Development of Predictive Models and Feature Selection Using LASSO and Elastic Net

Pushpike Thilakarathne, Martin Otava, Nolen Joy Perualila, Tatsiana Khamiakova, Adetayo Kasim, and Ziv Shkedy

I-BioStat

25 Sep 2012

Introduction

- Feature selection is an important scientific requirement of genomic studies when p is large than no.of samples (N « p)
- In most cases predictors are correlated
- Usually no.of predictors that relevant or informative are very few
- It is difficult to identify informative features when noisy predictors are present.

outline

1. *Introduction*
   - Shrinkage Methods
   - Validation of Performance
2. *Research Question on QSTAR Project*
   - Predictive Finger Prints
3. *Prediction Gene Module using Fingerprints*
4. *Application of LASSO and Elastic Net*
5. *Leave One Out and K-Fold Cross Validations for LASSO*
6. *3-Fold Cross Validations for LASSO*
   - Mostly selected 24 features
7. *Relaxed LASSO (Meinshausen, 2007)*
   - Results based on Relaxed LASSO
   - 3-fold cross validations for Relaxed LASSO
8. *Elastic Net*
   - Choice of mixing parameter α
   - Results based on Elastic net and LOOCV
   - Results based on 3-fold CV
9. *Summary of Results*
   - presence and absence of finger prints
10. *Wrap up*

LASSO and Elastic Net (l_1 and l_2 penalty

Lasso and Elastic Net

Linear regression methods for prediction and variable selection when the number of predictors exceeds the number of sample units ($p \gg N$).

Lasso: (l_1 penalty)

the largest number of predictors is equal to number of samples

$$\beta^{lasso} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

λ is chosen such that the mean squared prediction error is minimum

Elastic net: Lasso $\Longleftarrow$ Elastic net $\Longrightarrow$ Ridge regression

Number of predictors with non zero weights (coefficients) depend on the penalty

$$\beta^{ENet} = \arg\min_{\beta} \left(||\mathbf{y} - X\beta||^2 + \lambda \left(\alpha \sum_{i=1}^{p} |\beta_i| + \frac{1}{2}(1-\alpha) \sum_{i=1}^{p} \beta_i^2 \right) \right)$$

Elastic Net penalty is a mixture of the l_1 (lasso) and l_2 (ridge) penalties. α is the mixing parameter.

Assessment of prediction error

- 3-fold cross validations
- leave one out cross validations [LOOCV]

Predictive Fingerprints

Predictive model for gene expression using Finger Prints [FP] information (GC)

Information:

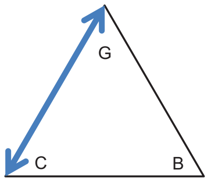
1. Gene module (Y)
2. FPs matrix (X)

Output:

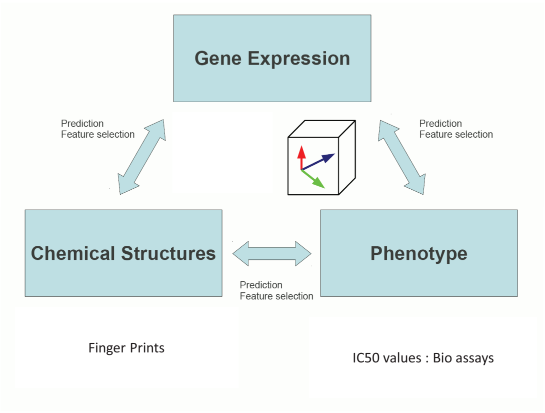
1. List of predictive features (FPS).
2. Predictive model

Methods:

LASSO and Elastic Net



QSTAR Project


Gene Expression

Chemical Structures

Phenotype

Prediction
Feature selection

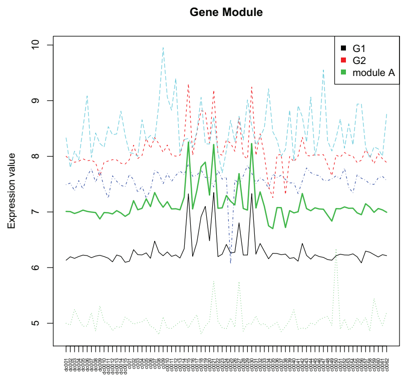
Prediction
Feature selection

Prediction
Feature selection

Finger Prints

IC50 values : Bio assays

Prediction of Gene Module using Fingerprints


Gene Module

Focused on a Gene Module A which is formed by average expressions of the two correlated genes.

