Gene Expression Level Association Analysis for Yeast Data

Group 5

1. Yeast Data

We explored the effects of different statistical methods by by the analysis of a yeast eQTL data set described by Brem and Kruglyak (2005), where $n = 112$ segregants were grown from a cross between two budding yeast strains, BY4716 and RM11-1a. For each of the segregants, gene expression was profiled on microarrays containing 6216 genes, and genotyping was performed at 2957 markers.

1.1 Data Preparation

1.1.1 Processing Markers Data

The DNA of the same organism has a very high similarity. For example, all humans are 99.9% identical and, of that tiny 0.1% difference. Considering the similarity of markers, we found that different markers in the data may have identical or differ at most few yeast samples after exploratory data analysis.

Inspired by Yin and Li (2011), we combined the markers into blocks such that markers with the same block differed by at most one sample, and one representative marker was chosen from each block. To accomplish that, we used hierarchical clustering with complete-linkage criteria and cut the clustering tree at distance $d = 1$. Note that the value of the marker data is only 0 or 1, so we can use any distance function, like Euclidean distance or Manhattan distance, to obtain the same result.

We can further reduce the number of markers, observing that different markers have different functions, and some markers have no effect on the gene expression we want to study. Thus a marginal gene–marker association analysis was then performed to identify markers that are associated with the expression levels of at least two genes with a p-value less than 0.05, resulting in a total of $p = 776$ markers.

1.1.2 Processing Expression Level Data

2. Methodology

Here we are going to introduce methods we used to select variables.

2.1 Unit-Response Regression

2.2 Grouped-Response Regression

2.3 Multi-Response Regression

References

Rachel B Brem and Leonid Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences*, 102(5): 1572–1577, 2005.

Jianxin Yin and Hongzhe Li. A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The annals of applied statistics*, 5(4):2630, 2011.