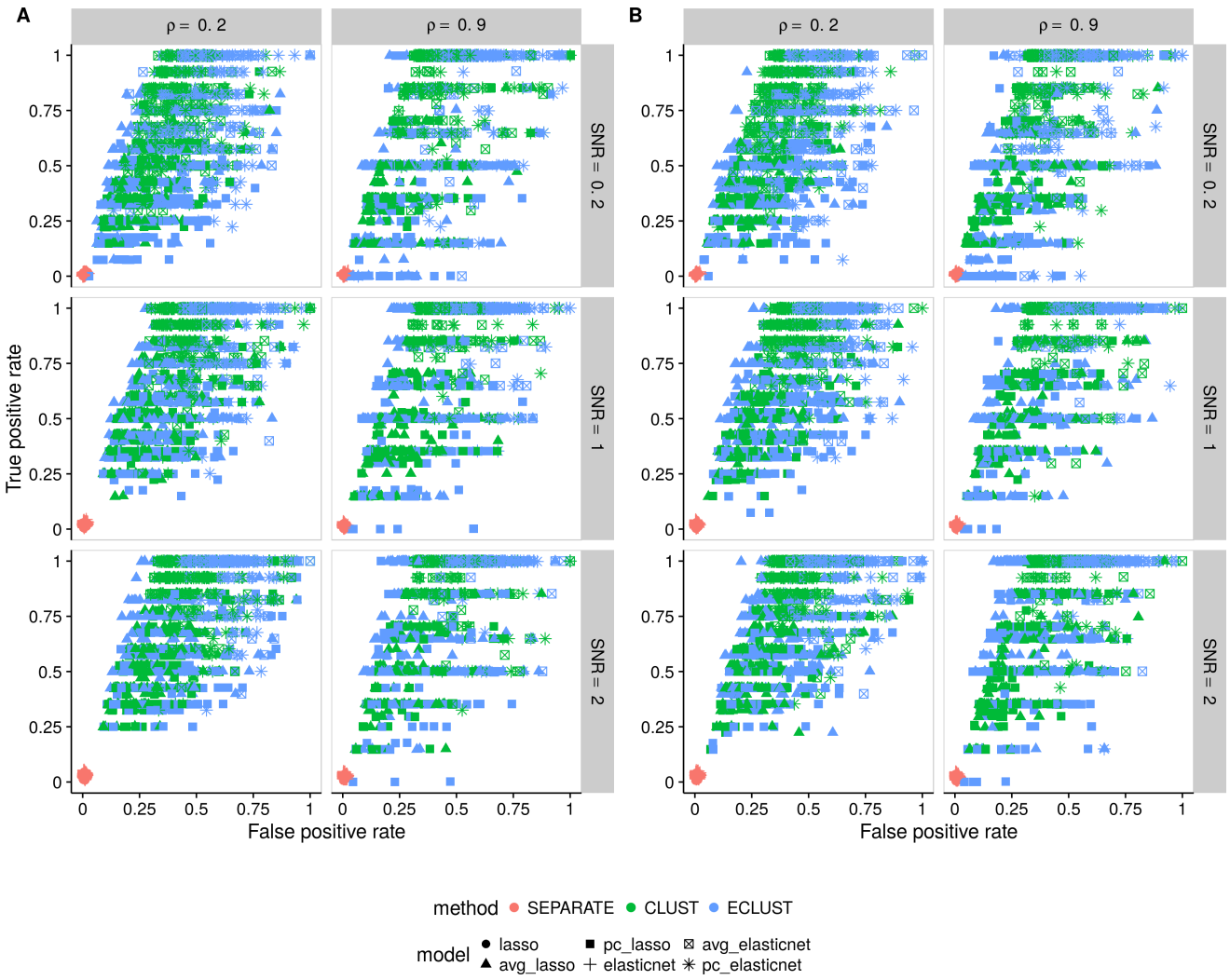


Figure S34: Simulation 2 – True positive rate vs. false positive rate based on the training set using the Pearson correlation as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. Each point represents 1 simulation run (there are a total of 200 simulation runs). Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