Data Structure

- Response: Expression levels of Gene Module A: mean gene expressions of the two correlated genes on 62 compounds
- Predictors: Unique fingerprint features 268 out of 16698 finger prints: each fingerprint feature is a binary vector indicating whether it is present or absent in the 62 compounds.

The Model

$$\mathbf{Y_i} = \beta_0 + \sum_{j=1}^{P} \beta_j \mathbf{FP_{ij}} + \varepsilon_i$$

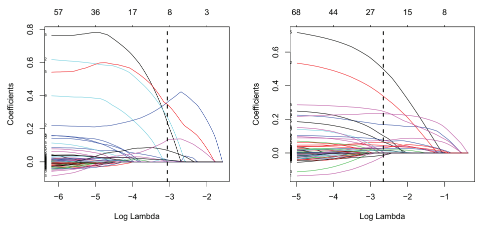
$$\mathbf{FP_{ij}} = \left\{ \begin{array}{ll} 1 & \text{if presents in the compound} \\ 0 & \text{Otherwise} \end{array} \right. .$$

Leave One Out cross validations for LASSO

assess the model in terms of prediction error:

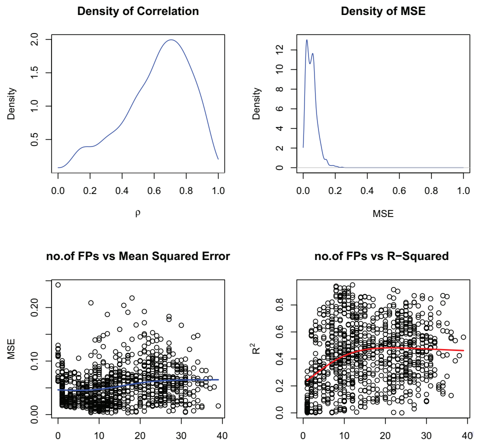
- the estimated R^2 value is 57.024%
- correlation between Y and $\hat{Y}_{-j}$ is 0.755.
- the estimated *MSE* is about 0.078

