Relaxed LASSO in GO space

Luis Pedro Coelho

Instituto de Medicina Molecular

*Suggested Page Limit: 3 pages.*

**Summary Sentence**: Relaxed LASSO after grouping gene expression values by GO term.

**Background/Introduction**

*Please try to address the following points:*

* *What is the motivation for your approach? This will include any previous work and observations that you have made about the NCI-DREAM data to suggest your approach is a good one. Provide the reader with an intuition of how you approached the problem*
* *What is the underlying methodology used (e.g., SVM or regression)?*
* *Where there any novel approaches taken in regards to feature selection, data imputation, ranking, etc?*

We approached this issue as being mainly a feature construction and selection issue.

The basis of this method is to bring in outside annotations (GO terms) in order to reduce the problem.

For this problem, I used penalized regression, in particular relaxed LASSO. Standard L₁ penalization achieves two goals: (1) sparsity in that many (even most) coefficients are exactly zero; and (2) regularization in that non-zero values are smaller (in absolute value) than in unregularized regression. Due to the very large number of variables, a very large penalization was needed for sparsity, which led to over-regularlization. Relaxed LASSO, by separating, the effects of sparsity and regularization achieves a better balance between these two.

**Methods**

Only gene expression data was used, with RNA-seq & microarray data being combined to form a average value.

1. Use the seq calls provided from the RNA-seq data to determine which genes are changing. Only these “active” genes will be used.
2. Preprocess the RNA-seq data with a log-transform:

r' = log(r + 1), where r' is the new value and r is the old one.

1. Combine the RNA-seq and the microarray data into a single prediction by averaging the z-scored values.
2. Any gene without both an RNA-seq and microarray measurement was discarded.

Now, we obtain a matrix of gene expression which contains only the active genes. We do not use this matrix directly. Instead we look up GO terms for all genes (ignoring the Cellular Component vocabulary). All genes which map to the same GO term are combined. We combine the genes by reducing with the following function

maxabs(a, b) = if |a| > |b| then a else b

That is, we keep the maximum in absolute terms, while preserving its sign. For learning, we further process the data by

* + 1. re-normalising to zscores (the maxabs transform does not preserve this property)
    2. selecting features which correlate with at least 20 drug signatures

Finally, we use relaxed lasso for the optimisation. A first lasso pass with λ=.000225010113525 is used (this is roughly 2⁻¹²) is used for feature selection, and a second pass with λ' = λ/10 is used for the final learning. An initial attempt to use cross-validation to learn λ led to massive over-fitting and a value in the middle of the range was chosen.

The optimisation uses coordinate descent for optimisation and ignores the regression error in the missing entries.

B\* = arg max\_B ½ 1/n ΣWᵢⱼ(Yᵢⱼ – (BX)ᵢⱼ)² – λ|B|

Where the weight of example i,j, Wᵢⱼ, is zero if the entry is missing.

**Conclusion/Discussion**

*This section should include a short summary and any insights gained during the algorithm. For example, which dataset was most informative? You can include future directions. You may also add some discussion on the general performance of your methodology (if you wish) and if there were pitfalls, what are they?*

The main driving force behind these ideas were the pressure for feature selection and dimensionality reduction.

One major problem with this approach is the need for setting parameters (the penalization factors λ). Cross-validation was a possible solution, but due to the small number of datapoints, the variation between different folds was enormous and the final result was very unstable, a value was then hard-coded. A more robust solution would have been desirable.

**References**

**Authors Statement**

LPC developed the methodology, implemented it, and wrote the report.