Memorial Sloan-Kettering Cancer Center

Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series

Year 2007 *Paper* 14

On Comparing the Clustering of Regression Models Method with K-means Clustering

Li-Xuan Qin* Steven G. Self†

*Memorial Sloan-Kettering Cancer Center, qinl@mskcc.org

†Fred Hutchinson Cancer Research Center, sgs@scharp.org

On Comparing the Clustering of Regression Models Method with K-means Clustering

Li-Xuan Qin and Steven G. Self

Abstract

Gene clustering is a common question addressed with microarray data. Previous methods, such as K-means clustering and hierarchical clustering, base gene clustering directly on the observed measurements. A new model-based clustering method, the clustering of regression models (CORM) method, bases the clustering of genes on their relationship to covariates. It explicitly models different sources of variations and bases gene clustering solely on the systematic variation. Both being partitional clustering, CORM is closely related to K-means clustering. In this paper, we discuss the relationship between the two clustering methods in terms of both model formulation and implications on other important aspects of cluster analysis. We show that the two methods can both be considered as solutions to a least squares problem with missing data but they each concern a different type of least squares. We also show that CORM tends to provide stable clusters across samples and is particularly useful if the cluster averages are used as predictors for sample classification. Finally we illustrate the application of CORM to a set of time course data measured on four yeast samples, which has a complicated experimental design and is difficult for K-means to handle.

1. INTRODUCTION

Advances in molecular technologies have led to an explosion of research to study how genomic alterations mediate disease etiology and progression. Gene expression microarrays allow simultaneous monitoring of thousands of genes at the mRNA level in tissue specimen from normal or disease samples. Clustering, a useful tool to look for unknown groupings of objects [1], has become an important part of the analysis of gene expression data, owing to the pioneering work of Eisen et al. [2]. The gene expression profile of a sample reflects a particular state of the sample, such as tissue type, disease status, and cell cycle phase [3-5]. By looking for clusters of genes that have similar expression levels across samples and sample states, researchers hope to better understand gene functions, genetic pathways, regulatory circuits, and ultimately disease etiology and treatment. Cluster analysis can also be used to cluster samples; we will focus on the problem of gene clustering in this paper.

Several methods have been applied to the problem of gene clustering. They can largely be classified to two categories: (1) algorithmic clustering methods, such as K-means clustering and hierarchical clustering [2, 6]; and (2) model-based clustering methods, such as the multivariate normal mixture model [7, 8]. These methods generally do not take into account of the experimental design, such as cross-sectional (CS) design, longitudinal with no replication (LNR) design, and longitudinal with replications (LWR) design.

A newly proposed clustering method, the clustering of regression models (CORM) method employs regression to model gene expression and clusters genes based on their relationship between expression levels and sample covariates [9]. Different from previous clustering methods, CORM partitions systematic variation from non-systematic variation and bases the clustering on systematic variation only. CORM is applicable to data collected under various designs for microarray experiments, including CS, LNR, and LWR designs.

K-means clustering is a commonly used clustering method for gene expression data [10]. In addition, K-means is a special case of a model-based clustering method – the multivariate normal mixture model, where the covariance matrix is diagonal (more precisely, scalar). Both K-means and CORM are partitional clustering methods, which concern the problem of the optimal partitioning of a given set of objects into a prespecified number of mutually exclusive and exhaustive clusters. In this paper we will investigate the relationship between K-means and CORM in terms of model formulation as well as other important but often overlooked aspects of cluster analysis, for example, selection of genes, characterization of clusters, and application of clusters. We will show that K-means and a special case of CORM can both be considered as solutions to a least squares problem, but they each concern a different type of sum of squares. Compared to K-means, CORM tends to find gene clusters that are stable across samples and thus provides a nice way to generate predictors for sample classification when averages of gene expression are used as predictors [11]. The rest of the paper is organized as follows. Section 2 briefly describes K-means and CORM. Section 3 discusses the relationship between K-means and CORM. Section 4 illustrates the application of CORM to a set of LWR data measured on four yeast samples, which has a complicated experimental design and is difficult for K-means to handle. Section 5 concludes the paper with some remarks.

2. METHOD

2.1. K-means Clustering

Given a set of objects, K-means clustering seeks a partition of all objects into K groups to minimize the total within group sum of squared Euclidean distance [12]. The minimum could, in theory, be found by searching over all possible clusterings; however, this approach is computationally prohibitive when the number of objects is large. An iterative procedure is instead adopted to search for the minimum. Specifically, K-means starts with an initial value for the cluster centers, then iterates between the cluster-assigning step (each object is assigned to the closest cluster center) and the cluster-center-recalculating step (each cluster center is updated as the average of objects assigned to that

cluster), until convergence. It has been pointed out that K-means is equivalent to assuming a multivariate normal mixture model with component distributions having the same scalar covariance matrix and equal mixture proportion, and then fitting the model using an EM algorithm to maximize the classification likelihood [13]. Here notation is introduced for LWR data, including CS data and LNR data as special cases. Let $\mathbf{y}_{gi}$ denote the vector of expression levels for gene g and sample i, $\mathbf{y}_g = (\mathbf{y}^T_{g1}, \ldots, \mathbf{y}^T_{gm})^T$ the vector of expression levels for gene g for sample 1 through sample m, G the number of genes, and K the number of clusters. Let u_g denote the cluster membership for gene g. The model underlying K-means can be written as

$$\mathbf{y}_g \mid (u_g = k) = \boldsymbol{\mu}_k + \boldsymbol{\varepsilon}_g$$
$$\boldsymbol{\varepsilon}_g \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{I})$$

where $\boldsymbol{\varepsilon}_g$ is the vector of measurement errors, $\mathbf{I}$ is an identity matrix, and u_g is a random variable on (1, 2, . . . , K) with probabilities $\pi_k = 1/K$. Cluster memberships are considered as missing data in the EM algorithm: cluster-assigning step corresponds to E-step and cluster-center-recalculating step to M-step.