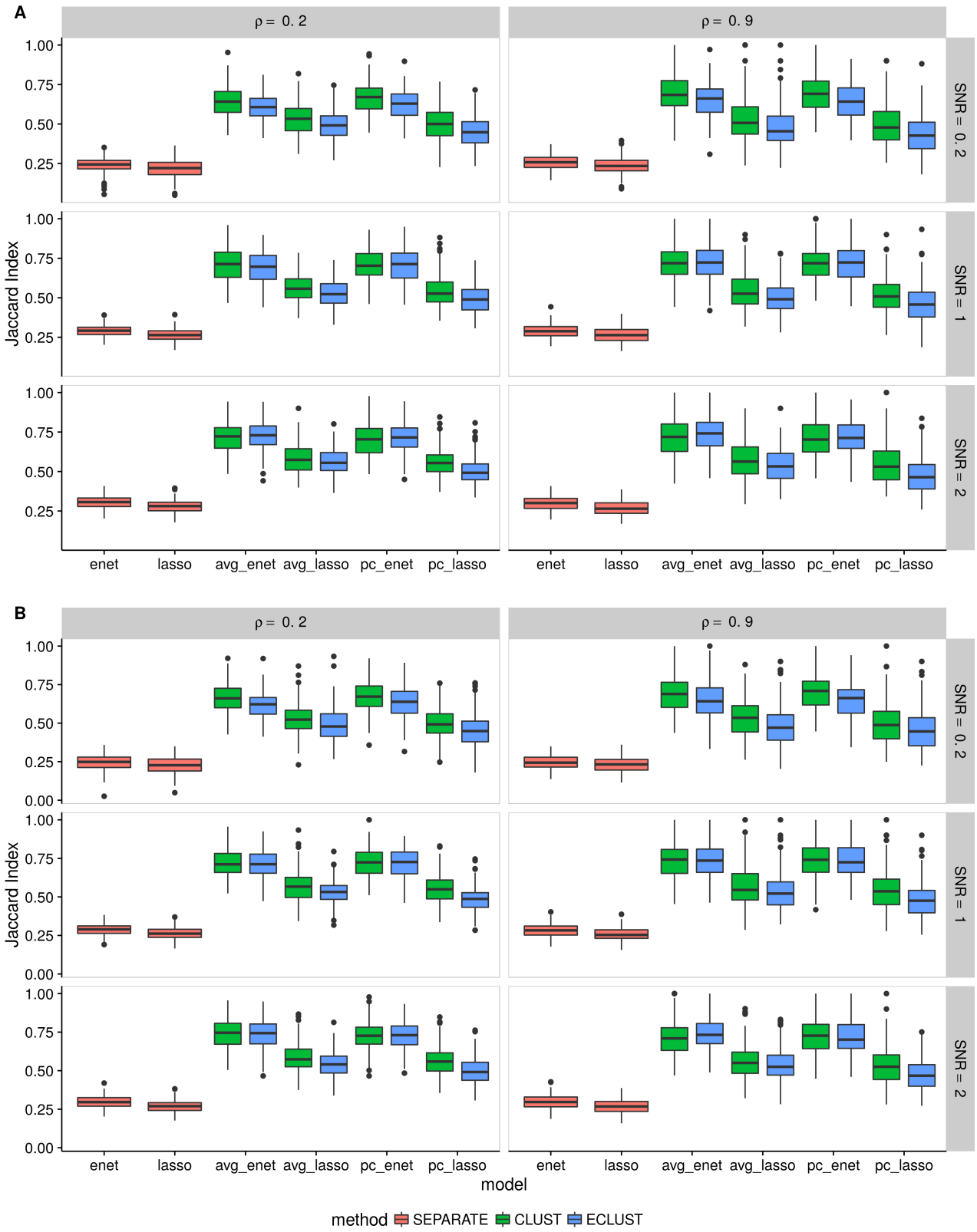


Figure S35: Simulation 2 – Average Jaccard Index from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. We fit the model to each of the 10 CV folds resulting in 10 sets of 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.

