Predicting phenotypes from microarrays using amplified, initially marginal, eigenvector regression (AIMER)

Daniel J. McDonald
Indiana University, Bloomington
mypage.iu.edu/∼dajmcdon

30 July 2017

The paper:

R package:

My www:

Motivation

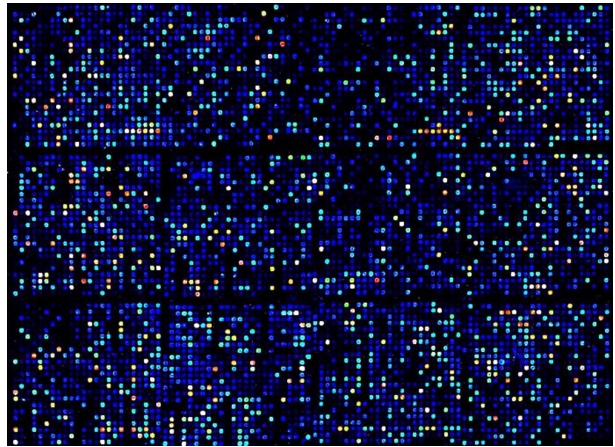
Sparse linear model $Y = \mathbf{X}\beta + \epsilon$

We want to:

- Predict survival (or other phenotype) from genetic information.
- Find predictive genes.

Difficulties:

- Lots of genes (p large) but few patients (n small)
- Computationally slow
- Estimates are poor without structural assumptions



METHODS

- Supervised techniques
  - Lasso ($-$ doesn't respect groups of genes, $+$ sparse solutions)
  - Ridge ($+$ maintains correlated groups, $-$ not sparse)
  - OLS (Not unique, not sparse)
  - Group lasso ($-$ requires knowledge of groups)
- Unsupervised ($-$ all ignore the supervisor)
  - PCA ($+$ finds linear groups)
  - Spectral clustering ($+$ finds nonlinear groups)
  - Hierarchical clustering
- Semisupervised
  - PCR ($+$ finds groups, $-$ nonsparse, $-$ inconsistent estimator)
  - Supervised PCA ($+$ finds groups, $+$ sparse, $-$ strong assumptions)[1]
  - Gene shaving, tree harvesting, others
  - AIMER—finds groups, sparse, weaker assumptions

[1] Assumes that $\text{Cov}(X_j, y) = 0 \Rightarrow \widehat{\beta}_j = 0$

- Treat this like a matrix approximation problem: $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ is $p \times p$. PCA is bad (computationally, statistically)
- (Unsupervised) Matrix approximation literature suggests performing computations on a subset of the columns (see references later).
- (Supervised) We use those that have high marginal correlation with Y: $\mathbf{X}_A$.
- Comparison for reconstruction:

AIMER
$$\begin{pmatrix} \mathbf{X}_A^T\mathbf{X}_A & \mathbf{X}_A^T\mathbf{X}_{A^C} \\ \mathbf{X}_{A^C}^T\mathbf{X}_A & \mathbf{X}_{A^C}^T\mathbf{X}_{A^C} \end{pmatrix}$$

PCR
$$\begin{pmatrix} \mathbf{X}_A^T\mathbf{X}_A & \mathbf{X}_A^T\mathbf{X}_{A^C} \\ \mathbf{X}_{A^C}^T\mathbf{X}_A & \mathbf{X}_{A^C}^T\mathbf{X}_{A^C} \end{pmatrix}$$

SPCA
$$\begin{pmatrix} \mathbf{X}_A^T\mathbf{X}_A & \mathbf{X}_A^T\mathbf{X}_{A^C} \\ \mathbf{X}_{A^C}^T\mathbf{X}_A & \mathbf{X}_{A^C}^T\mathbf{X}_{A^C} \end{pmatrix}$$

- Now just use standard approximation techniques:
  - Compute the SVD
  - Perform some computations to get coefficient estimates
  - “extend” these to the rest
  - Threshold for selection