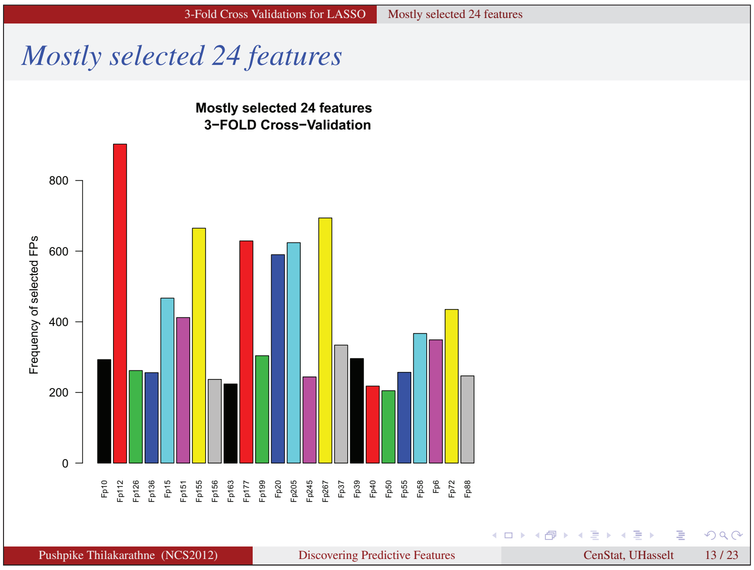
LASSO and Elastic Net



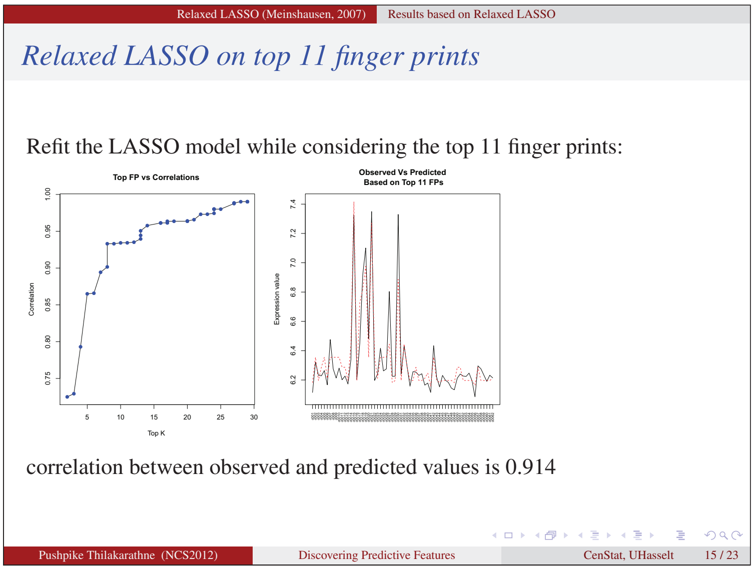
10 finger prints selected with LASSO (left panel) while Elastic net (right panel) with mixing parameter, $\alpha = 0.3$, selects 21 finger prints. LASSO results is subset of the Elastic Net results.

3-Fold Cross Validations for LASSO





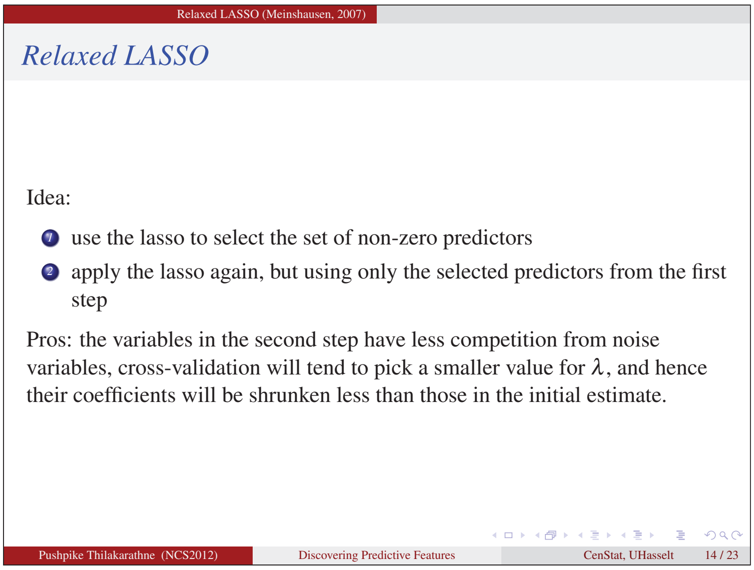
Mostly selected 24 features



Relaxed LASSO on top 11 finger prints

Refit the LASSO model while considering the top 11 finger prints:

correlation between observed and predicted values is 0.914

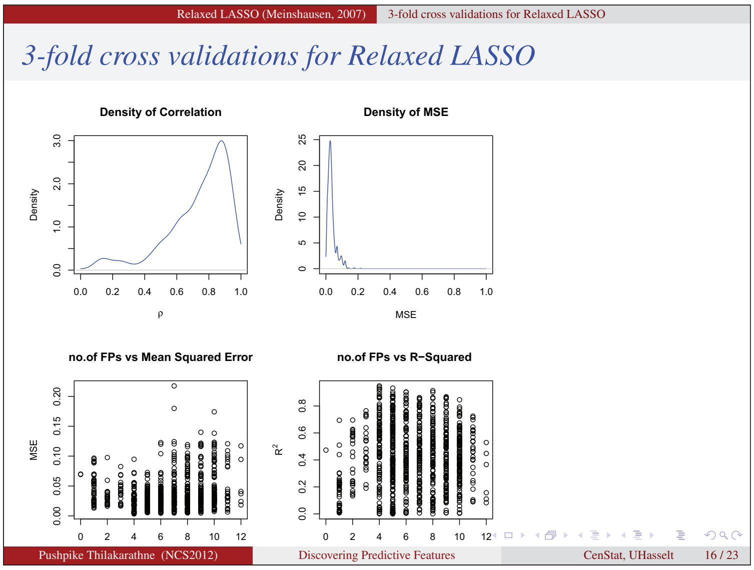


Relaxed LASSO

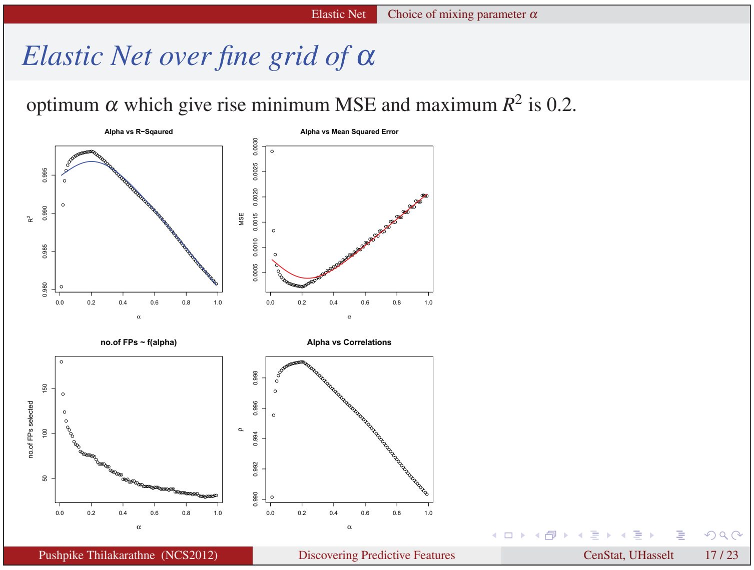
Idea:

1. use the lasso to select the set of non-zero predictors
2. apply the lasso again, but using only the selected predictors from the first step

Pros: the variables in the second step have less competition from noise variables, cross-validation will tend to pick a smaller value for λ, and hence their coefficients will be shrunken less than those in the initial estimate.