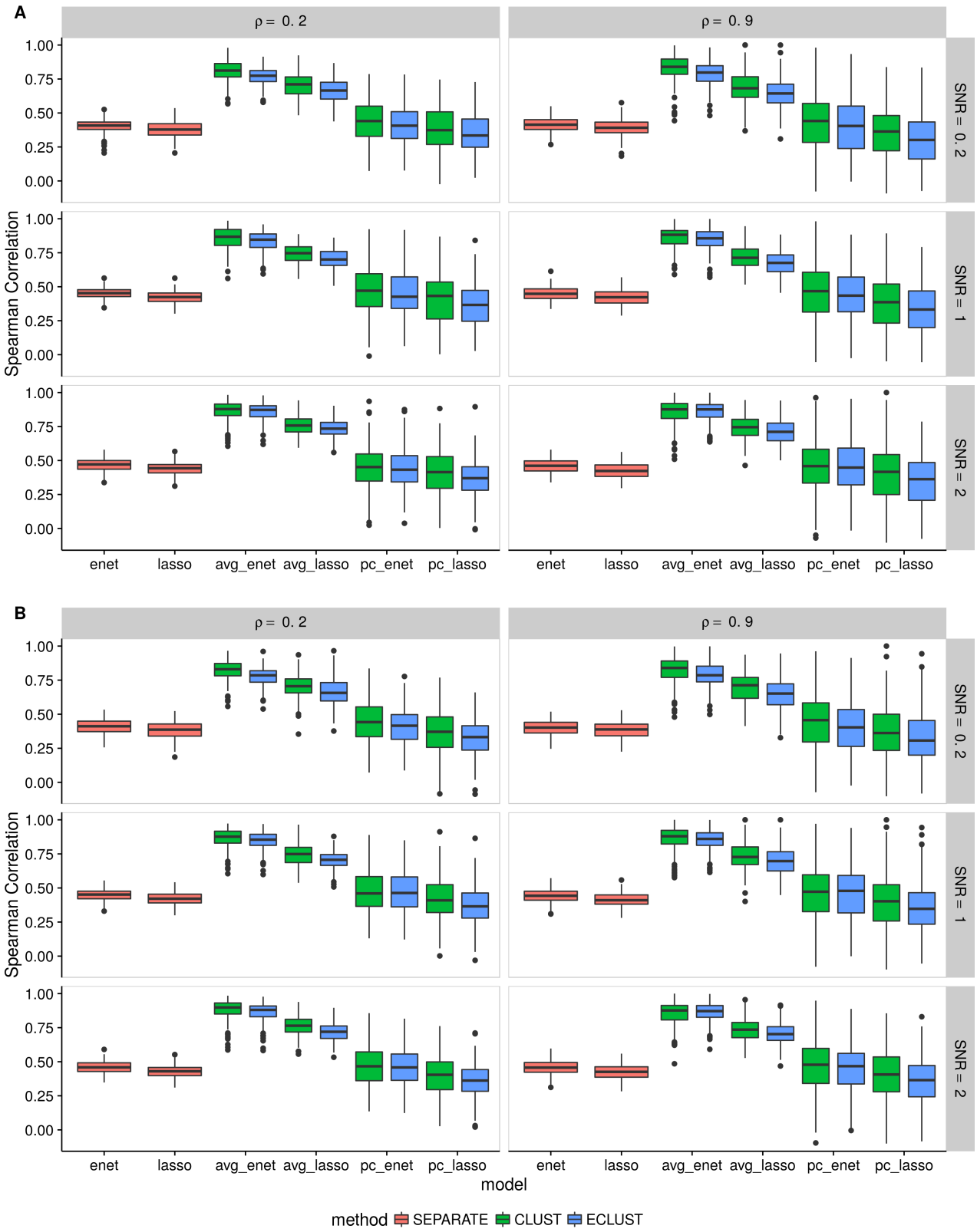


Figure S36: Simulation 2 – Average Spearman correlation from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. We fit the model to each of the 10 CV folds resulting in 10 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.