There are two reasons that we chose to compare CORM with K-means but not MNM in this paper. One is that K-means is more commonly-used and is more familiar to researchers working with microarray data. The other is that MNM with diagonal covariance matrix is more appropriate for the purpose of clustering genes when the experimental design is cross sectional, where elements are samples and can be reasonably assumed to be independent of each other.

2.2. The CORM Method

For the problem of differential expression analysis, the regression modeling framework has been employed to characterize systematic variation in the expression profile of each gene and distinguish it from random variation. Differential expression is identified by contrasting expression levels measured under different experimental conditions or by

identifying dependencies on concomitantly measured covariates. The resulting estimated regression models can provide an accurate and precise description of expression profiles. Similarly, the regression model framework can be used for the problem of gene clustering: systematic variation is separated from random variation and gene clustering is based solely on the systematic part of the variation. We call this as the clustering of regression models method (CORM).

Let $\mathbf{X}_{gi}$ ($n_{gi} \times p$) denote the design matrix for gene g and sample i, $\mathbf{F}_{\boldsymbol{\beta}k,\xi k}$ the conditional distribution of genes in cluster k given the covariates with parameters $\boldsymbol{\beta}_k$ and ξ_k, $\boldsymbol{\beta}_k$ ($p \times 1$) the vector of regression coefficients, and $\mu(.;\ .)$ the regression function. The model underlying CORM can be written as

$$\mathbf{y}_{gi} \mid (\mathbf{X}_{gi}, u_g = k) \sim \mathbf{F}_{\boldsymbol{\beta}k,\xi k}$$
$$E(\mathbf{y}_{gi} \mid \mathbf{X}_{gi}, u_g = k) = \mu(\mathbf{X}_{gi}; \boldsymbol{\beta}_k)$$

where u_g is a random variable on (1, 2, . . . ,K) with probabilities (π_1, π_2, . . . , π_K). Complete specification of the CORM modeling framework requires identification of the error structure (parameterized by ξ), which depends on the form of the regression model. The specific form of the regression model used for CORM is flexible. For example, it can be the linear model, the linear mixed model, the nonlinear model, and the nonparametric regression model. Its choice should depend on the experimental design and the scientific question. The EM algorithm can be used to fit the CORM model [14, 15]. Implementation details can be found in [9] for the clustering of linear models (CLM) method and the clustering of linear mixed models (CLMM) method.

3. COMPARISON

3.1. Comparing K-means and CORM

K-means and CORM are similar in that they both seek a partition of objects, as opposed to a hierarchical tree, and are both implemented through an iterative EM algorithm.

However, the two methods base clustering on different features of a gene. The feature of interest for K-means is the vector of sample-specific expectations for a gene. For each sample-specific expectation, sample size is 1 and genes in the same cluster are used as replicates for its estimation. K-means does not make any assumption on the relationship between the expected expression level and the covariates and is 'model-free' in this respect. The feature of interest for CORM is the vector of regression parameters shared by samples for a gene. It separates systematic variation from random variation and increases clustering precision especially when the sample size is large (Figure 1). Note that although CORM is 'model-based' in terms of modeling expression levels with covariates, the regression model itself can be either parametric or nonparametric (for example, use of spline basis for modeling longitudinal data).

[Figure 1 about here.]

The different gene features considered by K-means and CORM also have implications on other important issues of cluster analysis. We will comment on three such issues here, one for before gene clustering and two for after gene clustering.

(a) Selection of genes. A microarray provides measurements on thousands of genes, but it is common to select a small subset (tens to hundreds) of genes to cluster, especially for partitional clustering. One reason to select a subset is to keep the computation manageable and fast. Another reason is to try to exclude the uninformative genes to prevent them from deteriorating the clustering. For K-means, however, 'uninformative' is not well defined. One might select the most variable genes. However, on one hand, it does not distinguish genes with large signal and genes with large noise when including genes; on another hand, it does not distinguish genes with small signal and genes with small noise when excluding genes. For CORM, one could first select informative genes using a per-gene regression model and a significance cutoff appropriately adjusted for multiplicity [16], and then cluster genes with significant systematic variation to find those that are similarly associated to the covariates. CORM and regression-based differential

expression analysis can thus form an integrated framework for the analysis of microarray data.

(b) Characterization and interpretation of clusters. After clustering genes, it is useful to determine the cluster signatures for the identified clusters. Often they are set to be the cluster centers. CORM clusters can be identified by their regression coefficients and have a specific interpretation depending on the experimental design. For example, we can tell whether a gene cluster tends to be up-regulated or down-regulated comparing diseased samples to normal samples. The interpretability of CORM clusters allows a more interpretable comparison of genes clusters identified in different data sets with similar experimental designs – not only the clustering of genes can be compared but also the characteristics of the clusters.

