1-25-2005

The Clustering of Regression Models Method with Applications in Gene Expression Data

Li-Xuan Qin
University of Washington, lqin@u.washington.edu

Steven G. Self
Fred Hutchinson Cancer Research Center, sgs@scharp.org

1. Introduction

Microarray technology provides the ability to measure the expression levels of thousands of genes at once (Schena et al., 1995; Nguyen et al., 2002). For microarray experiments, two common objectives are to detect differentially expressed genes and to cluster co-expressed genes. Differentially expressed genes can serve as disease-specific markers for disease diagnosis in clinical research (Pepe et al., 2003), while co-expressed genes can contribute to our understanding of the regulatory network of gene expression (Eisen et al., 1998).

To answer either question, repeated measurements are needed. In microarray experiments, repeated measurements are often obtained by measuring the expression levels of (1) multiple samples at one time point, or (2) a single sample at multiple time points, or (3) multiple samples each at multiple time points, which we refer to here as type I, II, and III data, respectively. For type I and type III data, samples can be either homogeneous or heterogeneous.

In the past decade, many statistical methods have been proposed for the differential expression question. (Note that "differential expression" was originally used to refer to the comparison of expression levels between *two* conditions, but it is now often used for a more general regression setting, e.g., the comparison among *two or more* conditions or patterns of expression *over time*.) Examples of analytic methods for type I, II and III data include the two-sample t test and its modified versions (Tusher et al., 2001; Baldi and Long, 2001) and ANOVA (Dudoit et al., 2002; Kerr, 2003), the single

pulse model (Zhao et al., 2001), and the linear mixed model with B-spline basis for time (Storey et al., 2004), respectively. Despite the different forms that those per-gene models assume, they all share the common feature of modeling the expression data with a gene-specific regression function, i.e., a mean function dependent on covariates and gene-specific parameters, and differential expression of each gene is assessed by formal inference procedures applied to the gene-specific model parameters. For example,

(a) The two-sample t test assumes a simple linear regression model, for gene g, with covariate $\boldsymbol{x} = (1, x_D)^T$ and regression coefficient $\boldsymbol{\beta}_g = (\beta_{g0}, \beta_{g1})^T$:

$$y_{gi} = \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_g) + \epsilon_{gi} = \beta_{g0} + x_{Di}\beta_{g1} + \epsilon_{gi},$$

where y_{gi} is the expression value for sample i and gene g, x_{Di} is the indicator of disease status for sample i, and ϵ_{gi} is the measurement error.

(b) The linear mixed effects model with spline basis for multiple-sample time series data models the expression value for gene g and sample i at time t_l, y_{gil}, as

$$y_{gil} = \mu(\boldsymbol{x}_{il}; \boldsymbol{\beta}_g) + \epsilon_{gil} = \boldsymbol{x}(t_{il})^T\boldsymbol{\beta}_g + \boldsymbol{x}(t_{il})^T\boldsymbol{b}_{gi} + \epsilon_{gil},$$

where $\boldsymbol{x}_{il} = \boldsymbol{x}(t_{il})$ is a vector of spline basis (Green and Silverman, 1994) evaluated at time t_{il} for sample i, $\boldsymbol{\beta}_g$ is the fixed effects for gene g, $\boldsymbol{b}_{gi}$ is the random effects for gene g and sample i, and ϵ_{gil} is the measurement error.

Motivated by this common regression modeling structure for per-gene models, we propose a new clustering method that employs the same regression model structure so as to provide a natural complementary analysis technique. This method – the clustering of regression models (CORM) method – groups genes that share a similar relationship to the covariate(s). The CORM method provides a unified framework for a family of model-based clustering methods that specifically incorporate experimental design information and other (biologically or clinically) interesting covariates. It can be applied to a wide range of data types.