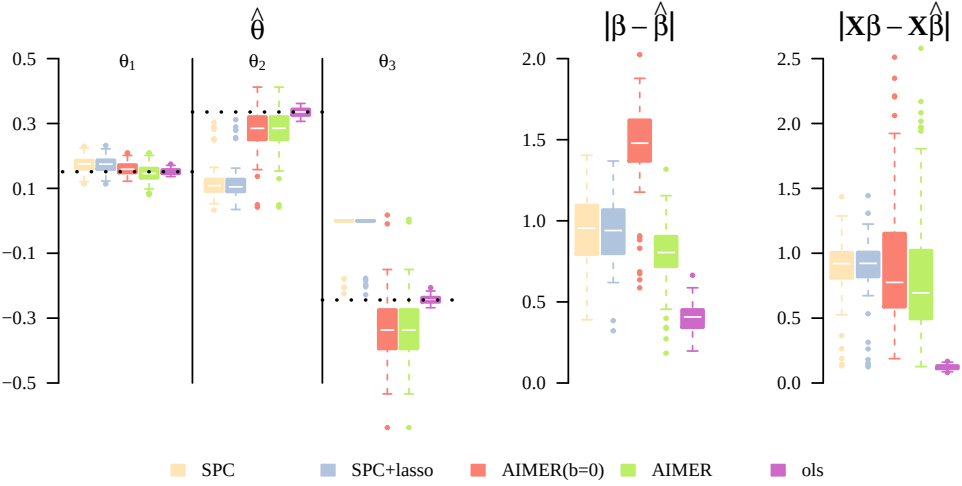
AIMER

Algorithm: Amplified, Initially Marginal, Eigenvector Regression (AIMER)

1: **Input:** centered design matrix $\mathbf{X}$, centered response Y, thresholds $t_*, b_* \geq 0$, integer d.
2: **Compute** marginal correlation t_j between X_j and Y for all j.
3: **Set:** $A = \{j : |t_j| > t_*\}$, $\widetilde{\mathbf{X}} = [\mathbf{X}_A,\ \mathbf{X}_{A^c}]$, $\mathbf{F} = \widetilde{\mathbf{X}}^\top \mathbf{X}_A$.
4: **SVD:** $\mathbf{F} = \mathbf{U}(\mathbf{F})\mathbf{\Lambda}(\mathbf{F})\mathbf{V}(\mathbf{F})^\top$.
5: **Set:** $\widehat{\mathbf{V}}_{[d]} = \mathbf{U}_{[d]}(\mathbf{F})$, $\widehat{\mathbf{\Lambda}}_{[d]} = \mathbf{\Lambda}_{[d]}(\mathbf{F})^{1/2}$, $\widehat{\mathbf{U}}_{[d]} = \mathbf{X}_{new}\widehat{\mathbf{V}}_{[d]}\widehat{\mathbf{\Lambda}}_{[d]}^{-1}$.
6: **Estimate:** $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{V}}_{[d]}\widehat{\mathbf{\Lambda}}_{[d]}^{-1}\widehat{\mathbf{U}}_{[d]}^{T}Y$.
7: **Threshold:** $\widehat{\boldsymbol{\beta}}(b_*) := \widehat{\boldsymbol{\beta}}\mathbf{1}_{(b_*,\infty)}(|\widehat{\boldsymbol{\beta}}|)$.
8: **Return:** Coefficient estimates $\widehat{\boldsymbol{\beta}}(b_*)$.

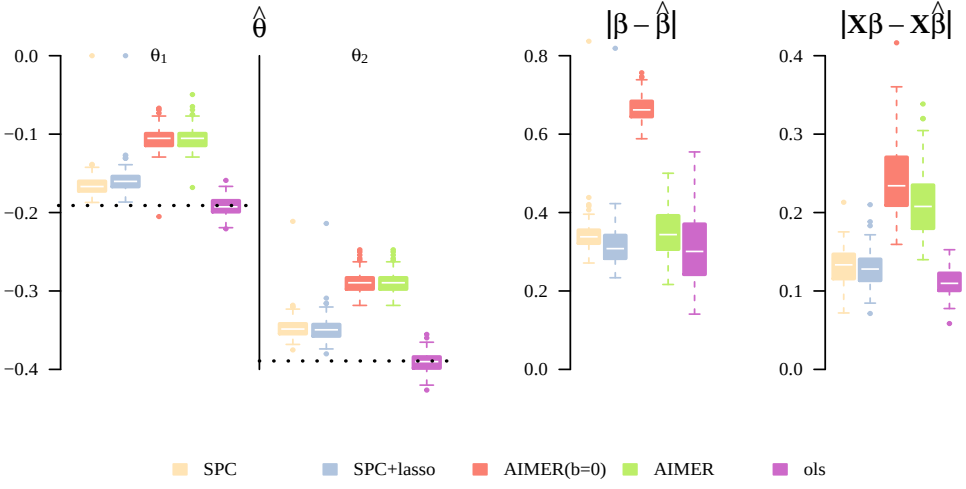
Favorable: $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V} + \mathrm{N}_{n\times p}(0,1), \quad Y = \mathbf{U}\Theta + \mathrm{N}_n(0,1)$