(c) Application of clusters. Average of genes in the same cluster has been proposed to act as predictors for sample classification [11]. CORM tends to find clusters that more stable across samples, as we will show later. In addition, CORM, but not K-means, provides an explicit prediction rule for new genes that are measured on a new set of samples.

CORM provides an alternative clustering method for scenarios when K-means has limitations. For example, while applicable to both CS data and LNR data, K-means does not distinguish the two experimental designs. K-means cannot naturally handle LWR data – profiles of a gene need to be averaged or connected first. K-means might not use all information in the data; for example, in a longitudinal study, it considers time points to be exchangeable and ignores their ordering and correlation. Unlike K-means, CORM can naturally deal with missing value on any gene or sample (under the assumption of missing at random) as well as imbalanced experimental design (for example, different sampling times for different samples in a longitudinal study). Moreover, CORM can easily incorporate technical replicates together with biological replicates in a hierarchical manner.

The gains of CORM depend on the truth of the regression model and its robustness to model misspecification. Ideally, the design of an experiment determines the gene-related feature available for clustering and hence informs parameterization of the regression model for CORM. Experimental design should be chosen to produce the feature that most likely reflects biological clusters of interest. For example, a longitudinal design can be used to find clusters of genes that behave similarly across time, while a cross-sectional design can be used to find clusters of genes that behave similarly across different levels of covariate (for example, disease stage).

3.2. Comparing K-means and CLM

The CLM method can be applied to CS data to find genes that have similar expression levels across a set of, homogeneous or heterogeneous, samples. In CS data, a single expression value is measured for a gene on a sample; hence, $\mathbf{y}_{gi}$ reduces to y_{gi} and $\mathbf{X}_{gi}$ to $\mathbf{x}_{gi}$. The underlying model for K-means and CLM can be written, respectively, as

$$y_{gi} \mid (u_g = k) = \mu_{ki} + \varepsilon_{gi}$$
$$\varepsilon_{gi} \sim N(0, \sigma^2)$$
and
$$y_{gi} \mid (\mathbf{x}_i, u_g = k) = \mathbf{x}^T_i \boldsymbol{\beta}_k + \varepsilon_{gi}$$
$$\varepsilon_{gi} \sim N(0, \sigma^2_k)$$

K-means is closest to CLM when CLM assumes a common variance for measurement errors ($\sigma^2_k = \sigma^2$) and a common mixture proportion ($\pi_k = 1/K$) for all clusters, as well as uses the classification likelihood for the EM algorithm. Under this specific scenario, the only difference between the two models is that, for each cluster, K-means assumes a different mean for each sample, while CLM assumes the same mean for samples at the same covariate level.

Both K-means and CLM can be considered as solutions to a least squares problem. Take CS data on a group of homogeneous samples as an example. Given the data, the sum of

squared distance between individual expression levels, y_{gi}, and the global average across genes and samples, y, is fixed. This sum of squares can be decomposed into three components:

$$\sum_{g,i}(y_{gi} - y)^2 = \sum_{g,i}(y_{gi} - y_{ki})^2 + \sum_{g,i}(y_{ki} - y_k)^2 + \sum_{g,i}(y_k - y)^2 \qquad (1)$$

where y_{ki} stands for the sample-specific cluster mean for sample i and cluster k and y_k for the cluster mean averaged across samples for cluster k. K-means seeks a partition of genes to minimize the first component, while CLM seeks a partition to minimize the sum of the first and second components. The second component measures the variability of sample-specific cluster centers among samples, which suggests that CLM seeks a partition that has stable cluster centers across samples. Stable cluster centers are particularly desired if they are further used to form sample classifiers [11].

Alternatively, the sum of squares can be decomposed into another three components:

$$\sum_{g,i}(y_{gi} - y)^2 = \sum_{g,i}(y_{gi} - y_g)^2 + \sum_{g,i}(y_g - y_k)^2 + \sum_{g,i}(y_k - y)^2 \qquad (2)$$

where y_g stands for the gene-specific mean averaged across samples for gene g. The sum of the first two components in equation (1) is equivalent to the sum of the first two components in equation (2). The first component in equation (2) measures the non-systematic variation, while the second component measures the systematic variation. The first component is independent of gene clustering (that is, variation unrelated to clustering); hence, CLM bases the clustering only on the second component – the systematic variation between genes (that is, variation related to clustering).

4. APPLICATION

To study the regulation of cell cycle in yeast, the Breeden Lab at Fred Hutchinson Cancer Research Center studied gene expression of cell cycle for both wild type (WT) yeast and a single mutant (SM) yeast with Yox1 knocked out. They used α factor for cell

synchronization and measured 6,227 ORFs at 5-min intervals for 120 min. cDNA microarray was used with a common reference mRNA and log ratios are used to measure expression levels. Replicate measurements were obtained for both WT yeast and SM yeast. We are interested in the co-expression behavior of cell cycle dependent genes. Using the three microarray data sets on yeast cell cycle published by [5], Zhao et al. (2001) identified a set of 256 genes to be cell cycle dependent in at least two out of the three data sets using a per-gene regression modeling approach [17]. We focused on these 256 periodic genes in our analysis.

