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

Kevin H Eng
eng@stat.wisc.edu
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

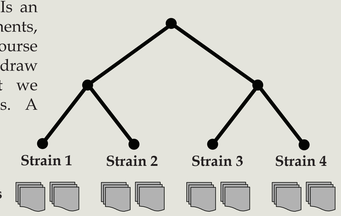
Audrey Gasch
University of Wisconsin, Madison
Department of Genetics

Sündüz Keleş
keles@stat.wisc.edu
University of Wisconsin, Madison
Department of Statistics
Department of Biostatistics and Medical Informatics

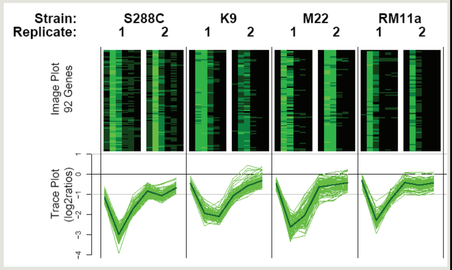
Data Motivation

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

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.



Example Plot. The cluster to the right represents a set of genes found to have significantly different responses to heatshock among four yeast strains. Each strain has two replicate time courses with six arrays each in the image plots (time goes left to right). The trace plots show each gene's time profile (averaged over replicates) and also plots the mean profile. We can see the added dimension in multi-species data: instead of just finding separate time patterns, we also need to consider species specific differences.



Model Statement

- **Interest in gene function versus differential expression.** We employ clustering techniques instead of a gene-by-gene analysis, choosing to model and group together genes with similar effects.
- **Time course measurements.** Our *model based clustering* approach improves on the standard hierarchical clustering strategy by parameterizing the mean profile, accounting for systematic variation, and giving us a measure of clustering uncertainty.
- **Measurements across many species.** An extra random effect accounts for species variation, adding the possibility of inference on parameters, *effects due to phylogeny*.
- **Mixed Effects clustering model.** Conditional on cluster membership, we model each gene's expression with a mixed effects model:

$$Y_{gr}\,\big|\,(U_g = k) = X\beta_k + Ma_{gk} + Wb_{gk} + e_{gkr}$$

$$\beta_k = \left(\beta_0,\ \beta_1^{(T)},\ldots,\beta_{T-1}^{(T)},\ \beta_1^{(S)},\ldots,\beta_S^{(S)}\right) \quad \text{(fixed effects for time and species)}$$

$$a_{gk} = (a_1,\ldots,a_T)\,\big|\,(U_g = k) \sim N(0, A_k) \quad \text{(time random effect)}$$

$$b_{gk} = (b_1,\ldots,b_S)\,\big|\,(U_g = k) \sim N(0, B_k) \quad \text{(species random effect)}$$

$$e_{gkr}\,\big|\,(U_g = k) \sim N(0, \sigma_k^2 I) \quad \text{(measurement error)}$$

Marginal Model

$$Y_k = Y_{gr}\,\big|\,(U_g = k) \sim N(X\beta_k, MA_kM + WB_kW + \sigma_k^2 I), \quad Y_{gr} = \sum_k \pi_k Y_k$$

Conditional Model

$$Y_{gr}\,\big|\,(U_g = k, a_{gk}, b_{gk}) \sim N(X\beta_k + Ma_{gk} + Wb_{gk}, \sigma_k^2 I)$$

- The time random effect covariance, A_k, allows covariance between time points and the species random effect covariance, B_k, captures the dependence between species.
- Model fitting proceeds by an *EM algorithm*, which returns parameter estimates and cluster memberships.

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 refit 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 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 : otherwise

$$b_{gk} \sim \begin{cases} N(0, \sigma^2 I) & \text{under } H_0 \\ N(0, \Sigma) & \text{under } H_1 \end{cases}$$

A test for neutral evolution. Neutral evolution hypotheses represent a particular tree structure and thus a form for the covariance. We estimate B_{ne} the covariance for a known set of "psuedogenes" and fit the model assuming the same covariance up to a scale constant. Under H_0, the data contain no signal unexplained by neutral evolution.

H_0 : "Consistent with Neutral Evolution"
H_1 : otherwise

$$b_{gk} \sim \begin{cases} N(0, \rho B_{ne}) & \text{under } H_0 \\ N(0, \Sigma) & \text{under } H_1 \end{cases}$$

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.

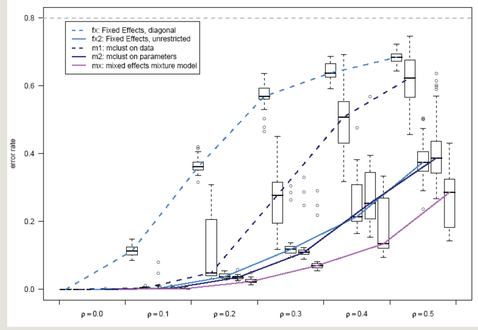
Method	Marginal Model	Mean	Covariance
mclust on data	$Y_{gr} \sim N(\mu_g, \Sigma_g)$	of data only	on data
mclust on parameters	$\beta_g \sim N(\theta_g, \Sigma_g)$	ad hoc	on parameters
fixed effects, diagonal	$Y_{gr} \sim N(X\beta_k, \sigma_k^2 I)$	parameterized	diagonal
fixed effects, general	$Y_{gr} \sim N(X\beta_k, \Sigma)$	parameterized	on parameters
mixed effects model	$Y_{gr} \sim N(X\beta_k, V_k)$ $V_k = WA_kW + MB_kM + \sigma_k^2 I$	parameterized	parameterizes factors

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. ρ=0 parameterizes the fixed effects model while increasing ρ increases the variance. At ρ=1 we have set the variance so that no method should be able to detect separate clusters.

