Gene Analysis

Jackson Curtis and Brandon Carter

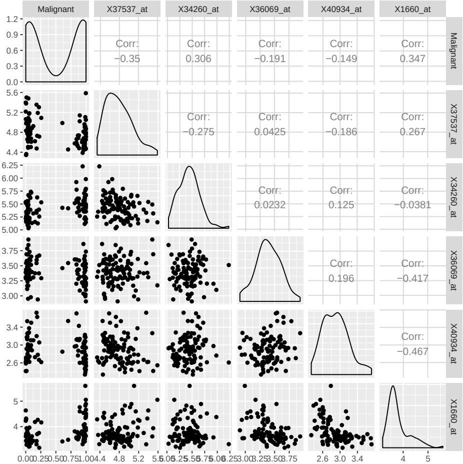
February 21, 2018

Introduction to Gene Data

- Genes encode life
- Genes are tightly regulated on when and how they are expressed
- Cancer arises when this regulation is interfered with, causing uncontrolled growth of cells

Data Features

102 subjects, 5159 cellular responses per subject. Hard to explore, impossible to estimate coefficients in a linear model



Goal

- Rank genes by the influence they have on tumor malignancy
    - Explore different ways of reducing data to prevent overfitting
    - After standardizing, coefficient's magnitude is a measure of variable importance
    - Compare the different methods and find cross-section of important genes

Regularization Models

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 + \lambda \sum_{p=1}^{P} \mathrm{Size}(\beta_p)$$

- λ is a tuning parameter, shrinks the values of $\hat{\boldsymbol{\beta}}$ toward zero
- Ridge Regression : $\mathrm{Size}(\beta_p) = \beta_p^2$
- Lasso Regression : $\mathrm{Size}(\beta_p) = |\beta_p|$

LASSO vs Ridge Regression

- Covariates are centerd and scaled
- Penalized Least squares reduces the flexibility of the model
  - introduces bias, but reduces the variance of our estimates.
  - λ can be chosen through cross validation
- In LASSO the coefficients are zeroed out as λ increases
- In Ridge coefficients approach zero

Regularization Assumptions

- No distirbutional assumptions
- Linearity

Checking Linearity

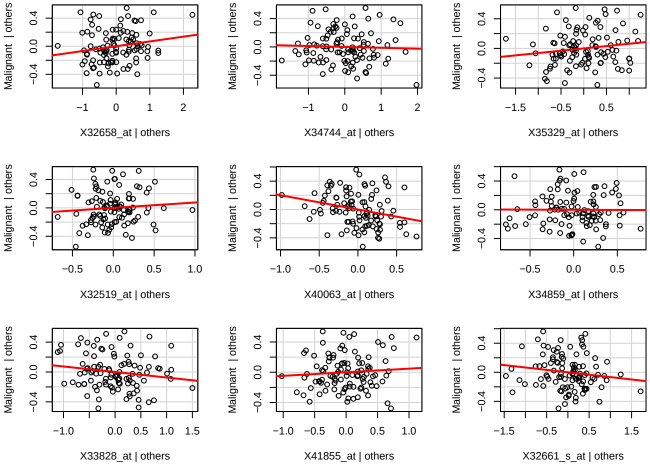


Figure: Added Variable Plots

Principal Component Analysis

Main idea: Use linear combinations of Xs to summarize bulk of information into less than n variables ($p < n$).

Benefits of Eigenvector decomposition:

- Eigenvectors corresponding to largest eigenvalues create axes with maximum variation
- Eigenvectors are orthogonal (no colinearity)
- Just a transformation of Xs, so all principles of linear models still apply
- Straightforward to back transform

PCA Assumptions

Need colinearity in Xs (if independent, combining Xs doesn't create better summary)!

$$Z = X\Psi$$

Ψ is the matrix of the first M eigenvectors. Z is N x M (where $M < n$), so

$$y = \beta_0 + Z\theta + \epsilon$$

has a normal least squares solution.

Partial Least Squares Model

- Similar to PCA, but instead of using eigenvectors, construct linear combinations of Xs weighted by their correlation with the response.
- Intuitively: Xs that strongly correlate with the response will play a bigger role in your Z-matrix.
- Columns of Z can be orthogonalized by regressing X on Z_1, taking residuals, and constructing new columns by measuring correlations between X and the residuals.

