`glmnet`

Kate Cook

ML4Bio software presentation

2010.02.04

Features

- Linear regression library for R
- Makes regression models and predictions from those models
- Lasso and elastic net regression via coordinate descent (Friedman 2010)
- Very fast
  - FORTRAN-based
  - exploits sparsity in input data
- Simple to use

Availability & installation

- install.packages("glmnet")
- GPL licensed
- Citation, manual etc:

http://cran.r-project.org/web/packages/glmnet/index.html

  – Or just google "glmnet"…

Regularization (review?)

- p features, n observations
- $y = X\beta + \varepsilon$
- Want to minimize the sum of squared errors:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - x_i' \hat{\beta})^2$$

- To reduce overfitting, add a penalty term
- Now we minimize:

$$\sum_{i=1}^{n} (y_i - x_i^T b)^2 + \lambda P_\alpha (b_1, \ldots, b_p)$$

Ridge, LASSO, and elastic net regularization are related

- Ridge regression, LASSO, and elastic net are part of the same family with penalty term:

$$P_\alpha = \sum_{i=1}^{p} \left[\frac{1}{2}(1-\alpha) b_j^2 + \alpha |b_j| \right]$$

- $\alpha = 0$ → ridge regression
- $\alpha = 1$ → LASSO
- $0 < \alpha < 1$ → elastic net!

Features of LASSO and elastic net regularization

- Ridge regression shrinks correlated variables toward each other
- LASSO also does feature selection
  - if many features are correlated (eg, genes!), lasso will just pick one
- Elastic net can deal with grouped variables

One more detail

- Elastic net formulation above is actually the "naïve elastic net"
  - Doesn't perform well in practice
  - Parameters are penalized twice
- How to fix it?

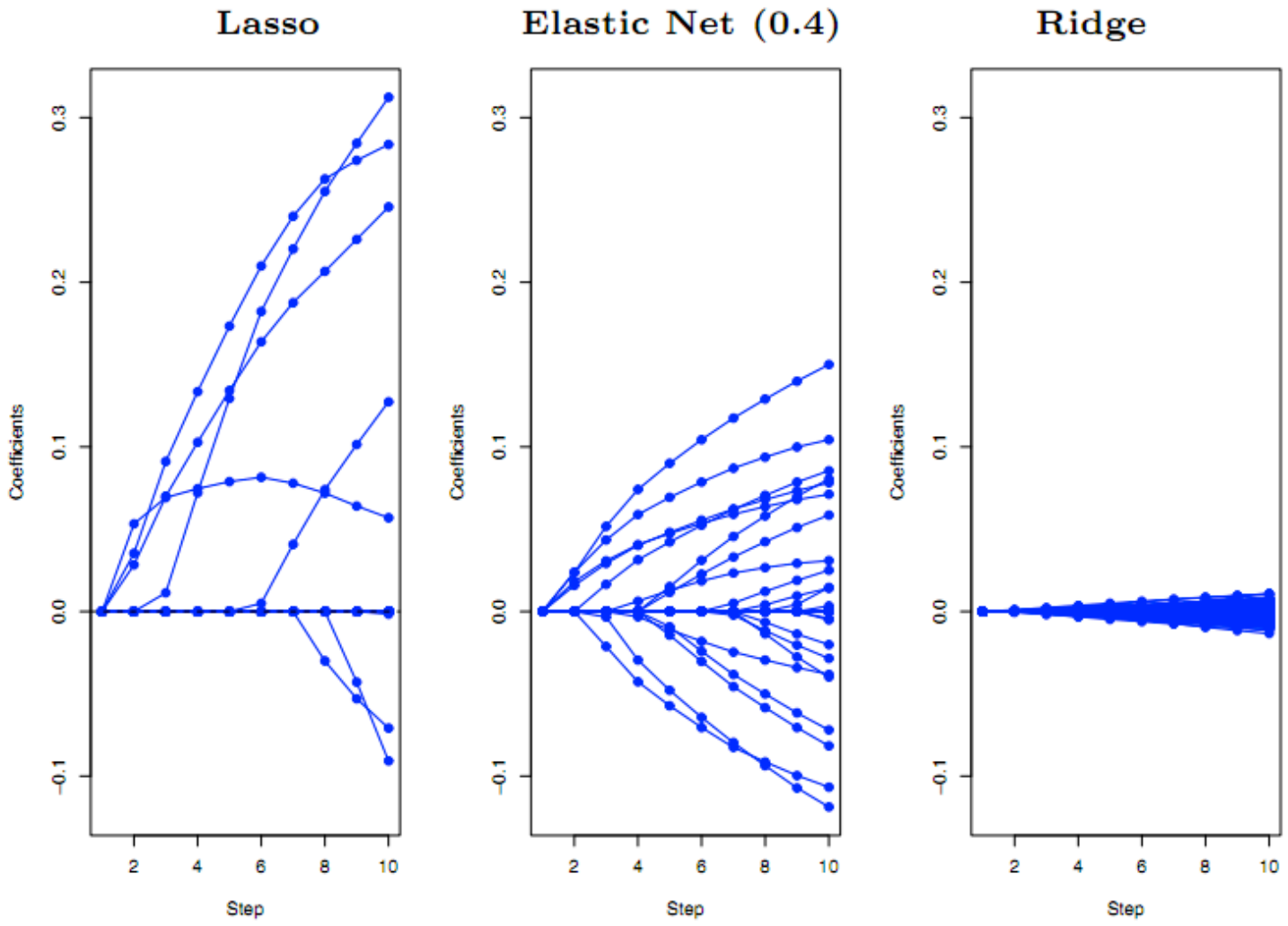
Penalty = $(1-\alpha)\left|\boldsymbol{\beta}\right|_1 + \alpha\left|\boldsymbol{\beta}\right|^2$