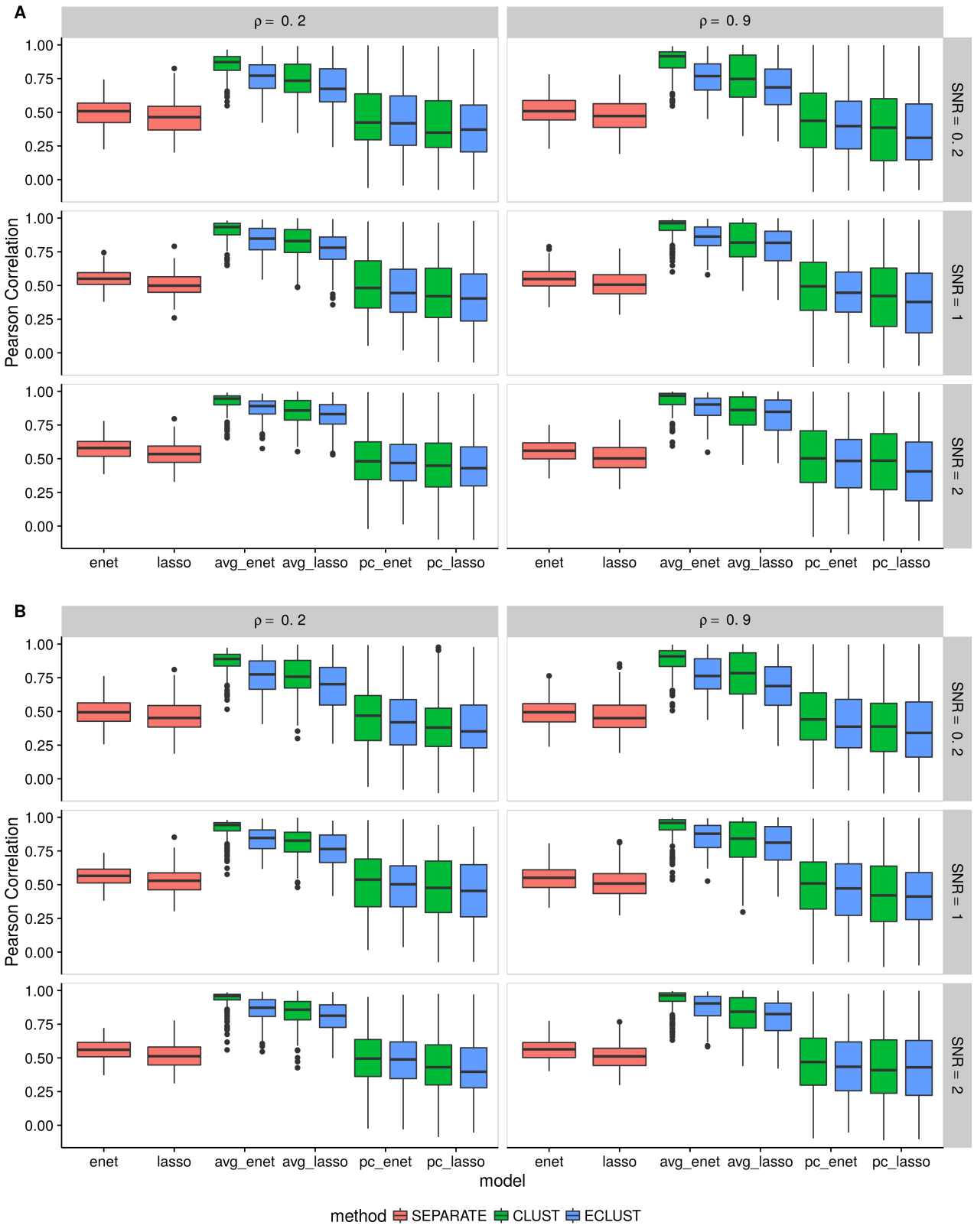


Figure S37: Simulation 2 – Average Pearson correlation from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. (A) $\alpha_j \sim \text{Unif}\,[0.4, 0.6]$, (B) $\alpha_j \sim \text{Unif}\,[1.9, 2.1]$. We fit the model to each of the 10 CV folds resulting in 10 sets 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

