Linear Data Fusion with drCCA

[1]Abhishek Tripathi, [2]Arto Klami, [2]Samuel Kaski
[1]Department of Computer Science, University of Helsinki
[2]Laboratory of Computer and Information Science, Helsinki University of Technology

Motivation

- A fast, linear method to combine multiple data sets
- All data sets measure the same entity, for example, genes, over different conditions
- Noise reduction by discarding data-specific variation
- Emphasize what is interesting in the data by keeping shared variation
- First step of exploratory data analysis, any method can be applied to the result

Methods

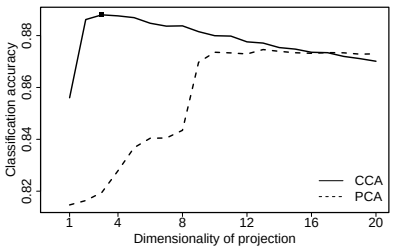
- Center and whiten all data sets to remove data-specific variation
- Perform PCA on column-wise concatenation $\boldsymbol{Z}$ of data sets to capture shared variation
- The d-dimensional combined representation is given by $\boldsymbol{P_d = ZV_d}$, projecting $\boldsymbol{Z}$ onto first d principal directions $\boldsymbol{V_d}$
- d is selected by a statistical test based on randomization

Connection to Canonical Correlation Analysis:

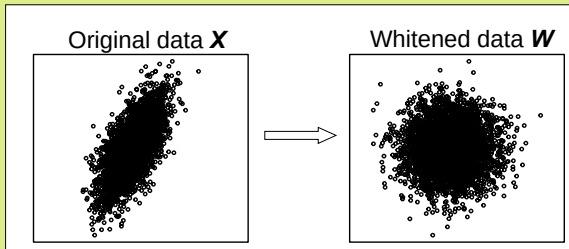
- Generalized CCA (gCCA) can be used for the same task
- Perform gCCA on data sets and sum the scores to compute $\boldsymbol{P_d}$
- Given $\boldsymbol{U_{i,d}}$ be the first d canonical weights for data matrices $\boldsymbol{X_i}$, $\boldsymbol{P_d = \sum_i X_i U_{i,d}}$

Classification of Cell cycle regulated genes

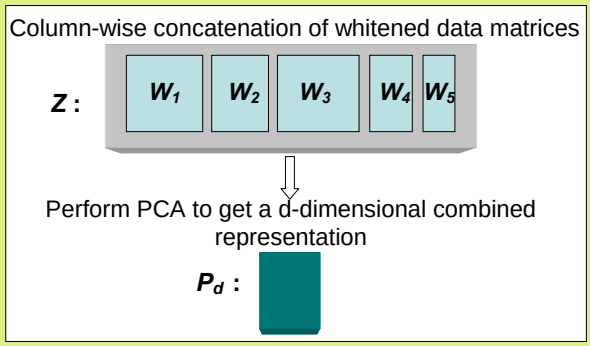
Data: We study yeast (*Saccharomyces cerevisiae*) cell cycle regulation based on 5 different data sets[2], using the power spectrum of fourier-transformed data. The task is to detect cell-cycle regulated genes in a classification setting.



Results: K-nearest neighbor classifier is used in leave-one-out manner, and the classification accuracies of CCA-based method is compared with the accuracy of directly using PCA on the collection of data sets. Clearly, CCA-based method provides a better representation for separating cell-cycle regulated genes from the others.



$\boldsymbol{W = X \Sigma^{-1/2}}$, $\boldsymbol{\Sigma}$: covariance matrix of $\boldsymbol{X}$



Implementation

Implementation of the method in R is available at http://www.cis.hut.fi/projects/mi/software/drCCA

Characterizing the commonalities in different Leukemia subtypes

Data: Gene expression measurements of 5 different subtypes BCR-ABL, E2A-PBX1, MLL, TEL-AML1 and T-ALL, of pediatric acute lymphblastic leukemia [1]. We analyzed 22,283 genes for 31 patients. RMA is used for preprocessing. The task is to characterize the commonolities in the subtypes.

Results: We picked the genes with the highest variance in the combined representation, and studied enrichments of GO terms from the biological processes category. The most enriched terms were related to immune response. Direct application of PCA on the collection found the same GO terms, but the enrichments were less significant in 9/10 cases.

References

1. Ross et al, Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, **102**(8), 2951–2959.
2. Spellman et al, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, **9**,3273–97.