PLS Assumptions

- PLS uses simple correlations to build components, so colinearity can cause problems.
- Like PCA, normal regression assumptions apply.
- Reduces bias but increases variance

Model Residuals

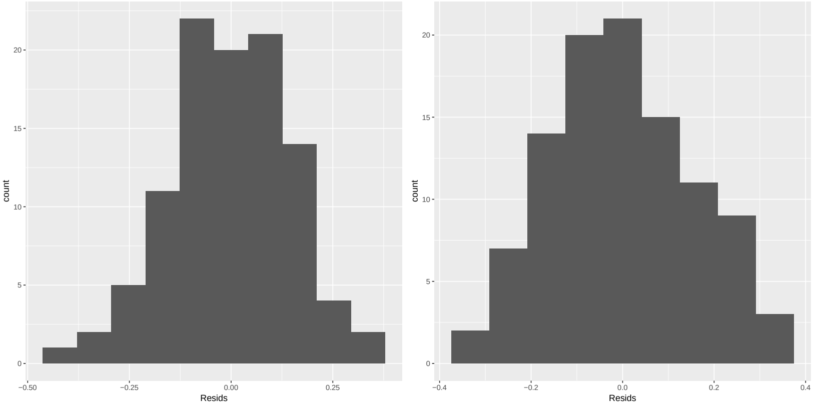


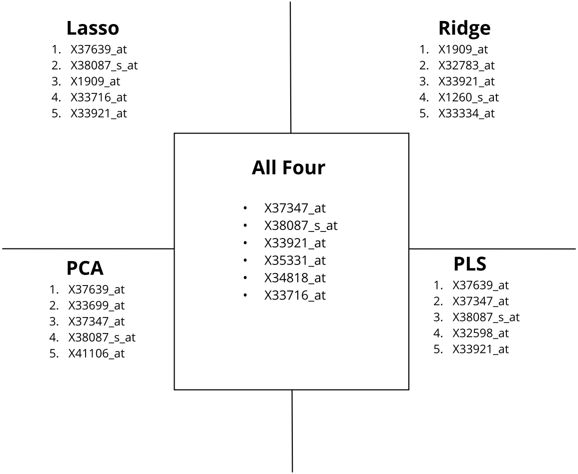
Figure: PCA (left) and PLS (right) show fairly normal residuals.

Report Model Fits

LASSO	Ridge	PCA	PLS
0.932	1.000	0.893	0.883

Table: R^2 values

Top Genes



Lasso

1. X37639_at
2. X38087_s_at
3. X1909_at
4. X33716_at
5. X33921_at

Ridge

1. X1909_at
2. X32783_at
3. X33921_at
4. X1260_s_at
5. X33334_at

All Four

- X37347_at
- X38087_s_at
- X33921_at
- X35331_at
- X34818_at
- X33716_at

PCA

1. X37639_at
2. X33699_at
3. X37347_at
4. X38087_s_at
5. X41106_at

PLS

1. X37639_at
2. X37347_at
3. X38087_s_at
4. X32598_at
5. X33921_at

Confidence Intervals on Top Genes

gene	LASSO	Ridge	PLS	PCA
X33716_at	(-0.0821, -0.0073)	(-0.1253, -0.0138)	(-0.002, -0.0017)	(-0.0021, -0.0019)
X33921_at	(-0.081, -0.0169)	(-0.178, -0.0688)	(-0.0025, -0.002)	(-0.0028, -0.0022)
X34818_at	(-0.0408, 0.0172)*	(-0.102, 0.0088)*	(-0.0024, -0.002)	(-0.0024, -0.0021)
X35331_at	(5e-04, 0.0463)	(-0.0118, 0.0861)*	(0.0016, 0.0019)	(0.0022, 0.0025)
X37347_at	(-0.0418, 0.0064)*	(-0.1024, -0.0161)	(-0.0031, -0.0025)	(-0.0031, -0.0025)
X38087_s_at	(-0.2229, -0.0942)	(-0.1344, -0.0186)	(-0.0028, -0.0024)	(-0.0029, -0.0026)

Table: Intersect of top 50 genes by magnitude from all four methods.

* Confidence interval contains zero.

Model Strengths/Weaknesses

- Different methods coming to same conclusions give us high confidence in the overlapping genes
- Model assumptions are hard to validate and contradictory (colinearity in PCA/PLS)
- No statistical method to combine results for different methods
- No single confidence intervals

Conclusion

- With limited resources begin to research those genes selected by all models.
- Explore genes with highest magnitude of effect from each method.
- Explore relationships between genes.