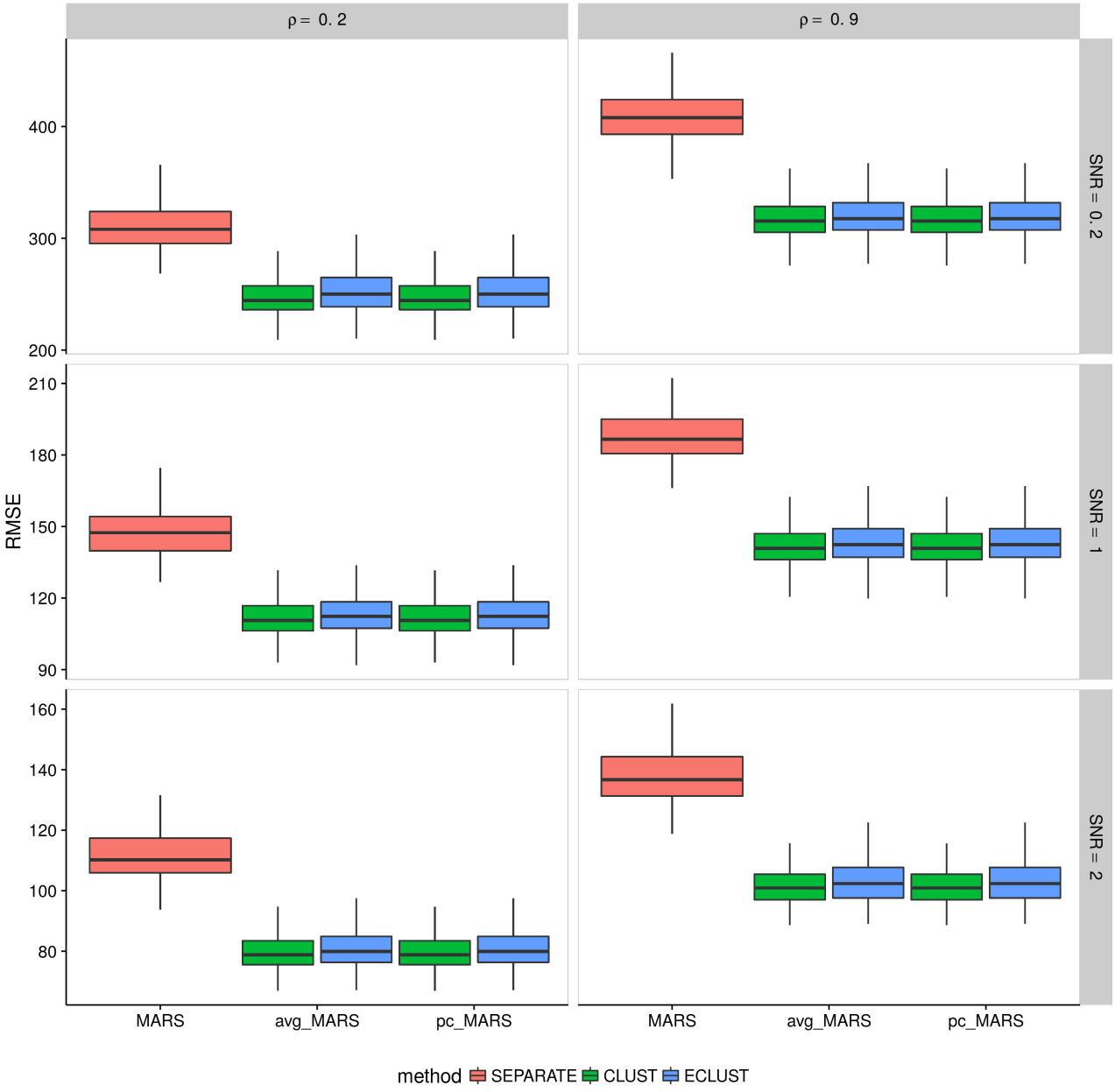


Figure S38: Simulation 3 – Root mean squared error on an independent test set using the Pearson correlation as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