The primary goal of our analysis is to cluster genes that have similar expression patterns among WT yeast. As a secondary goal, we also clustered genes using both WT yeast and SM yeast to identify genes whose expression patterns are changed by the mutation. Unlike K-means clustering, CLMM can explicitly accommodate both the replication and the sample covariate (mutation status). In addition, CLMM can naturally deal with the imbalanced experimental design: WT had one bad time point at 105 min and SM had three at 25 min, 40 min, and 55 min, where bad time points were assessed by the Breeden Lab based on technical considerations and were removed from the cluster analysis. There was also missing data: 41 measures belonging to 17 genes for WT data and 17 measures belonging to 17 genes for SM data were clearly outliers based on judgments from the Breeden Lab and were most likely due to technical failures of the measurement procedure rather than reflecting true biological variation that should be modeled. See supplementary materials for details.

4.1. Cluster WT Data

The CLMM method was applied to cluster the 256 genes using WT data. The design matrix for fixed effects was the B-spline basis for time 0-120 min with 7 equally spaced knots. The number of knots was set to be 7 to allow a flexible modeling of the expression profiles and at the same time to avoid overfitting. Within a reasonable range, the clustering results were not sensitive to the number of knots for the B-spline basis. The number of samples is small in this data and does not allow the application of

bootstrapping-based methods, such as the bootstrapped maximum volume measure [9]; hence we fit the CLMM model for several numbers of clusters, including K=6, K=7, and K=8. As K increased from 6 to 7, a group of 17 genes was separated from a loose cluster and formed a new cluster. According to the clustering estimated in [5] based on the time to the first peak, all 17 genes belong to the G2/M cluster. When K increased from 7 to 8, the major change in gene clustering was that cluster 1 for K=7 was split into two smaller clusters. The clustering for K=8 (Figure 2) seemed to describe the data better than that for K=6 and K=7. We will focus on the clustering for K=8 in the following discussion.

We did model checking by plotting the model residuals and the BLUPs (see supplementary materials). Estimated variance of the residuals is fairly constant across time for each of the clusters. Also there is no obvious pattern across time in the residuals, except clusters 3 and 4. Estimated variance of the BLUPs is also reasonably constant across elements of the random effects for each of the clusters. To further explore how the clusters are located to each other and how tight each cluster is, we calculated the eigenvalues for the observed expression data and used the two eigenvectors corresponding to the largest two eigenvalues to display the observed data and the estimated cluster centers with genes in the same cluster highlighted in the same color. The two vectors explain 64.11% of the total variation in the observed data. Figure 3 shows that the clusters partition the samples well, except that cluster 8 overlaps several others and is relatively loose itself.

[Figure 2 about here.]
[Figure 3 about here.]

4.2 Cluster Both WT Data and SM Data

CLMM was also applied to cluster genes using both WT data and SM data. Figures 4 and 5 show the clustering result when CLMM is fit with eight clusters. To gain more insights into the underlying biology, the estimated profiles are compared between WT yeast and SM yeast for each cluster (Figure 6). For example, in cluster 1, periodicity is maintained

in SM yeast but with a smaller magnitude, which suggests that the mutation may have turned off a repressor for genes in this cluster.

[Figure 4 about here.]
[Figure 5 about here.]
[Figure 6 about here.]

To identify genes whose clustering status is changed by the mutation (that is, 'differentially clustered' genes), we compared the clustering using both WT and SM data and that using WT data only (table 1). This is an empirical approach to identify differentially clustered genes and we would like to further rigorously address this problem in the future. The two clusterings differ mostly in their clusters 1 and 2 and their detailed GO annotation is provided in supplementary materials.

[Table 1 about here.]

5. REMARKS

Both K-means and CORM are useful tools for clustering genes using expression data. K-means makes no assumption about the relationship between expression levels and sample covariates. It is intuitive and has produced reasonable results in applications [6,10]. K-means is especially useful to explore the data when no prior knowledge is available on genes' relationship to covariates. CORM assumes a regression relationship between gene expression and covariate. When the assumption holds, CORM is able to provide more precise clustering and cluster center estimates. Moreover, CORM is capable of naturally handling data with complicated experimental design, for example, longitudinal with replications design, unbalanced time points, and missing data.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grants 1 U01 AI46703, R01 AG014358, and 2 R37 AI29168 to SGS and a fellowship from the Merck Research Laboratories to LXQ. We thank Dr. Linda Breeden and Dr. Pramila Tata for providing the yeast data. We also thank Dr. John Storey and Dr. Venkatraman Seshan for helpful discussion.

REFERENCES

1. Kaufman L and Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. New York: Wiley, 1990.
2. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. PNAS 1998; **95**: 14863–14868.
3. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; **270**: 467–470.
4. Perou CM, Srlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000; **406**: 747–752.
5. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell 1998; **9**: 3273–3297.
6. Tavazoie S, Hughes J, Cho R, Church G. Systematic determination of genetic network architecture. Nature Genetics 1999; **22**: 281–285.
7. Yeung K, Fraley C, Murua A, et al. Model-based clustering and data transformations for gene expression data. Bioinformatics 2001; **17**: 977–987.
8. Ghosh D and Chinnaiyan A. Mixture modelling of gene expression data from microarray experiments. Bioinformatics 2002; **18**: 275–286.
9. Qin LX and Self SG. The clustering of regression models method with application in gene expression data. Biometrics 2006; **62**: 526–533.
10. Cho RJ et al. A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell 1998; **2**: 65–73.
11. Park M, Hastie T, Tibshirani R. Averaged gene expressions for regression. Biostatisitcs doi:10.1093/biostatistics/kxl002 2007.