Previous model-based clustering methods for gene clustering include the multivariate normal mixture model (Yeung et al., 2001) for type I and II data and the clustering of mixed-effects model with B-spline basis method (Luan and Li, 2003) for type II data. The multivariate normal mixture model assumes that the vector of expression values for a gene is distributed as a mixture of multivariate normals and does not incorporate experimental design information, e.g., disease status, time ordering, and (biological) replicates. Luan and Li's method is a special case of the CORM method with the regression model being the mixed-effects model and the number of samples being one. Clustering methods were also proposed in the Bayesian hierarchical model framework (Ramoni et al., 2002; Wakefield et al., 2003; Oh and Raftery, 2003). We will focus on likelihood-based methods in this paper.

The outline of this paper is as follows. In section 2, we describe the CORM method, including the model, the model fitting, and a new method to select the number of clusters. In section 3 the CORM method is applied

to a breast cancer dataset and a yeast cell cycle dataset. Some discussion and remarks are provided in section 4.

2. The Clustering of Regression Models Method

2.1 *The model*

The CORM method uses regression to model systematic variation in gene expression levels as with per-gene methods but, in addition, assumes that gene clusters exist and that genes belonging to the same cluster share the same values of regression coefficients.

We now introduce some notation for type III data, which includes type I and type II data as special cases. Let $\boldsymbol{y}_{gi}$ $(n_{gi} \times 1)$ denote the expression values measured for gene g and sample i, $\boldsymbol{X}_{gi}$ $(n_{gi} \times p)$ the design matrix for gene g and sample i, and $\boldsymbol{\epsilon}_{gi}$ $(n_{gi} \times 1)$ the vector of measurement error. Assume that there are K clusters. Let $\boldsymbol{F}$ be the conditional distribution of $\boldsymbol{y}_g$ with mean $\mu_{\boldsymbol{F}}$ and parameters $\boldsymbol{\xi}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T$, $\mu(.;.)$ the regression function, and $\boldsymbol{\beta}_k$ $(p \times 1)$ the regression coefficient vector for genes in cluster k. Denote the cluster membership for gene g as u_g; or equivalently, write u_g as $\boldsymbol{u}_g = (u_{g1}, \ldots, u_{gK})^T$, of which only the u_gth element equals 1 and the others equal 0. The essential modeling elements underlying the CORM method can be written as

$$\begin{gathered} \boldsymbol{y}_{gi} | (u_g = k, \boldsymbol{X}_{gi}) \sim \boldsymbol{F}_{\boldsymbol{\beta}_k, \boldsymbol{\xi}_k}, \\ \text{with } \mu_{\boldsymbol{F}} = \mu(\boldsymbol{X}_{gi}; \boldsymbol{\beta}_k), \\ \boldsymbol{u}_g \sim Multinomial(\boldsymbol{\pi}), \\ \text{with } \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T, \pi_k \geq 0, \text{ and } \sum_{k=1}^{K} \pi_k = 1. \end{gathered} \tag{1}$$

Complete specification of the CORM modeling framework requires identifi-

cation of the error structure which, in turn, depends on the particular form of the regression model in (1).

The specific form of the conditional distribution function $\boldsymbol{F}$ used for the CORM method is flexible; to name a few, it can be the classical linear model, the linear mixed model, the generalized linear model, and the nonlinear regression model. The choice of the regression model should depend on the nature of the data and the specific scientific question; it should also be informed by the per-gene analysis. For example, if one wants to use type I data to identify groups of genes that have similar expression levels in both cancer samples and normal samples, one can use the clustering of linear models (CLM) method to model the expression data for gene g and sample i as

$$\begin{aligned} y_{gi}|(u_g = k, \boldsymbol{x}_{gi}) &= \boldsymbol{x}_{gi}^T \boldsymbol{\beta}_k + \epsilon_{gi}, \\ \epsilon_{gi} &\sim N(0, \sigma_k^2), \\ \boldsymbol{u}_g &\sim Multinomial(\boldsymbol{\pi}), \end{aligned} \tag{2}$$

where $\boldsymbol{x}_{gi} = (1, x_{Di})^T$ and x_{Di} is the indicator of disease status for sample i. The measurement error variance, σ_k^2, can be either cluster-specific or common to all clusters. If samples are measured repeatedly across time, one may use the linear mixed effects model to model the correlation between measurements on the same sample and to identify genes that have similar expression profiles over time. Specifically, the model for the clustering of linear mixed