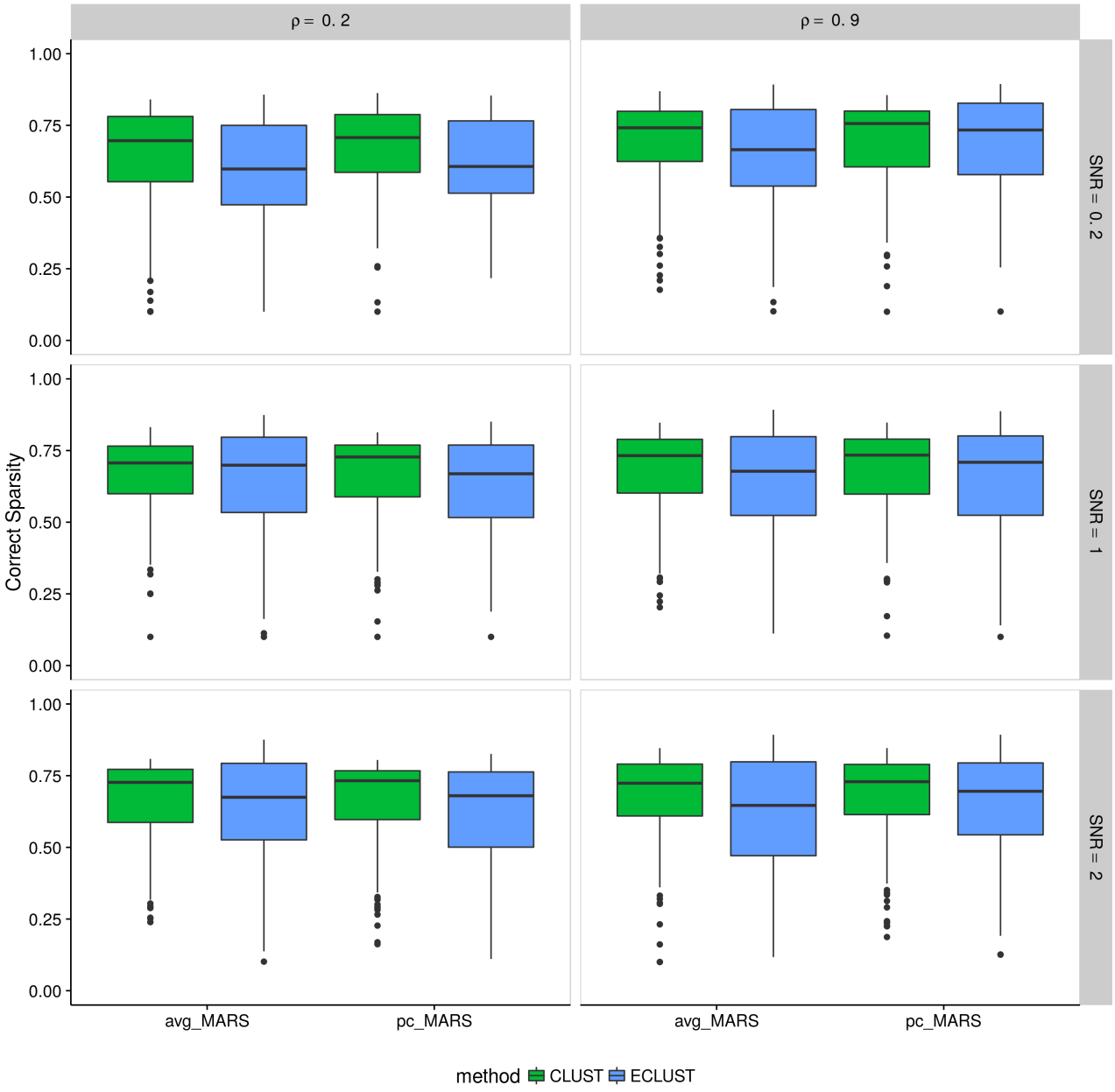


Figure S39: Simulation 3 – Correct Sparsity based on the training set using the Pearson correlation as a measure of similarity from 200 simulation runs. Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