12. MacQueen J. Some methods for classification and analysis of multivariate observations. In Proc 5th Berkeley Symp Math Stat Probability, 1965; 281–297.
13. Celeux G and Govaert G. A classification em algorithm for clustering and two stochastic versions. Computational Statistics and Data Analysis 1992; **14**: 315–332.
14. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. JRSS-B 1977; **39**: 1–22.
15. Fraley C and Raftery AE. How many clusters? which clustering method? answers via model-based cluster analysis. Computer Journal 1998; **41**: 578–588.
16. Storey JD. A direct approach to false discovery rates. JRSS-B 2002; **64**: 479–498.
17. Zhao, LP, Prentice R, Breeden L. Statistical modeling of large microarray data sets to identify stimulus-response profiles. PNAS 2001; **98**: 5631–5636.

Table 1. Compare the clustering using both WT and SM data and that using WT data.

	Both 1	Both 2	Both 3	Both 4	Both 5	Both 6	Both 7	Both 8
WT 1	**13**	1	3		2			3
WT 2	21	**33**						
WT 3	3		**22**					1
WT 4				**11**				
WT 5		2			**30**	2		
WT 6		1			4	**50**		
WT 7							**17**	
WT 8		4			2	8		**23**

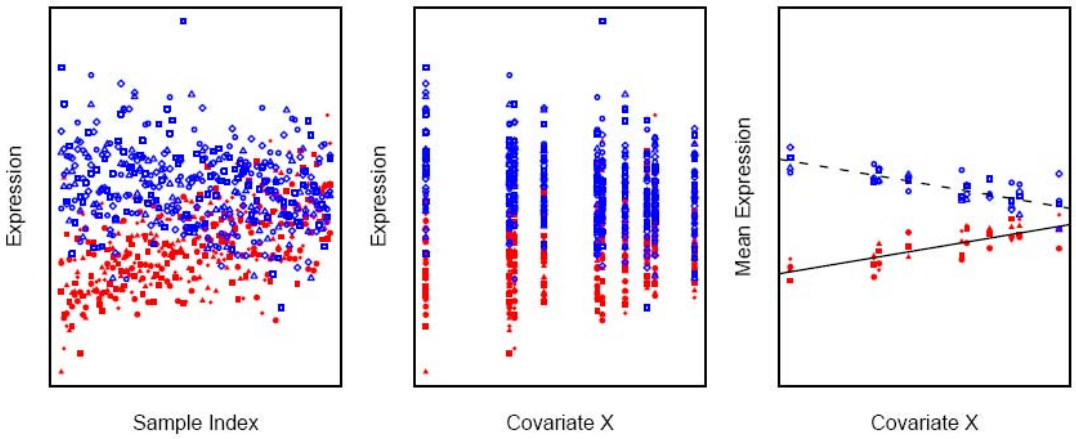

Figure 1. Compare K-means and CLM. Data is simulated for eight genes and 100 samples. Eight genes belong to two clusters. Each of the ten levels of covariate X has ten samples. Left panel plots gene expression *versus* sample index. Middle panel plots gene expression *versus* covariate X. Right panel plots average gene expression for samples at the same X level *versus* covariate X. Symbols represent genes and colors represent gene clusters.