models (CLMM) method models type III data as

$$\begin{aligned} \boldsymbol{y}_{gi}|(u_g = k, \boldsymbol{X}_{gi}) &= \boldsymbol{X}_{gi}\boldsymbol{\beta}_k + \boldsymbol{\epsilon}_{gi}, \\ \boldsymbol{\epsilon}_{gi}|(u_g = k) &\sim MVN(0, \boldsymbol{V}_{gi}(\boldsymbol{\xi}_k)), \\ \boldsymbol{V}_{gi}(\boldsymbol{\xi}_k) &= \boldsymbol{Z}_{gi}\boldsymbol{D}_k\boldsymbol{Z}_{gi}^T + \sigma_k^2\boldsymbol{I}, \\ \boldsymbol{u}_g &\sim Multinomial(\boldsymbol{\pi}), \end{aligned} \tag{3}$$

where $\boldsymbol{D}_k$ is the variance matrix for random effects in cluster k and $\boldsymbol{I}$ is an identity matrix. The random effects variance and the measurement error variance can be either cluster-specific or common to all clusters. The design matrix for fixed effects $\boldsymbol{X}_{gi}$ could be the spline basis of time and the design matrix for random effects $\boldsymbol{Z}_{gi}$ is usually a subset of $\boldsymbol{X}_{gi}$. The CLMM model can also be applied for type II data with $i = 1$. To our knowledge the CORM method is the only method available for gene clustering with type III data.

It is worthy to note that one can apply the CORM method to cluster data with multiple outcomes in general; gene expression data is only one of the possible applications.

2.2 *Model fitting*

Because the cluster membership indicator u_g's are not observable, we treat them as missing data and fit the CORM model with the EM algorithm using mixture or classification likelihood (Dempster et al., 1977). Implementation details depend on the specific type of the regression model used. We have implemented the fitting procedures for the CLM method and the CLMM method, which are presented in Appendix A and B, respectively.

To obtain a starting value for the parameters, a regression model is fitted for observations on each gene and the corresponding gene-specific parameters

are clustered either randomly or via an empirical clustering procedure (e.g., K-means). Then set the cluster centers to be $\boldsymbol{\beta}_k^{(0)}$ and the proportion of genes in the corresponding cluster to be $\pi_k^{(0)}$. The model fitting via EM algorithm should be carried out with different starting values to guard against identifying a local maxima as the global maxima. Different starting values can be obtained by re-clustering gene-specific regression parameters (using the same or different clustering methods) or by clustering gene-specific regression parameter estimates based on random subsets of the original sample.

Besides parameter estimates, the fitting procedure also generates an estimate for the cluster membership u_g's. The estimated cluster membership not only offers insight about the underlying structure of the large number of genes, but also can form a classifier for samples or a predictor for future samples.

2.3 *Selecting the number of clusters*

So far we have assumed that the number of clusters K is known; however, this is rarely true. One important aspect of cluster analysis is to empirically determine the number of clusters. Many existing methods for this problem focus on the within cluster dispersion and exploit the so called "elbow" phenomenon (Milligan and Cooper, 1985; Gordon, 1999). New methods for selecting the number of clusters have recently been proposed in the context of sample clustering with gene expression data, including Tibshirani *et al.* (2001) and Dudoit and Fridlyand (2002). These methods base the choice of K on the reproducibility of sample clustering and require the definition of some measure of agreement between two clusterings.

In real data, however, there is no "true" number of clusters, but only a

choice of a useful value of K that results in stable and replicable results that provide a good (e.g., efficient) fit to the data. As such we consider choice of K a matter of empirically informed judgment rather than a matter of formal inference and prefer informal selection based on specified guidelines than a purely automated approach. Two features are critical in guiding our selection of K: efficiency of data description and stability of cluster centers. For the CORM method, cluster centers correspond to the cluster-specific regression coefficient $\boldsymbol{\beta}_k$'s.

