A mixed effects clustering model for multi-species time course gene expression data

Kevin H Eng  
University of Wisconsin, Madison  
Department of Statistics

Dan Kvitek  
University of Wisconsin, Madison  
Department of Genetics

Grace Wahba  
University of Wisconsin, Madison  
Department of Statistics

Audrey Gasch  
University of Wisconsin, Madison  
Department of Genetics

Sündüz Kele  
University of Wisconsin, Madison  
Department of Statistics

Data Motivation

Time course design. Arrays are repeated measurements of a single species over time. The complete set comprises the experimental replicate, a time course.

Arrays  
Time 1  Time 2  Time 3  Time 4  Time course

Multi-species time course design. Is an extension of time course experiments, where we have multiple time course measurements for many species. We draw the phylogeny to emphasize that we expect correlation between species. A replicate is now several time courses.

Simulation Studies

Methods for comparison. Each of the following clustering models differs in its consideration of the cluster mean or covariance structure, and thus its marginal model. The first is a vanilla application of mclust (Fraley and Raftery 2002) to the data, then we consider an exploratory procedure where we fit gene-wise anova models and cluster their parameter estimates. Finally we compare the mixed effects clustering to two fixed effects clustering models, one demonstrates what happens when we ignore the covariance structure and the other shows the limitations of the fixed effects model fitted to the data with array-level heterogeneity.

Model Marginal Model  Mean Covariance

mclust on data  $Y_i \sim N(k_i, \Sigma_i)$  means only on data  $\Sigma_i$

mclust on parameters  $p_i \sim N(k_i, \Sigma_i)$  means on parameters  $\Sigma_i$

fixed effects, diagonal  $Y_i \sim N(k_i, \Sigma_{diag})$  mean parameters diagonal $\Sigma_{diag}$

fixed effects, general  $Y_i \sim N(k_i, \Sigma_{full})$  mean parameters $\Sigma_{full}$

mixed effects  $Y_i \sim N(k_i, \Sigma)$  mean $\Sigma$ parameterized $\Sigma$

Simulation 1: Effect of Random Effects Variance.

We generate data from the mixed effects model where we add in random effects variance. We choose k=5 well separated clusters of 200 genes each. We parameterize the fixed effects model while increasing $\sigma^2$ increases the variance. At $\sigma^2=1$ we have set the variance so that no method should be able to detect separate clusters.

Simulation 2: Effect of Clustering Noise.

There is biological evidence that some genes are really singleton clusters, that is, they do not belong in a cluster. We simulate a clustering scenario where we vary $\sigma^2$, the percentage of genes which truly belong in clusters, and measure the ability of each method to identify noise and the effect on the clustering result (in the set of genes which are not noise). Data are drawn from a moderately hard problem (p=52) with k=5 clusters, but we substitute some whole clusters for genes which do not belong to a cluster (but still have significant effects). Since all the methods admit a measure of cluster membership, for N many noise genes we select the N bottom ranked genes, and measure classification as a singleton or as a clustable gene.

A mixed effects clustering model for multi-species time course gene expression data

Data Analysis

Heat shock stress response data. Our example data come from four strains of S. cerevisiae subjected to heat shock stress. Each time course consists of 6 time points and there are two replicates per time course per strain.

Penetrating step. In order to fit the model, we first fit gene-wise fixed effects anova models and compute F-tests for main effects. If a gene has a significant F-test for time (FDR corrected), it is included in the model fitting. If a gene does not have a significant time test, but does have a significant F-test for strain effect, we consider it in a separate category "strain only effect."  

Model Fitting. We used the model using the clustering results from mclust on parameters and pick an appropriate number of clusters by BIC.

Alignment with GO categories. We expect some attenuation in association with functionally annotated categories due to the intended separation of different effects in different strains. Below are two clusters found to be enriched for GO annotation "purrine metabolic process." Of the 14 genes in the category those two clusters account for 7 of them, 2 have a "strain only effect" and the remaining 5 appear in other clusters. Thus a significant portion of the purrine genes are heavily repressed (up to 8 fold) in every strain except the lab-strain (S288C). Accumulating baseline expression data suggests that when they are unstressed, the strains show no significant difference in expression.

Inference on Phylogeny

Hypothesis Testing. The relevant phylogenetic information in for each cluster is contained in the species random effects, which are, by assumption, normal random variables. We can get the conditional model for each cluster under different assumptions and carry out the tests with an appropriate form of the LRT.

A test for effect. We test for the presence of a phylogenetic effect by comparing the log likelihoods of the following models. If we conclude $H_0$, then the data are consistent with an independent tree.

$H_0$: "No detectable phylogenetic effect"  
$H_1$:  "Consistent with Neutral Evolution"  
$H_2$:  otherwise

$G(X)$ under $H_0$  
$G(X)$ under $H_1$

References and Acknowledgements


3. KVE is supported by the following grants: NSF DBS 0604752 (GR), P20101(1-12752 (CA).

Work in progress

- Implementing model selection within model fitting
- Implementing the testing framework
- Mining clustering results for biological information