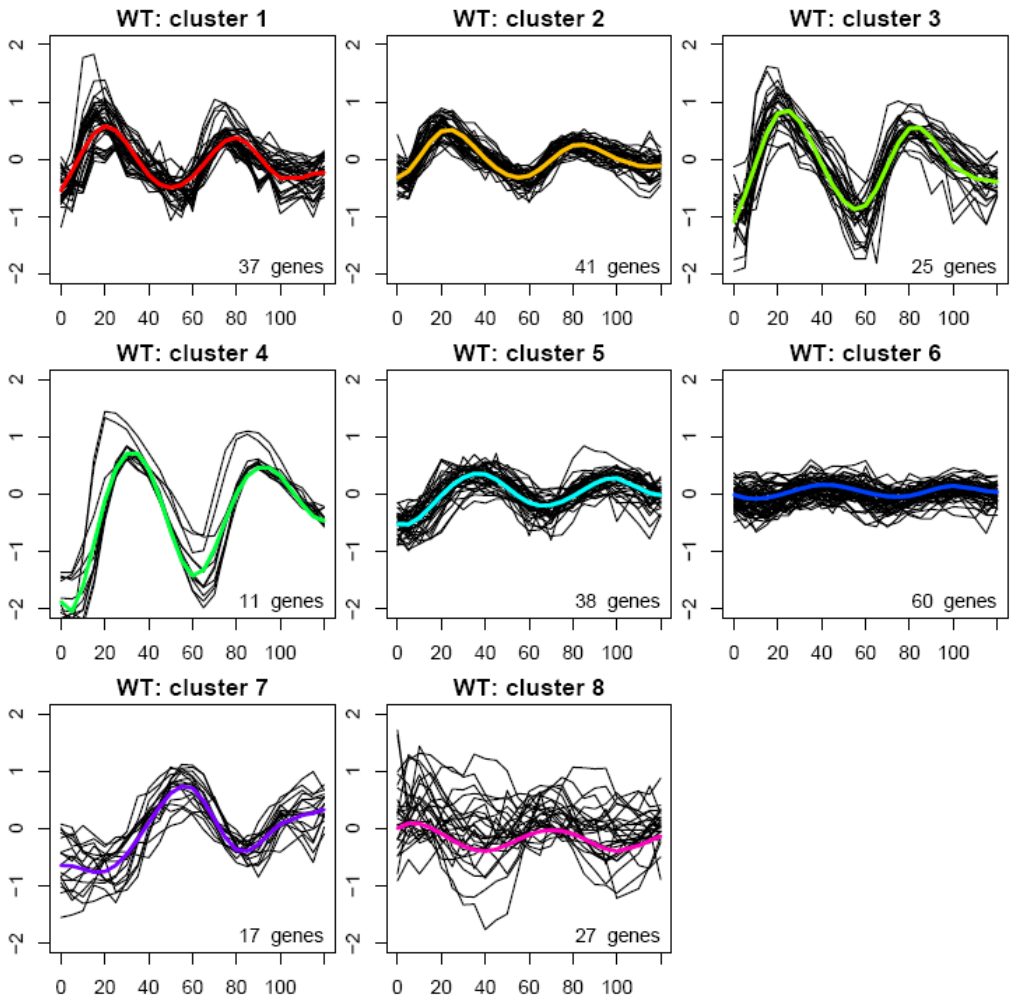

Figure 2. Cluster the 256 genes using WT data. Genes were clustered to eight clusters. Each panel plots the fitted profile (colored line) of one cluster and the observed profiles (black line) of genes in that cluster averaged across the two WT samples *versus* time in minutes. The number of genes in each fitted cluster is labeled at the lower right corner of each panel.

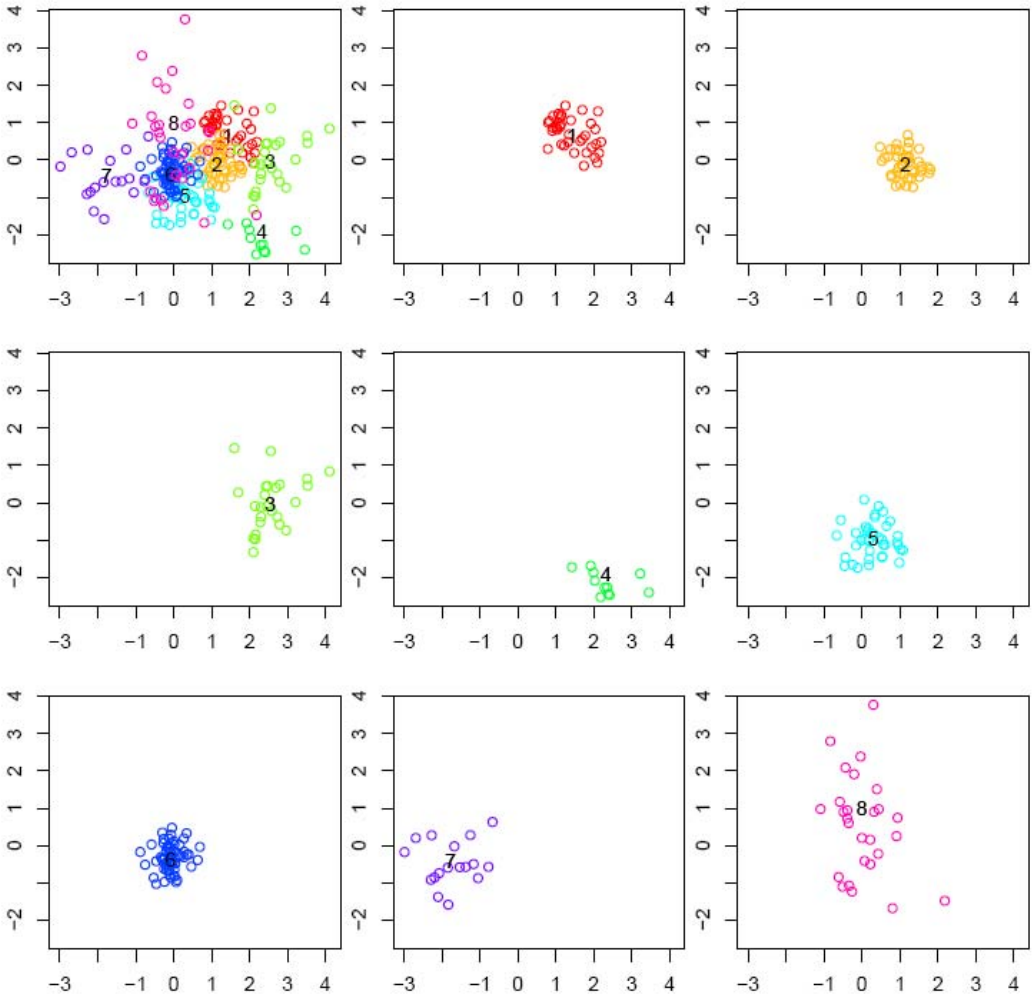

Figure 3. Display the CLMM-based clustering for WT data. X axis and Y axis are the top two eigenvectors. Panel one plots all genes with genes in the same cluster plotted in the same color. Numbers 1-8 indicate the eight cluster centers. The other eight panels plot genes by cluster.