$= \lambda_2\left|\boldsymbol{\beta}\right|^2 + \lambda_1\left|\boldsymbol{\beta}\right|_1$ where $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$

$$\hat{\boldsymbol{\beta}}(\text{elastic net}) = (1+\lambda_2)\hat{\boldsymbol{\beta}}(\text{naive elastic net}).$$

Zou and Hastie 2005

Graphically…



Hastie, http://www-stat.stanford.edu/~hastie/TALKS/glmnet.pdf

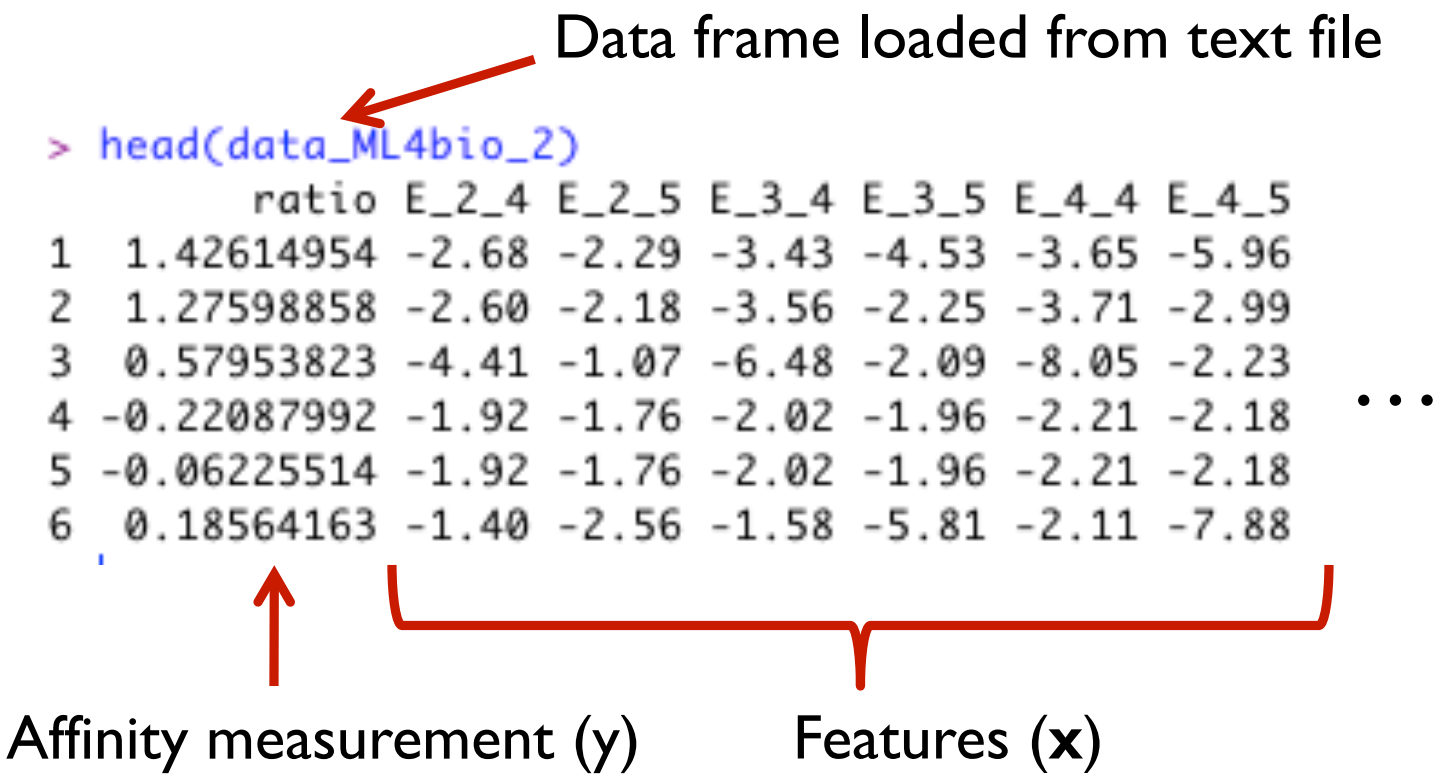
Example – intro

- Determining RNA sequence features predictive of binding to an RNA-binding protein
- Apply LASSO regression to model binding
- Use cross-validation to select the best λ

$$\sum_{i=1}^{n} (y_i - x_i^T b)^2 + \lambda P_\alpha (b_1, \ldots, b_p)$$

- Train model on first 10,000 points, test on last ~5,000

Example - data



• Note: features must be numeric (use dummy variables for categorical data)

Example – fitting the model

- Syntax:

```
Fit <- cv.glmnet(X, y, …)
```

```
