A Model for Interpretable High Dimensional Interactions

Bhatnagar SR[1,2], Yang Y[3], Blanchette M[4], Greenwood CMT[1,2]

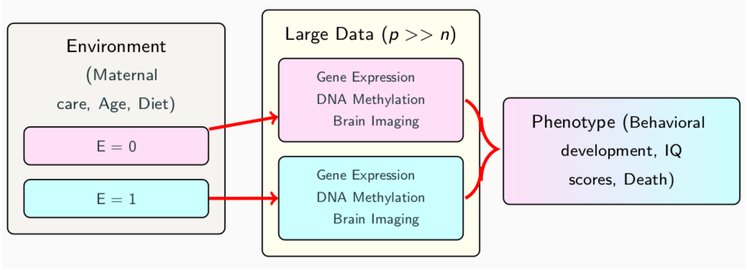
[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University
[2]Lady Davis Institute, Jewish General Hospital, Montreal, QC
[3]Department of Mathematics and Statistics, McGill University, QC
[4]School of Computer Science, McGill University

Summary

- Environmental exposures may induce subtle system-wide changes in high-dimensional genomic data such as gene expression or epigenetic measures
  - Can such situations be exploited to improve prediction models?
- Large system-wide changes are observed in many environments and hence this assumption can possibly be exploited to aid analysis of high dimensional data
- We develop and implement a multivariate penalization procedure for predicting a continuous or binary disease outcome while detecting interactions between high dimensional data ($p >> n$) and an environmental factor. R software: http://sahirbhatnagar.com/eclust/
  - Dimension reduction is achieved through leveraging the environmental-class-conditional correlations
  - Also, we develop and implement a strong heredity framework within the penalized model

Motivation

FIGURE 1: Microarray study of COPD. Top: Heatmap of Pearson correlations. Bottom: Heatmap of gene expression data (2,900 genes) rows are genes and columns are subjects. There are 7 subjects in each group, matched on COPD case status, gender and age.



Methods

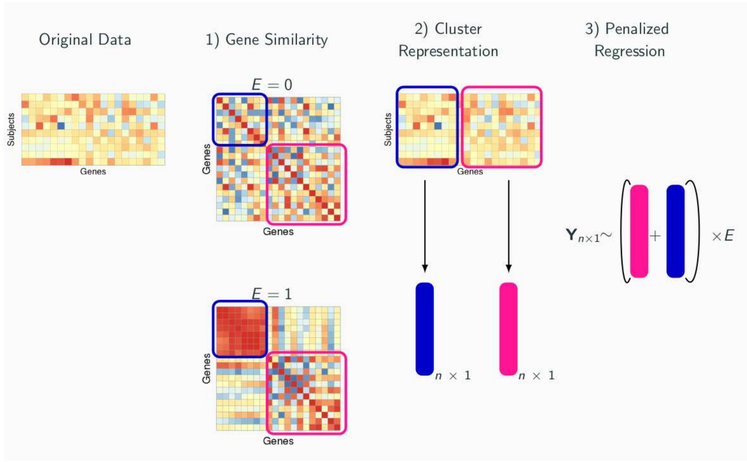


FIGURE 3: Method overview. First step involves measuring gene similarity in both exposure groups. We then cluster these and create a cluster representation. The last step involves entering these terms in a penalization model that follows the strong heredity principle [1]

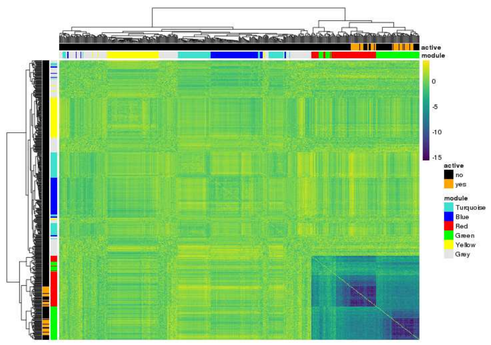


FIGURE 4: Clustering based on Fisher's Z transformation of exposure dependent correlations. Let ρ_{ijk} be the correlation between genes i and j in class k. Transform the correlations into z values: $z_{ijk} = 0.5log[(1+\rho_{ijk})/(1-\rho_{ijk})]$. The Z-test statistic is given by $|z_{ij0} - z_{ij1}|/\sqrt{1/(n_0-3)+1/(n1-3)} \sim \mathcal{N}(0,1)$