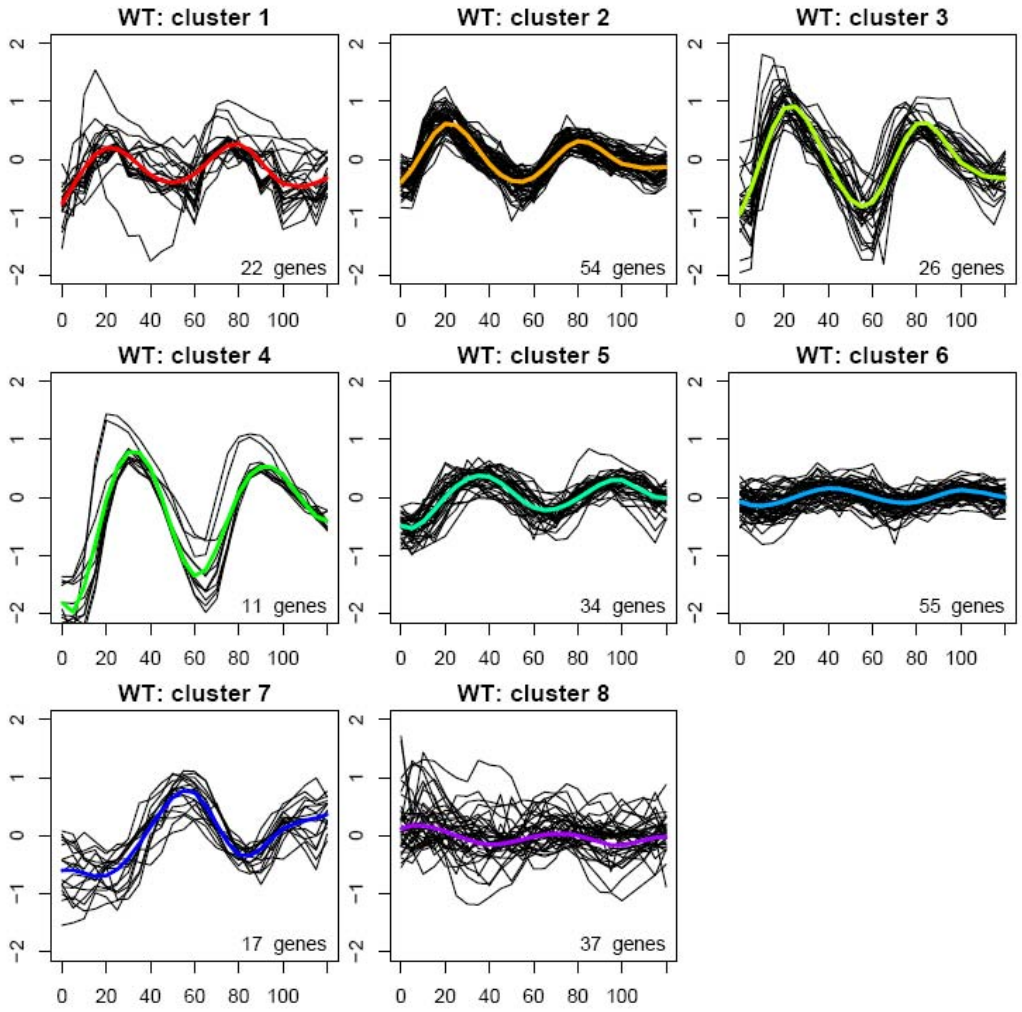

Figure 4. Cluster the 256 genes using both WT and SM data: WT profiles.