> fit1_mmodel.cv<-cv.glmnet(mmodel_ML4bio[1:10000,],
data_ML4bio$ratio[1:10000], alpha=1)
```

Y vector

LASSO

X matrix

Example – making predictions

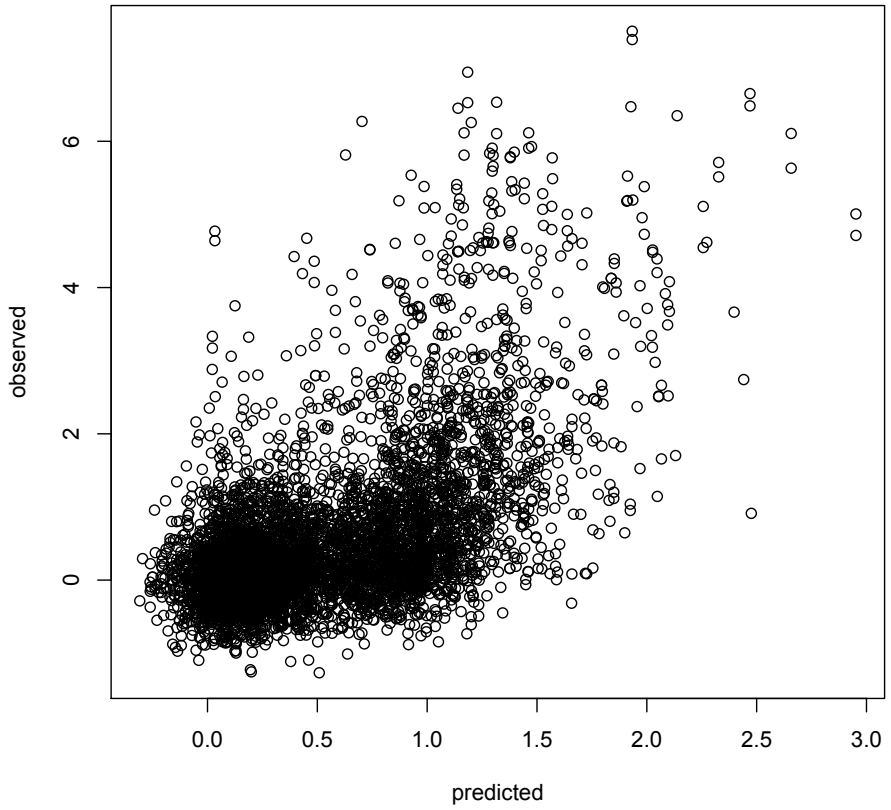
- Syntax:

```
Pred <- predict(Fit, newX)
```

```
> pred_fit1<-predict(fit1_mmodel.cv, mmodel_ML4bio
[10001:15490,])
> cor(pred_fit1, data_ML4bio$ratio[10001:15490])
          [,1]
1 0.5660715
```

Example – comparing predicted to observed

```
> plot(pred_fit1, data_ML4bio$ratio
[10001:15490],xlab="predicted",ylab="observed")
```



Stuff that helped me understand how this works

- http://www-stat.stanford.edu/~hastie/TALKS/glmnet.pdf
  - Theory behind LARS and coordinate descent, speed trials, biological examples
- Friedman, Hastie & Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent*, J Stat Soft, 2010
- Zou and Hastie, *Regularization and Variable Selection via the Elastic Net*, J Royal Stat Soc B, 2005