To measure the efficiency of data description, we choose to use Bayesian Information Criterion (Schwarz, 1978; Fraley and Raftery, 2002):

$$BIC_K = 2\text{log-likelihood} - p\log(N),$$

where p is the number of covariates in the model and N is the number of observations. Stability of a random variable is often quantified by its variance. However, for each fitted cluster center, the variance is often a *covariance matrix*, say $\boldsymbol{\Sigma}_k$, and there is a covariance matrix for each of the K clusters. To obtain a single numerical summary for stability of cluster centers, we propose a new measure – Bootstrapped Maximum Volume (BMV):

$$BMV_K = \max_{k=1,\ldots,K}\{volume(\hat{\boldsymbol{\Sigma}}_k)\}.$$

The BMV measure first summarizes each covariance matrix with its *volume* (Banfield and Raftery, 1993) and then summarizes the set of volumes for K clusters with their maximum value.

Specifically, the BIC/BMV method is to,

(i) for each candidate K, compute the BIC value;

(ii) for each candidate K, compute the BMV value by

   a. drawing B bootstrap samples and for each bootstrap sample
      - fitting the CORM model;
      - relabeling the estimated coefficient vector for this bootstrap sample to minimize its Euclidean distance to the estimated coefficients for the original data;
   b. computing the covariance matrix for the estimated coefficients obtained in (a) for each of the K clusters;
   c. computing the maximum volume of the K covariance matrices;

(iii) select the number of clusters to be the K value that has a large BIC value and a small BMV value.

3. Data Analysis

We now illustrate the application of the proposed methodology using two real datasets from microarray experiments, one by Zhao *et al.* (2004) and the other by Spellman *et al.* (1998). We will briefly describe the two studies here and the readers are refered to the original publications for more detailed reports.

3.1 *Breast Cancer Data*

Zhao *et al.* (2004) studied the gene expression profiles of two major histological types of breast cancer – invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC). They analyzed the expression profiles of 38 IDC samples and 21 ILC samples, using cDNA arrays spotted for 42,000 clones. A common reference sample is used for all arrays. With SAM analysis (Tusher et al., 2001), they identified a total of 474 clones that were differentially expressed between IDCs and ILCs, representing 354 unique genes. To illustrate our methodology, we applied the BIC/BMV method and the CLM

method to the log ratios of those 354 genes (measurement of multiple clones corresponding to the same gene are averaged).

The BIC/BMV method shows that $K = 9$ and $K = 14$ offer better trade-off between BIC and BMV measures than other candidate values of K (Figure 1). Figure 2 shows the clustering results when the CLM model is fitted with 9 clusters and with 14 clusters. When $K = 9$, genes CRBPIV, FABP4, LPL, and PLIN are grouped in cluster 3, which are involved in Lipid/fatty acid transport and metabolism; genes HIST1H2AL, HIST1H2BD, HIST2H2AA, HIST2H2BE, HIST1H2BK, HIST1H2BL, and HIST3H2A are all involved in nucleosome assembly, but the first four genes are grouped in cluster 9, while the last three in cluster 7. It also shows that all but two clusters fitted for $K = 14$ are roughly subsets of one of the clusters fitted for $K = 9$, which suggests that subgroups are further separated as K increases from 9 to 14 and provides a somewhat hierarchical view of clustering for those genes.

[Figure 1 about here.]

[Figure 2 about here.]

3.2 *Yeast Cell Cycle Data*

Spellman *et al.* (1998) monitored the genome-wide mRNA levels for 6,108 yeast ORFs at 7-min intervals for 119 min, relative to a reference mRNA from an asynchronous yeast culture, in a cell culture synchronized by *cdc15*, *cdc28*, or α factor. These three datasets were analyzed by Zhao *et al.* (2001) using the single pulse model; a total of 256 genes were identified to oscillate significantly in at least two datasets. For illustration purpose, we fit the

CLMM model to the log ratios of those 256 genes synchronized using α factor. Data from the last two time points are excluded from this analysis because the amount of missing data at these two time points implies poor data quality and because these two time points are at the end of the second cell cycle with weak cell cycle signal due to loss of synchrony. The design matrix for fixed effects is the B-spline basis with 7 equally spaced knots and so is the design matrix for random effects. The number of knots is set to be 7 to allow a flexible modeling of the curve and at the same time to avoid overfitting; within a reasonable range, the clustering results are not sensitive to the number of degrees of freedom for the B-spline basis.