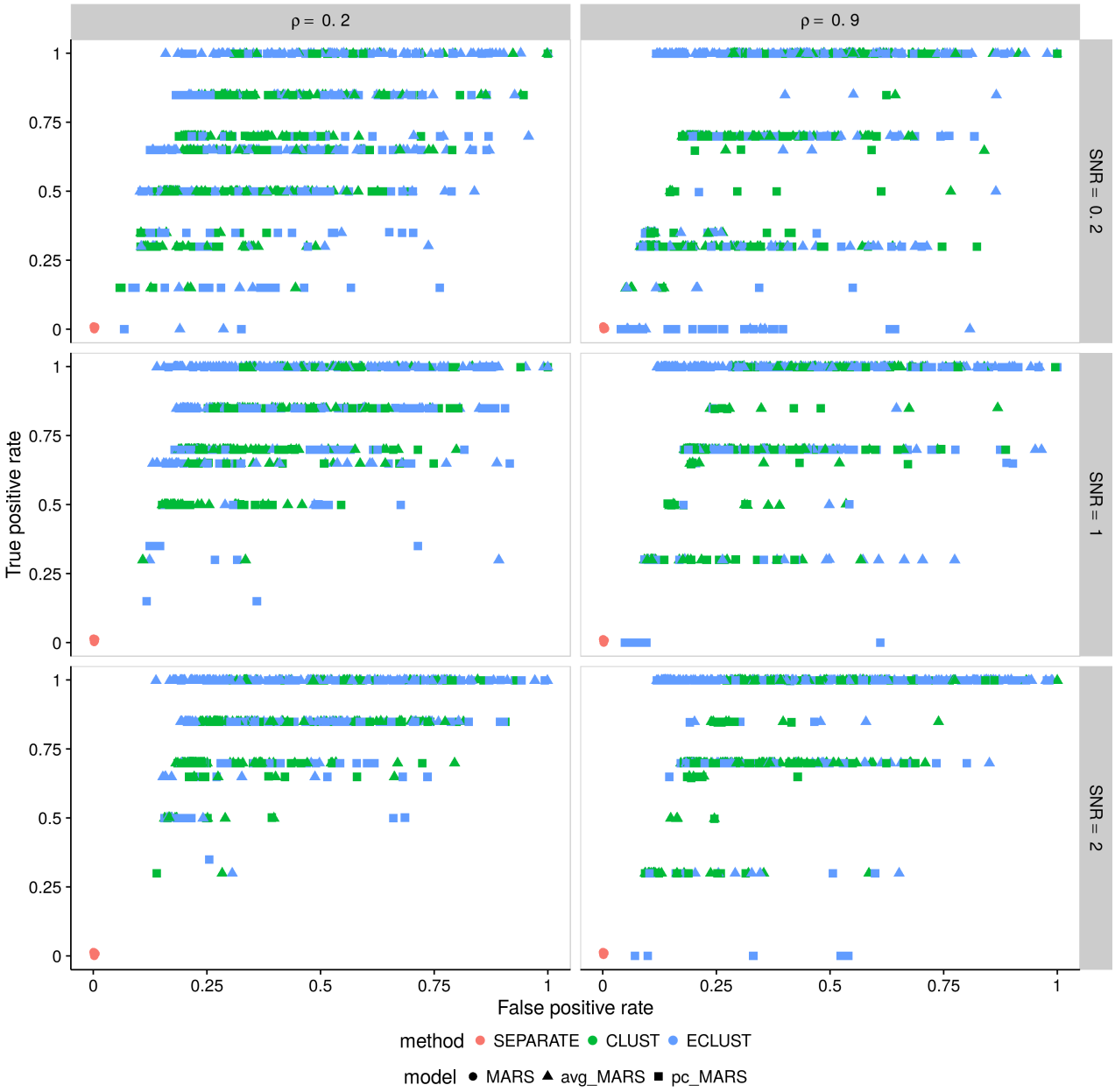


Figure S40: Simulation 3 – True positive rate vs. false positive rate based on the training set using the Pearson correlation as a measure of similarity. Each point represents 1 simulation run (there are a total of 200 simulation runs). Vertical panels represent varying correlation between active clusters. Horizontal panels represent different signal-to-noise ratios.

