**Relaxed Lasso in GO Space**

*Luis Pedro Coelho*

*Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa, Lisboa, Portugal*

**Summary Sentence**: We used relaxed lasso regression after grouping gene expression values by GO term.

**Background/Introduction**

Given the inputs and the continuous output variable, the problem can be posed as a simple regression problem. However, the number of input dimensions greatly exceeds the number of examples (just the gene expression data has over twenty-thousand dimensions, while only circa 30 training points are available per drug—the number is variable as some datapoints are missing). Therefore, we focused on reducing the number of dimensions in order to then be able to apply simple regularized regression to the problem.

We worked only with the gene expression data (microarray and RNA-seq). The fact that two different data sources were available for the same underlying phenomenon also led us to prefer this data. We grouped genes by their GO terms in order to obtain a smaller number of dimensions.

After projecting into GO terms, we used relaxed lasso (Meinshausen, 2007). 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 to achieve sufficient sparsity, which led to over-regularlization. Relaxed lasso uses a strong penalization to choose coefficients and then a smaller penalty to find their final values.

**Methods**

We first combined the microarray and RNA-seq data into a single measure per gene, and secondly projected to GO terms.

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,where is the new value and is the old one, followed by normalization to z-scores. Similarly, z-score the microarray data.
3. Combine the RNA-seq and the microarray data into a single prediction by averaging the z-scored values.
4. 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 keeping only the largest value (in absolute terms):

.

For learning, we further process the data by thresholding: we set to one whenever the feature at that point is two standard deviations away from the mean (in either direction). We then select only features that are significantly correlated with at least twenty different drug outputs (p-value < 0.01, estimated by a permutation test). This allows for a modicum of information transfer between different drugs, which are otherwise treated separately: features that are informative about many drugs are less likely to be statistical artifacts than if this measurement was applied independently for each drug.

Finally, we use relaxed lasso for the optimisation. A first lasso pass with λ=2⁻¹²·¹² 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.

On the output side, we normalize data by subtracting the per-drug average from each entry, so that we regress on a centered value. Finally, the optimisation uses coordinate descent for optimisation and ignores the regression error in the missing entries, i.e., we solve the following problem:

Whererepresents the weight of example . In our case, we set it to 1 if the example has data and to 0 if it is not (and we set to an arbitrary value). Therefore, missing entries are ignored. For optimisation, we used coordinate descent as proposed by Friedman et al. (2010) for this class of problems.

**Conclusion/Discussion**

The main driving force in the choices we made, was the pressure for feature selection and dimensionality reduction. For example, in the case of genes where one of the microarray or RNA-seq measurements was missing, we decided to discard them instead of relying on lower-quality measurements.

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 for the regularizer was then hard-coded. A more robust solution would have been desirable. Similarly, other choices in the methodology (for example, the function to aggregate genes by GO term) were evaluated by cross-validation, but it would have been preferable to be able to rely on internal metrics.

**References**

Nicolai Meinshausen, Relaxed Lasso, Computational Statistics & Data Analysis, Volume 52, Issue 1, 15 September 2007, Pages 374-393, ISSN 0167-9473, 10.1016/j.csda.2006.12.019.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent." Journal of statistical software 33.1 (2010): 1.

**Authors Statement**

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