Figure 3 shows the clustering result when the CLMM model is fitted with six clusters. Of these 256 genes, 229 genes have been previously classified into five clusters – G1, S, S/G2, G2/M, and M/G1 – by Spellman *et al.* (1998) based on the estimated time to the first peak (http://genome-www.stanford.edu/cellcycle/data/rawdata/). Figure 4 compares the fitted clustering using the CLMM method with Spellman *et al.*'s clustering (labeled as 'expected clustering'). Fitted cluster 1, 2, and 3 seem to divide G1 cluster into three clusters – early, middle, and late G1 clusters, while fitted cluster 5 consists roughly S/G2 cluster and G2/M cluster, suggesting that those two clusters have a similar expression profile. Figure 4 also shows that the fitted clusters seem to be shifting along the cell cycle phases and echoes the fact that the cell cycle is a continuous process.

[Figure 3 about here.]

[Figure 4 about here.]

4. Discussion

The CORM method can be considered as an application of finite mixture models (McLachlan and Basford, 1988). Traditional finite mixture models consider that an observation comes from a mixture of *marginal* distributions. In 1970's, switching regression models (Kiefer, 1978) were studied for the problem of switching regime, where some observations on an outcome variable come from one regression line and other observations come from another. Switching regression models consider that, conditional on the covariate, an observation on the outcome variable comes from a mixture of *conditional* distributions. Now microarray technology offers a wealth of data with multiple outcome variables. The CORM method assumes that (observations on) an outcome variable comes from a mixture of *conditional* distributions. The CORM method groups outcome variables, while the switching regression model groups observations on a single outcome variable. For the CORM method, the availability of outcome-specific regression parameter estimates makes it relatively easy to find starting values for the EM algorithm.

The CORM method can naturally accommodate missing data on any gene or any sample, while algorithmic clustering procedures (e.g., K-means) and the multivariate normal mixture model cannot. Computer programs for the CLM method and the CLMM method have been implemented in R. Alternatively, one can also implement the estimation of the CORM method in a Bayesian framework using standard computer packages, such as BUGS.

The BMV measure can be used for not only the CORM method but also many other model-based clustering methods (together with the BIC measure or some other goodness-of-fit measure) or empirical clustering methods

(together with some cluster-tightness measure, e.g., the within-cluster dispersion). As an example, for K-means clustering, one can use the BMV measure together with the within-cluster sum of squares to choose a K value that has both a small variability and a small within-cluster dispersion.

In this paper, we have assumed that data are properly normalized and transformed (Yang et al., 2002). As with many parametric statistical methods, the choice of normalization and transformation methods can significantly alter the analysis results and need to be chosen carefully (Qin et al., 2004). After identifying differentially expressed genes and clustering co-expressed genes, the natural next step of analyzing microarray data would be to investigate the available functional information on the genes. Libraries of information are available from the public and private domains, among which is the Gene Ontology Consortium (http://www.geneontology.org/). By examining the functions of co-clustered genes, one can obtain biological insights into the pathways, functions of unknown genes, and pathogenesis of diseases.

In summary, the CORM method provides a flexible tool to cluster high-dimensional data and can be applied to a wide range of experimental designs and scientific questions. When used in combination with regression-based per-gene analyses of differential expression, it forms the basis for an integrated analytic framework for the analysis of microarray data. We expect this framework to be increasingly useful as microarray measurements are obtained over a broader and more complex set of experimental conditions.

Acknowledgements

This work was supported in part by grants 1 U01 AI46703 and 2 R37 AI29168

to SGS and a fellowship from the Merck Research Laboratories to LXQ. We thank Hongjuan Zhao and Stefanie Jeffrey at Stanford University for help with the breast cancer data.