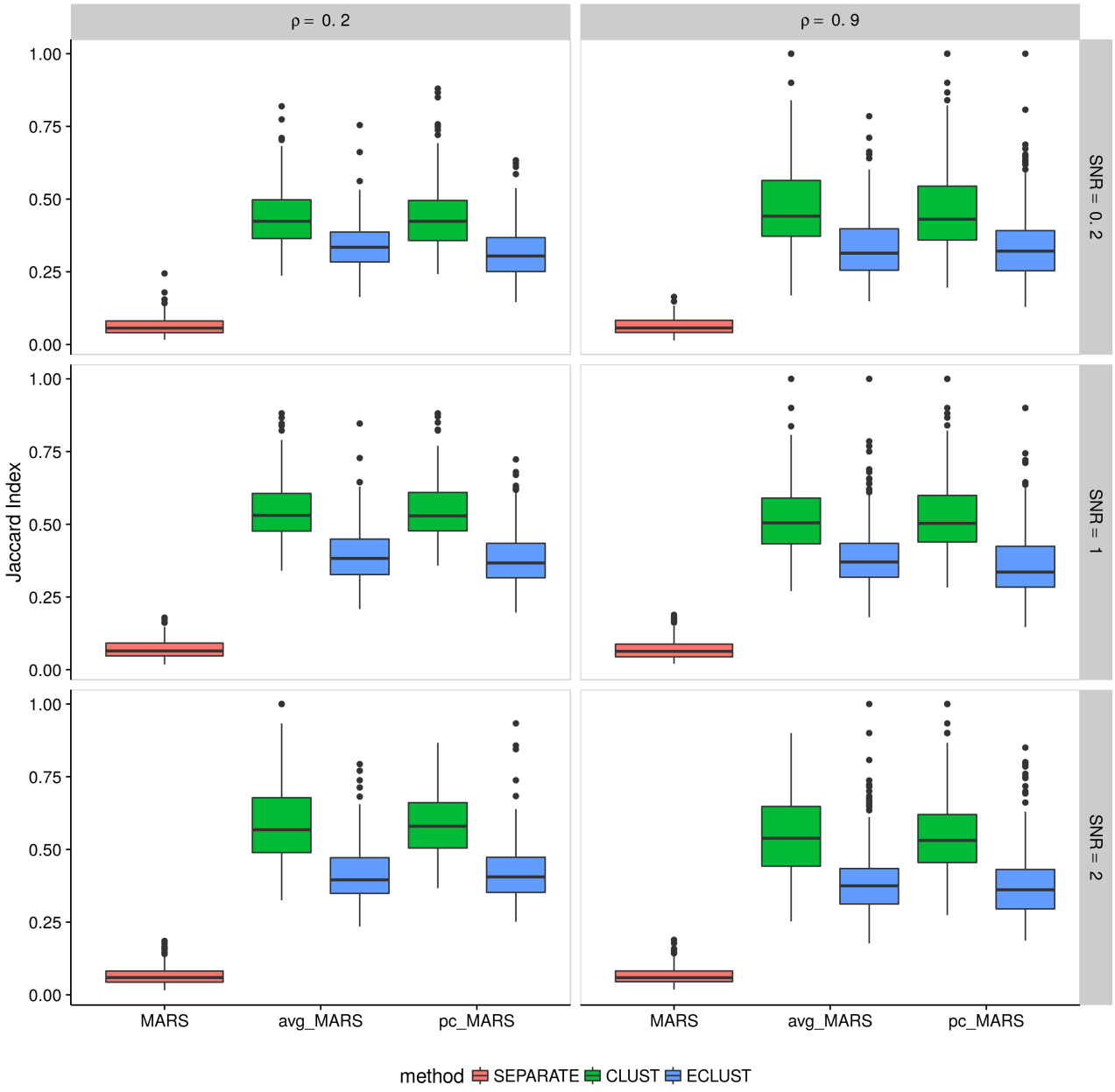
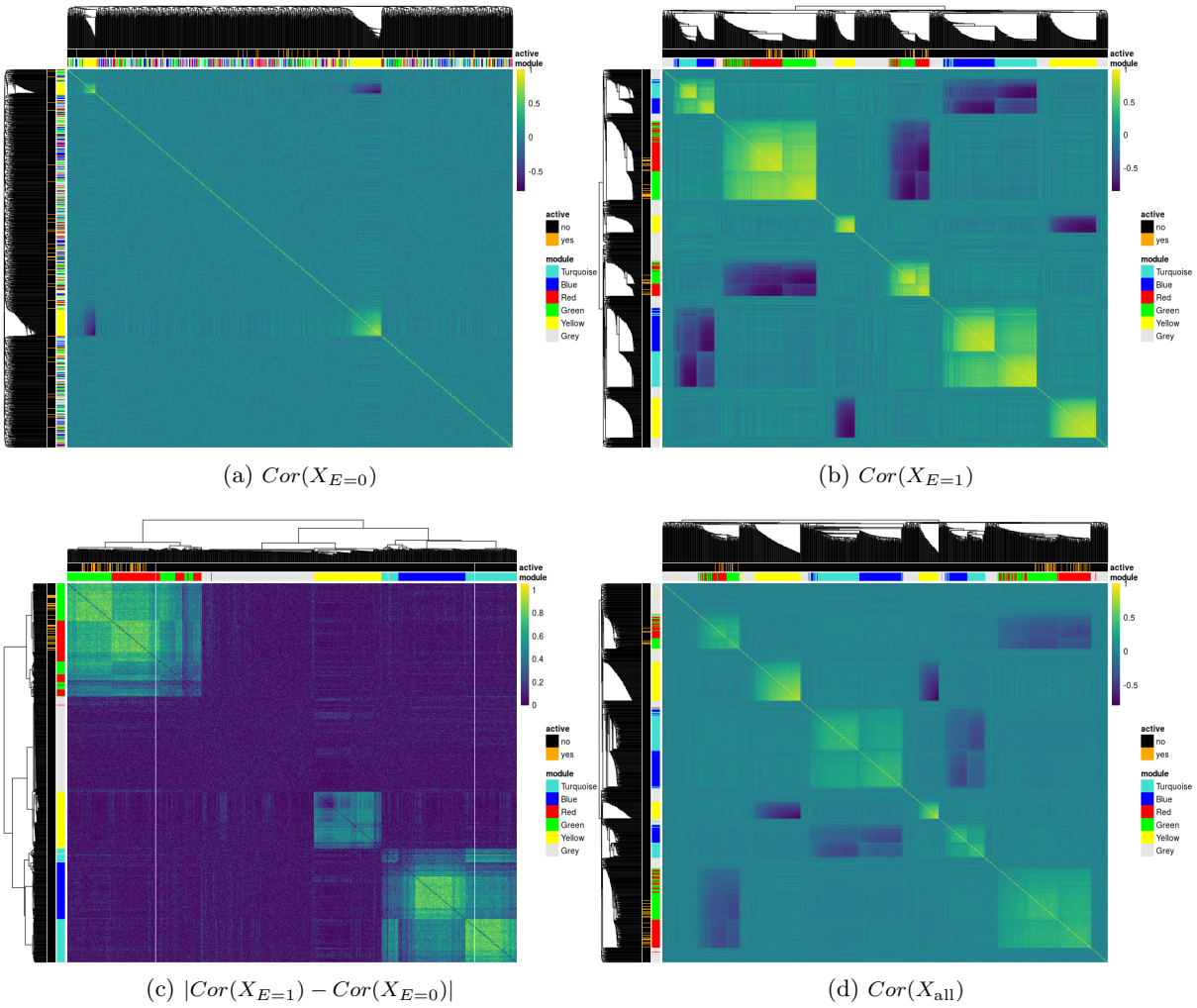


Figure S41: Simulation 3 – Average Jaccard Index from 10 CV folds of the training set using the Pearson correlation as a measure of similarity. We fit the model to each of the 10 CV folds resulting in 10 sets of selected predictors. We then calculate the Jaccard Index between all $\binom{10}{2}$ possible 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.

F Visual Representation of Similarity Matrices

F.1 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 S42: 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.

References

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555.

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