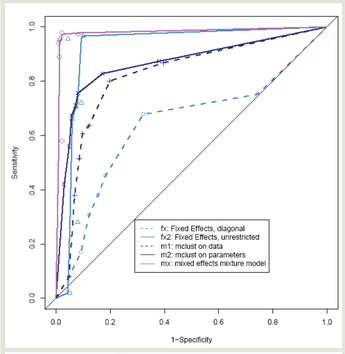
Data Generating Model :

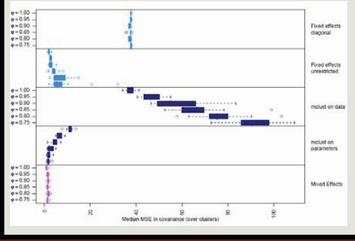
$$Y_{gr}\,\big|\,(U_g = k) = N(X\beta_k, V_k(\rho))$$
$$V_k(\rho) = \rho^2(WA_kW + MB_kM) + \sigma_k^2 I$$

Misclassification Rate. The plot on the left demonstrates that regardless of the difficulty of the clustering problem, methods which restrict the covariance perform poorly and methods which mis-parameterize the mean are penalized less. Note that in the fixed effects case (ρ =0) all methods perform well.



Simulation 2– Effect of Clustering Noise. There is biological evidence that some genes are really singletons, that is they do not belong in a cluster. We simulate a clustering scenario where we vary φ, 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 (ρ =0.2) with K=20 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 clusterable gene.





"ROC" type plot. We plot the median sensitivity and median 1-specificity over a number of simulations for varying φ. We make the noise call based on the gene's distance from its cluster center. It is clear that the mixed effects model outperforms the less specific models at picking singleton genes. Again, the methods which are flexible in their covariance structure do well.

Target Covariances. The boxplots below compare the clustering dependent empirical estimates of the covariances to the true covariances; they are the median over all clusters of the mean squared element-wise distance between the estimated covariance and the target covariance for each method. The two unrestricted covariances (fixed effects and mclust on the data) are increasingly sensitive to noise, while the diagonal covariance fixed effects model is insensitive to noise because its covariance is very inflexible. For mclust on parameters, it is notable that the decision to remove variation at the gene level is costly when the proportion of un-clusterable noise is low.

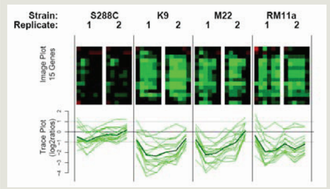
Data Analysis

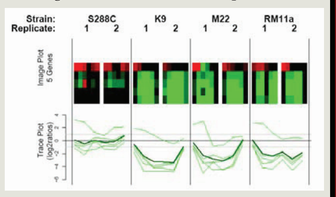
Yeast heatshock stress response data. Our example data come from four strains of *S. cerevisiae* subjected to heatshock stress. Each time course consists of 6 time points and there are two replicates per time course per strain.

Prescreening step. In order to fit the model, we first fit genewise 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."

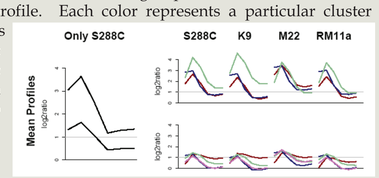
Model Fitting. We seed the model using the clustering results from mclust on parameters and pick an appropriate number of clusters by BIC.

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 are enriched for GO annotation "purine metabolic process." Of the 14 genes in the category these 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 purine genes are heavily repressed (up to 8 fold) in every strain except the lab strain (S288C). Accompanying basal expression data suggests that when they are unstressed, the strains show no significant difference in expression.

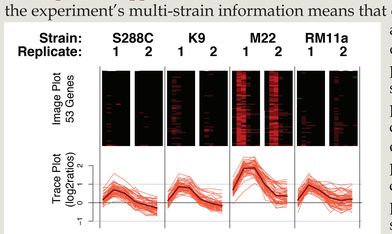




GO categories with strain information. Our clustering results inform the behavior of particular clusters in different strains as follows. Consider the set of induced Environmental Stress Response (iESR) genes from the study by Gasch et al. (2000). In our experiment, if we performed standard clustering on the iESR genes we would only see two clusters, plotted on the left, with similar mean profiles and differing only in intensity. When we add strain information, we see the seven strong patterns (more than 10 genes) in the plots on the right, divided into two groups based on which means correspond to which S288C Only profile. Each color represents a particular cluster Notice that we can now identify genes whose expressions differ across the strains. In particular the green curve in the upper plot is more induced in S288C and K9 than the other two strains.



Novel patterns appear with more strain information. Appropriate characterization of the experiment's multi-strain information means that different types of patterns might appear for the same genes in different strains. The following pattern shows a spike unique to a single strain. In contrast to the pattern in the clusters above, this cluster represents the identification of a consistent signal across potentially many functional categories, and provides a starting point for identifying the cause of such a difference.



Work in progress

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

References and Acknowledgements

- C Fraley and AE Raftery. (2002) "MCLUST: Software for model based clustering, discriminant analysis and density estimation." Technical report 415. University of Washington.
- A Gasch et al. (2000) "Genomic expression programs in the response of yeast cells to environmental changes." Mol. Biol. Cell. 11: 4241-4257.

- KHE is supported by the following grants: NSF DMS 0604572 (GW), PhARMA HG03747 (SK).

THE UNIVERSITY of WISCONSIN MADISON