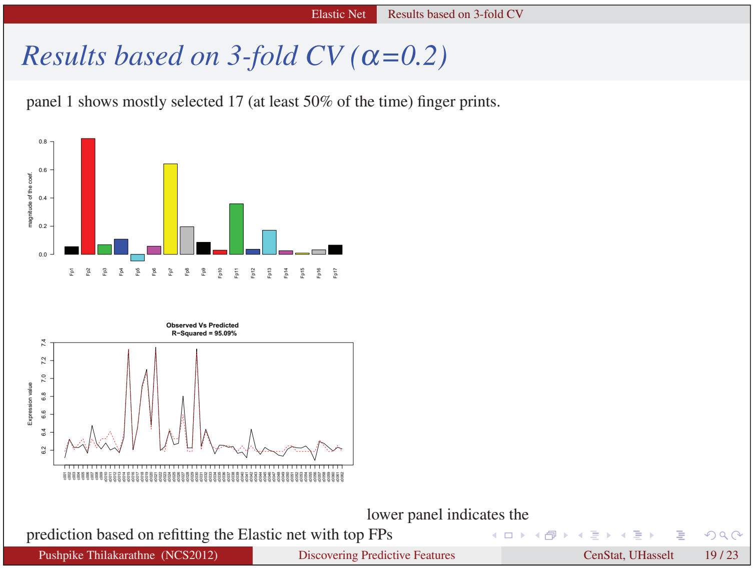


3-fold cross validations for Relaxed LASSO



Elastic Net over fine grid of α

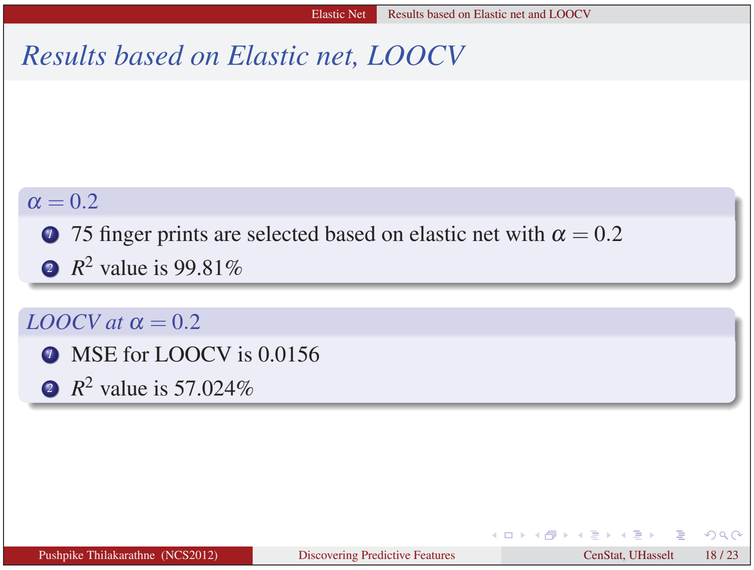
optimum α which give rise minimum MSE and maximum R^2 is 0.2.



Results based on 3-fold CV (α=0.2)

panel 1 shows mostly selected 17 (at least 50% of the time) finger prints.

lower panel indicates the prediction based on refitting the Elastic net with top FPs



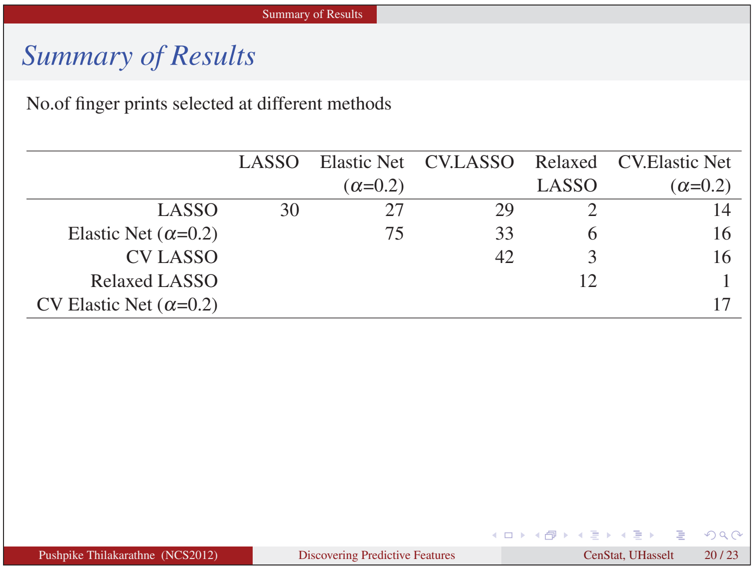
Results based on Elastic net, LOOCV

$\alpha = 0.2$

1. 75 finger prints are selected based on elastic net with $\alpha = 0.2$
2. R^2 value is 99.81%

LOOCV at $\alpha = 0.2$

1. MSE for LOOCV is 0.0156
2. R^2 value is 57.024%



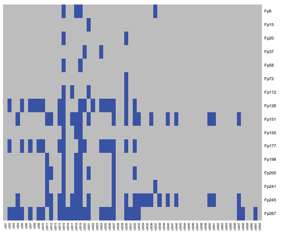
Summary of Results

No.of finger prints selected at different methods

	LASSO	Elastic Net (α=0.2)	CV.LASSO	Relaxed LASSO	CV.Elastic Net (α=0.2)
LASSO	30	27	29	2	14
Elastic Net (α=0.2)		75	33	6	16
CV LASSO			42	3	16
Relaxed LASSO				12	1
CV Elastic Net (α=0.2)					17

presence and absence of finger prints

if fingerprint is present 🟦
if fingerprint is NOT present ⬜



Thank you for your attention!
Dank u voor uw aandacht!

Winding up

- different methods lead to different finger print signatures
- there are over lapping fingerprints
- core fingerprint list: fingerprints common in 3-fold cross validated LASSO and 3-fold cross validated Elastic net ($\alpha = 0.2$)
- the Gene module A can be predicted using these finger print features and correlation between predicted and observed values are around 0.95
- need to check whether these finger prints represent interesting chemical compounds