References

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inference of gene changes. *Bioinformatics*, 17(6):509–519.

Banfield, J. and Raftery, A. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological*, 39:1–22.

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7):0036.1–0036.21.

Dudoit, S., Yang, Y., Callow, M., and Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111–139.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.

Gordon, A. D. (1999). *Classification*. Chapman & Hall Ltd.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall Ltd.

Kerr, M. K. (2003). Linear models for microarray data analysis: hidden similarities and differences. *Journal of Computational Biology*, 10:891–901.

Kiefer, N. M. (1978). Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica*, 46(2):427–434.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics*, 19(4):474–482.

McLachlan, G. and Basford, K. (1988). *Mixture models: inference and applications to clustering*. Marcel Dekker Inc.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179.

Nguyen, D. V., Arpat, A. B., Wang, N., and Carroll, R. J. (2002). Dna microarray experiments: biological and technological aspects. *Biometrics*, pages 701–717.

Oh, M.-S. and Raftery, A. (2003). Model-based clustering with dissimilarities: a Bayesian approach. Technical report, Department of Statistics, University of Washington, Seattle, WA.

Pepe, M., Longton, G., Anderson, G., and Schummer, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics*,

59:133–142.

Qin, L.-X., Kerr, K. F., and Contributing Members of The Toxicogenomics Research Consortium (2004). Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Research*, 32(18):5471–5479.

Ramoni, M., Sebastiani, P., and S., K. I. (2002). Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. USA*, 99:9121–9126.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297.

Storey, J. D., Leek, J. T., Xiao, W., Dai, J. Y., and Davis, R. W. (2004). A significance method for time course microarray experiments applied to two human studies. Technical report, Department of Biostatistics, University of Washington, Seattle, WA.

Tibshirani, R., Walther, G., Botstein, D., and Brown, P. (2001). Cluster validation by prediction strength. Technical report, Department of Statistics, Stanford University, Stanford, CA.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of

microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98(9):5116–5121.

Wakefield, J., Zhou, C., and Self, S. (2003). Modelling gene expression over time: curve clustering with informative prior distributions. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Statistics 7, Proceedings of the Seventh Valencia International Meeting*. Oxford University Press.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15.

Yeung, K., Fraley, C., Murua, A., Raftery, A., and Ruzzo, W. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987.

Zhao, H., Langerod, A., Ji, Y., Nowels, K., Nesland, J., Tibshirani, R., Bukholm, I., Karesen, R., Botstein, D., Borresen-Dale, A.-L., and Jeffrey, S. (2004). Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Molecular Biology of the Cell*, 15:2523–2536.

Zhao, L. P., Prentice, R., and Breeden, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl. Acad. Sci. USA*, 98(10):5631–5636.

Appendix A

Model fitting for the CLM method

For notational convenience, we drop the subsript for n_{gi} and m_g; the extension is straightforward. Let $\boldsymbol{y}_g = (\boldsymbol{y}_{g1}^T, \ldots, \boldsymbol{y}_{gm}^T)^T$, $\boldsymbol{Y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_G^T)^T$, $\boldsymbol{X}_g = (\boldsymbol{X}_{g1}^T, \ldots, \boldsymbol{X}_{gm}^T)^T$, $\boldsymbol{X} = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_G^T)^T$, $\boldsymbol{u} = (u_1, \ldots, u_G)^T$, and $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T, \boldsymbol{\xi}^T)^T$, where $\boldsymbol{\theta}$ represents all parameters involved in the component distributions and the definition of $\boldsymbol{\xi}$ depends on the form of $\mu()$.

To fit the CLM model, we need to estimate $\boldsymbol{\pi}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma_1^2, \ldots, \sigma_K^2)$. By treating $\boldsymbol{u}_g$'s as missing data, we can use the EM algorithm to find the MLE iteratively. The expected value of $\boldsymbol{u}_g$'s, i.e., the probability for one gene belonging to each cluster, is first calculated given the current parameter estimates (E-step); then the parameters are estimated given the current expected values of $\boldsymbol{u}_g$'s (M-step). We implemented the model fitting procedure with the mixture likelihood; alternatively, one can fit the CLM model with the classification likelihood with a slight modification to the E-step of the proposed algorithm.

