Randomized Elastic Net

Randomized elastic net in simple words means that you repeatedly add random penalties to the features (in our case most often genes) making it 'harder' for those features to get non-zero coefficients, you estimate coefficients, and again add random penalties, estimate coefficients… etc. You repeat this procedure N times. The number of times a feature had non-zero coefficient gives you a feeling how important this feature is, so you can rank-order features (genes) based on the number of times they had non-zero coefficients. Top ones are obviously best predictors.

Here is slightly modified explanation of randomized elastic net I wrote for shared project with extra comments in green:

Lambda and predictors selection (randomized elastic net) plus classification

Elastic net for patient vs control module genes.

Here we assume that we can explain disease status (i.e. healthy or ill) by gene expression:

$$y_j = \sum_{i=1}^{p} x_i \beta_i$$

where y_j denotes disease status for subject *j* (e.g. healthy or patient label)
x_i denotes expression of gene *i*
β denotes model coefficients
p denotes number of genes/predictors

Elastic net can be summarized as:

$$\beta^{\lambda,W} = argmin_{\beta \epsilon \mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X\beta}\|_2^2 + \frac{\lambda(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2 + \alpha\lambda\|\boldsymbol{\beta}\|_1$$

Here we have assumed α=0.5 and in order to select optimal λ parameter we have performed cross-validation using elastic net (in other words we create classifier with elastic net, classify patients and controls, and select lambda for which we get best model based on the one-standad-error criteria = see below).

For selected λ we have performed randomized elastic net implemented based on modified random lasso described in (*1*) to select best set of predictors (In this project we have selected ALL genes with non-zero coefficient in at least one run of randomized elastic net as a good predictors. For CRC project I have selected predictors that were selected more than 45% of the times – in general it should be modeled to get best cutoff for the predictors, at some point we worked on that but never finished – mainly lack of time and expertise… but I would be happy to discuss the theory if you're interested).

Using selected predictors and selected lambda we have performed final classification.
If you plan to do classification using elastic net I also attach function called '*classify_elastic_net()*'. I have added it at the end of the example code.

P-value for classifier is calculated using Wilcoxon rank-sum test based on the classifier output subject labels (You basically check if the discriminatory function score - classifier output -you get for patients differs significantly from the scores you get for healthy controls – done by the code).

Pseudocode for the randomized elastic net:

Standard normalize expression data (done by the code).
Select number of folds for cross-validation.
Run classification on all genes to get best lambda (we select lambda based on 1 standard error of 'deviance' criterion ("1 SE rule as selecting the most parsimonious model whose error is no more than one standard error above the error of the best model").
N times add random penalties to predictors and calculate coefficients.
Score how many times a predictor had non-zero coefficient.

Example how to run the code

See *Run_example_code_randomized_elastic_net.m*

References

1. N. Meinshausen, P. Bühlmann, Stability selection. *J. R. Statist. Soc. B* **72, Part 4**, 417–473 (2009).