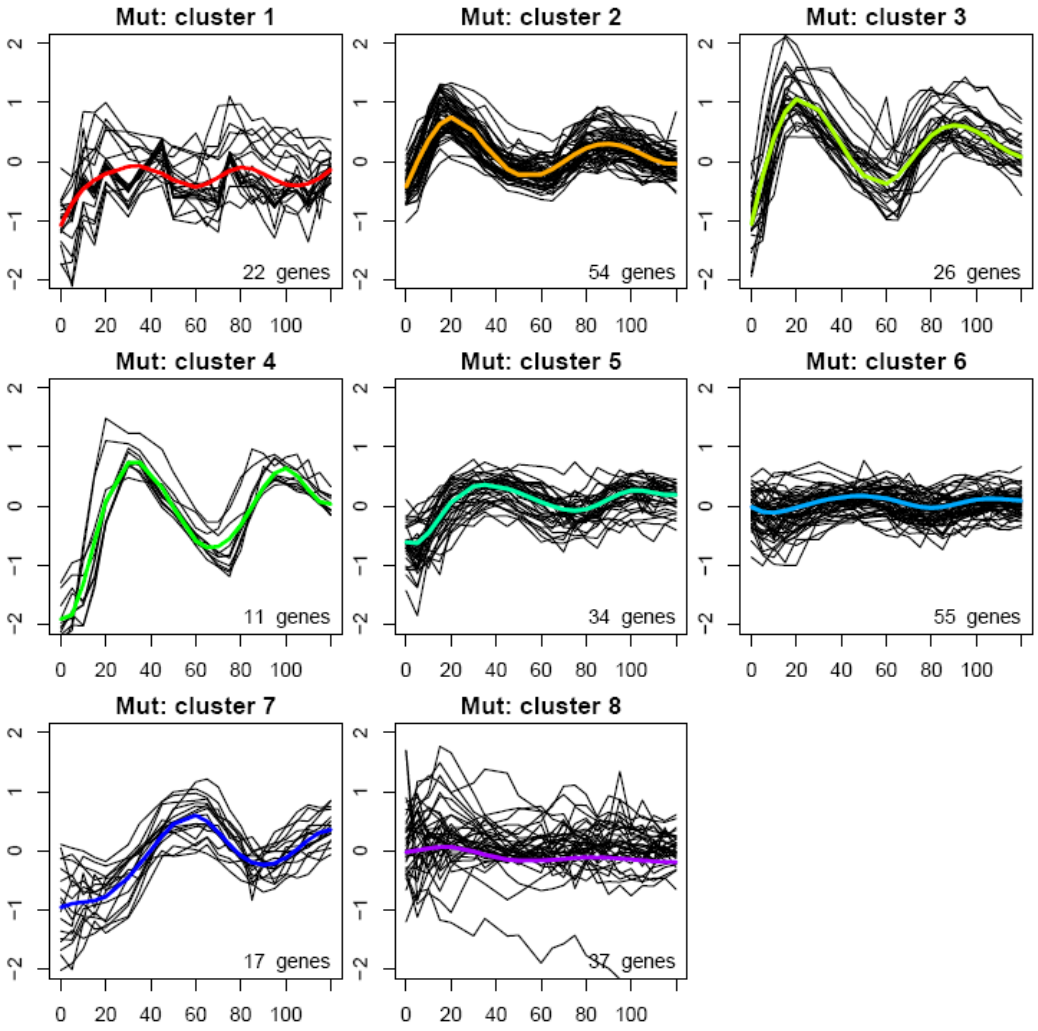

Figure 5. Cluster the 256 genes using both WT and SM data: SM profiles.

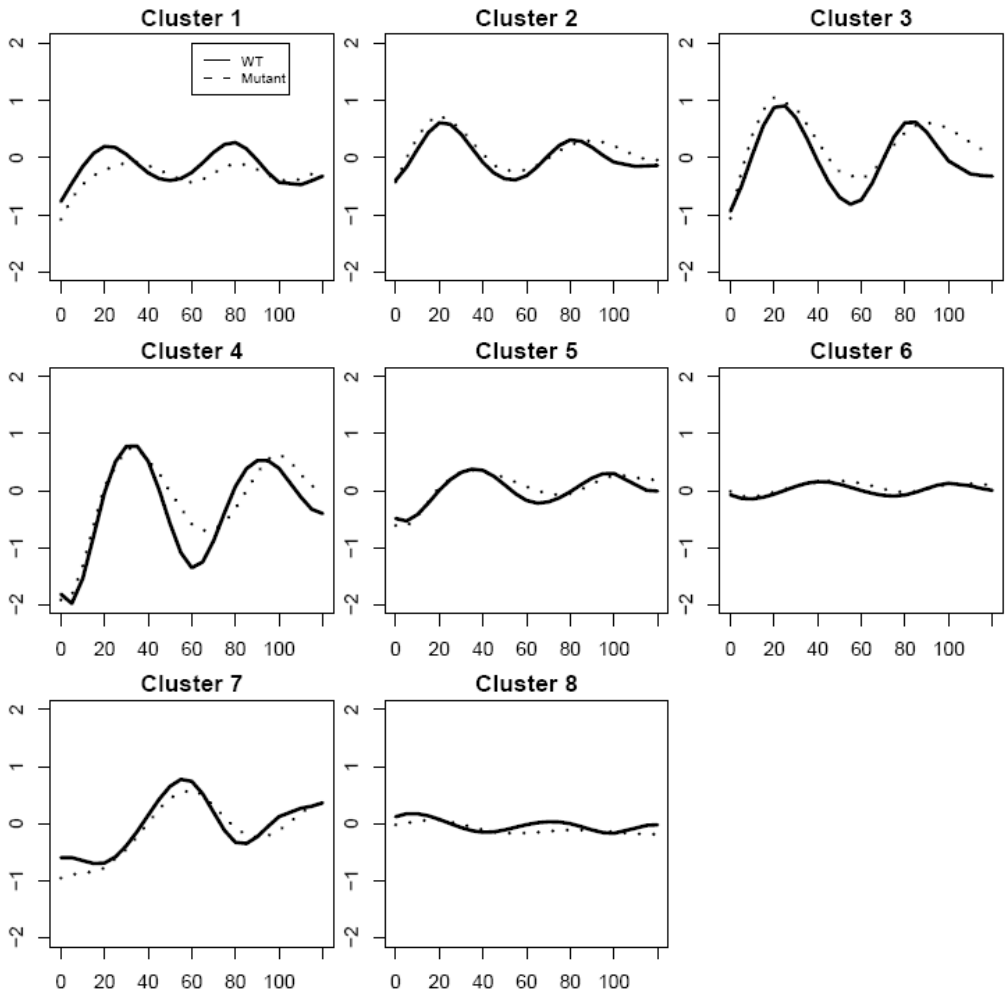

Figure 6. Cluster the 256 genes using both WT and SM data: compare the fitted WT profiles (solid line) and the fitted SM profiles (dotted line).