The complete data likelihood factors into two distinct parts, each involving a separate subset of parameters. Specifically,

$$\begin{aligned} f(\boldsymbol{Y}, \boldsymbol{u}|\boldsymbol{X}; \boldsymbol{\theta}, \boldsymbol{\pi}) &= f(\boldsymbol{u}|\boldsymbol{\pi}) f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{u}; \boldsymbol{\theta}) \\ &= \left[\prod_{g=1}^{G} \prod_{k=1}^{K} \pi_k^{u_{gk}} \right] \left[\prod_{g=1}^{G} \prod_{k=1}^{K} \left\{ f_k(\boldsymbol{y}_g|\boldsymbol{X}_g; \boldsymbol{\theta}) \right\}^{u_{gk}} \right]. \end{aligned}$$

Up to additive constants, the complete data likelihood can be written as

$$
\begin{aligned}
l(\boldsymbol{\theta}, \boldsymbol{\pi}; \boldsymbol{Y}, \boldsymbol{u} | \boldsymbol{X}) &= l_1(\boldsymbol{\pi}; \boldsymbol{u}) + l_2(\boldsymbol{\theta}; \boldsymbol{Y} | \boldsymbol{u}, \boldsymbol{X}), \\
l_1 &= \sum_{g=1}^{G} \sum_{k=1}^{K} u_{gk} \log(\pi_k), \\
l_2 &= -\frac{1}{2} \sum_{g=1}^{G} \sum_{k=1}^{K} \sum_{i=1}^{m} u_{gk} \left\{ \log(\sigma_k^2) + \frac{(y_{gi} - \boldsymbol{x}_{gi}^T \boldsymbol{\beta}_k)^2}{\sigma_k^2} \right\}.
\end{aligned}
$$

The E-step finds the expectation of the log likelihood function of $\boldsymbol{\theta}$ based on the complete data conditional on the observed data and $\boldsymbol{\theta}^{(t)}$ from the previous iteration, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. For the CLM method, this amounts to calculating $E(u_{gk} = 1|\boldsymbol{y}_g, \boldsymbol{\theta}^{(t)})$ since $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is linear in u_{gk}.

$$
\hat{u}_{gk} = E(u_{gk} = 1|\boldsymbol{y}_g) = \frac{\pi_k^{(t)} Pr(\boldsymbol{y}_g | u_{gk} = 1; \boldsymbol{\theta}^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} Pr(\boldsymbol{y}_g | u_{gk} = 1, \boldsymbol{\theta}^{(t)})},
$$

where $Pr(\boldsymbol{y}_g | u_{gk} = 1, \boldsymbol{\theta}^{(t)}) = \prod_{i=1}^{m} f(y_{gi} | u_{gk} = 1; \boldsymbol{\theta}^{(t)}) = \prod_{i=1}^{m} \phi(\frac{y_{gi} - \boldsymbol{x}_{gi}^T \boldsymbol{\beta}_k^{(t)}}{\sigma_k^{(t)}})$. $\phi(.)$ denotes the density function for the standard normal distribution.

The M-step updates $\boldsymbol{\theta}$ with the value $\boldsymbol{\theta}^{(t+1)}$ that maximizes$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. For the CLM method, the specific calculation of the M-step is

$$
\begin{aligned}
\hat{\pi}_k^{(t+1)} &= \sum_{g=1}^{G} \hat{u}_{gk} \\
\hat{\boldsymbol{\beta}}_k^{(t+1)} &= \Big(\sum_{g=1}^{G} \hat{u}_{gk} \boldsymbol{X}_g^T \boldsymbol{X}_g \Big)^{-1} \sum_{g=1}^{G} \hat{u}_{gk} \boldsymbol{X}_g^T \boldsymbol{y}_g \\
\hat{\sigma}_k^{2(t+1)} &= \Big\{ \sum_{g=1}^{G} \hat{u}_{gk} (\boldsymbol{y}_g - \boldsymbol{X}_g^T \boldsymbol{\beta}_k^{(t+1)})^T (\boldsymbol{y}_g - \boldsymbol{X}_g^T \boldsymbol{\beta}_k^{(t+1)}) \Big\} / \Big\{ \sum_{g=1}^{G} m \hat{u}_{gk} \Big\}
\end{aligned}
$$