$\mathrm{Cov}(X_1, y) \neq 0,\ \mathrm{Cov}(X_2, y) \neq 0$



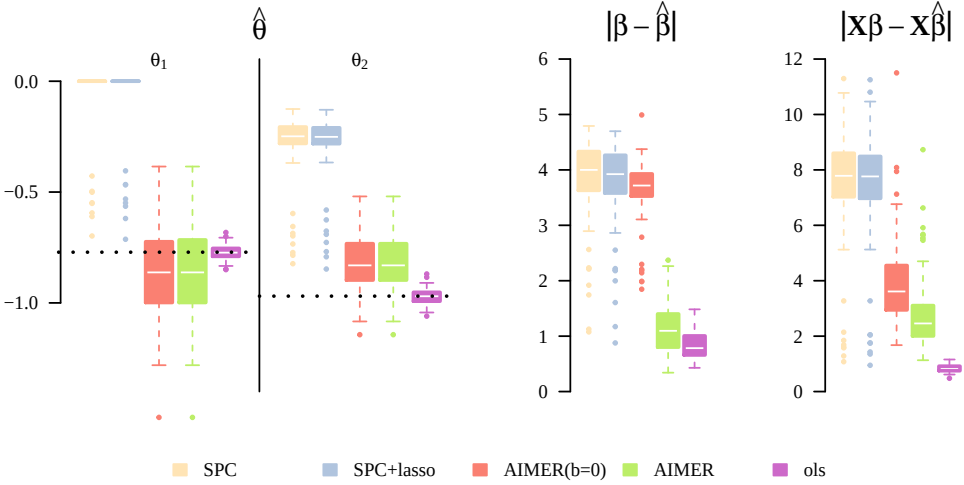
Unfavorable: $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V} + \mathrm{N}_{n\times p}(0,1), \quad Y = \mathbf{U}\Theta + \mathrm{N}_n(0,1),$

$\mathrm{Cov}(X_1, y) \neq 0,\ \mathrm{Cov}(X_2, y) \neq 0$



Unfavorable: $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V} + \mathrm{N}_{n\times p}(0,1), \quad Y = \mathbf{U}\Theta + \mathrm{N}_n(0,1),$

$\mathrm{Cov}(X_1, y) \neq 0,\ \mathrm{Cov}(X_2, y) \approx 0$



- Average MSE on random train/test splits for 4 standard genetics problems. Bolded values indicate the best predictive performance for each type of method for each dataset.

Methods	DLBCL	Breast cancer	Lung cancer	AML
lasso	0.6805	**0.6285**	0.8159	1.9564
ridge	**0.6485**	0.6407	**0.7713**	**1.9234**
SPC	0.6828	0.6066	**0.8344**	2.4214
AIMER	**0.6518**	**0.6004**	1.0203	**1.8746**

Examining DLBCL in detail:

- AIMER selects 26 predictive genes
- 16 genes have been related to lymphoma in the biology literature
- 19 genes have been identified via statistical techniques developed for the DLBCL data
- 4 genes (symbols ALDH2, CELF2, COL16A1, and DHRS9) are newly discovered to the best of our knowledge

CONCLUSIONS

- When to use AIMER?
  - Solve high-dimensional linear regression problem
  - Linear regression model: $Y = \mathbf{X}\beta + \epsilon$
  - Continuous response
  - The number of features is far larger than the number of samples

- Why AIMER?
  - Low computational cost
  - Often, more accurate prediction
  - Selects a small number of predictive features

The paper: R package: My WWW:



REFERENCES

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006) Prediction by supervised principal components. Journal of the American Statistical Association, **101**, 119–137.

Ding, L. and McDonald, D.J. (2017) Predicting phenotypes from microarrays using amplified, initially marginal, eigenvector regression. Bioinformatics, **33**, i350–i358.

Johnstone, I.M. and Lu, A.Y. (2009) Sparse principal components analysis.

Ma, P., Mahoney, M.W. and Yu, B. (2015) A statistical perspective on algorithmic leveraging. The Journal of Machine Learning Research, **16**, 861–911.

Paul, D., Bair, E., Hastie, T. and Tibshirani, R. (2008) "Preconditioning" for feature selection and regression in high-dimensional problems. The Annals of Statistics, **36**, 1595–1618.

Raskutti, G. and Mahoney, M. (2015) Statistical and algorithmic perspectives on randomized sketching for ordinary least-squares. Proceedings of the 32nd international conference on machine learning (icml) (eds F. Bach), & D. Blei), pp. 617–625. PMLR, Lille, France.