- Model: $g(\boldsymbol{\mu}) = \beta_0 + \underbrace{\beta_1 X_1 + \cdots + \beta_p X_p + \beta_E E}_{\text{main effects}} + \underbrace{\alpha_{1E}(X_1 E) + \cdots + \alpha_{pE}(X_p E)}_{\text{interactions}}$
- Strong Hierarchy Principle [1]: $\hat{\alpha}_{jE} \neq 0 \quad \Rightarrow \quad \hat{\beta}_j \neq 0 \quad \text{and} \quad \hat{\beta}_E \neq 0$
- Reparametrization [2]: $\alpha_{jE} = \gamma_{jE}\beta_j\beta_E$.
- Variable Selection:
  $\arg\min_{\beta_0,\boldsymbol{\beta},\boldsymbol{\gamma}} \frac{1}{2}\|Y - g(\boldsymbol{\mu})\|^2 + \lambda_\beta\left(w_1\beta_1 + \cdots + w_q\beta_q + w_E\beta_E\right) + \lambda_\gamma\left(w_{1E}\gamma_{1E} + \cdots + w_{qE}\gamma_{qE}\right)$
- Adaptive weights: $w_j = \left|\frac{1}{\hat{\beta}_j}\right|, \quad w_{jE} = \left|\frac{\hat{\beta}_j\hat{\beta}_E}{\hat{\alpha}_{jE}}\right|$
- Why strong heredity?
  - Statistical Power: large main effects are more likely to lead to detectable interactions than small ones
  - Interpretability: Assuming a model with interaction only is generally not biologically plausible
  - Practical Sparsity: $X_1, E, X_1 \cdot E$ (2 variables to measure) vs. $X_1, E, X_2 \cdot E$ (3 variables to measure).

Simulation Study Results

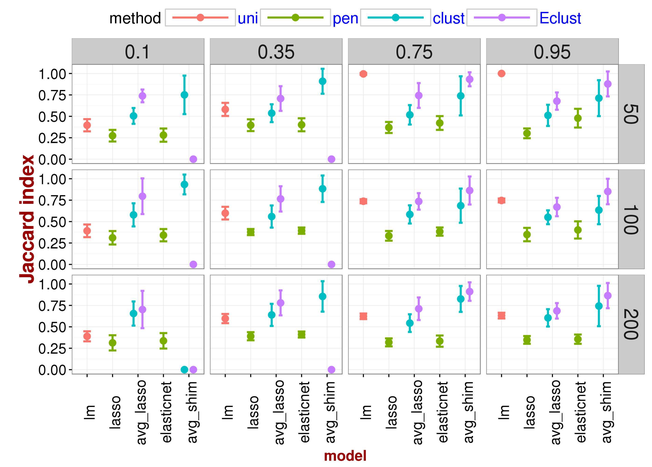


FIGURE 5: Stability of results: Average Jaccard distance from 10-fold cross validation. A Jaccard distance of 1 indicates perfect agreement between two sets while no agreement will result in a distance of 0.

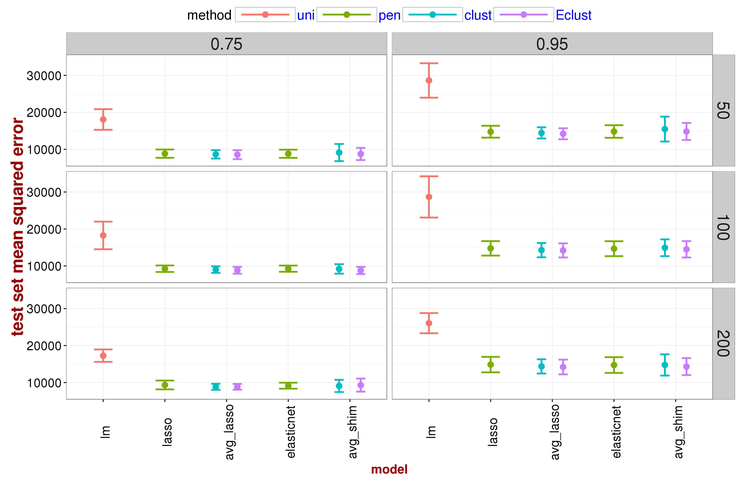


FIGURE 6: Prediction accuracy: Test set mean squared error

This work was supported by the Ludmer Centre for Neuroinformatics and Mental Health.
Software available at http://sahirbhatnagar.com/eclust/.

References

[1] Hugh Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36, 1996.
[2] Nam Hee Choi, William Li, and Ji Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364, 2010.