Appendix B

Model fitting for the CLMM method

To fit the CLMM model, we need to estimate $\boldsymbol{\pi}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{D}_1, \ldots, \boldsymbol{D}_K, \sigma_1^2, \ldots, \sigma_K^2)$. By treating both u_g's and $\boldsymbol{b}_{gi}$'s as missing data, we can use the EM algorithm to find the MLE iteratively. Cluster membership indicators and corresponding random effects are estimated given the current parameter estimates (E-step); then the parameters are estimated given the cluster assignment and the expected values of random effects from the previous E-step (M-step). The complete data likelihood factors into three distinct parts, each involving a separate subset of parameters.

$$\begin{aligned} l(\boldsymbol{\theta}, \boldsymbol{\pi}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_G, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_G, \boldsymbol{u}) = & \, l_1(\boldsymbol{\pi}; \boldsymbol{u}) \\ & + l_2(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_K; \boldsymbol{b}_1, \ldots, \boldsymbol{b}_G | \boldsymbol{u}) \\ & + l_3(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma_1^2, \ldots, \sigma_K^2; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_G | \boldsymbol{b}_1, \ldots, \boldsymbol{b}_G, \boldsymbol{u}), \end{aligned}$$

where $\boldsymbol{b}_g = (\boldsymbol{b}_{g1}^T, \ldots, \boldsymbol{b}_{gm}^T)^T$.

Furthermore, the three components of the complete data likelihood can, up to additive constants, be written as

$$\begin{aligned} l_1 &= \sum_{g=1}^{G} \sum_{k=1}^{K} u_{gk} \log(\pi_k), \\ l_2 &= -\frac{1}{2} \sum_{g=1}^{G} \sum_{k=1}^{K} \sum_{i=1}^{m} u_{gk} \{\log(|\boldsymbol{D}_k|) + \boldsymbol{b}_{gi}^T \boldsymbol{D}_k^{-1} \boldsymbol{b}_{gi}\}, \\ l_3 &= -\frac{1}{2} \sum_{g=1}^{G} \sum_{k=1}^{K} \sum_{i=1}^{m} u_{gk} \{\log(|\sigma_k^2 \boldsymbol{I}|) + \sigma_k^2 \boldsymbol{r}_{gik}^T \boldsymbol{r}_{gik}\}, \end{aligned}$$

where $\boldsymbol{r}_{gik} = y_{gi} - \boldsymbol{X}_{gi} \boldsymbol{\beta}_k - \boldsymbol{Z}_{gi} \boldsymbol{b}_{gi}$.

The E-step finds $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. For the CLMM method, this is to calculate

$$\begin{aligned}
\hat{u}_{gk} &= E(u_{gk}=1|\boldsymbol{y}_g;\boldsymbol{\theta}^{(t)}) = \frac{\pi_k^{(t)} Pr(\boldsymbol{y}_g|u_{gk}=1,\boldsymbol{\theta}^{(t)})}{\sum_{k=1}^{K}\pi_k^{(t)} Pr(\boldsymbol{y}_g|u_{gk}=1,\boldsymbol{\theta}^{(t)})},\\
\hat{\boldsymbol{b}}_{gik} &= E(\boldsymbol{b}_{gi}|\boldsymbol{y}_g,u_{gk}=1;\boldsymbol{\theta}^{(t)}) = \boldsymbol{D}_k^{(t)}\boldsymbol{Z}_{gi}^T\boldsymbol{V}_{gik}^{-1(t)}(\boldsymbol{y}_{gi}-\boldsymbol{X}_{gi}\boldsymbol{\beta}_k^{(t)}),\\
\hat{\boldsymbol{b2}}_{gk} &= E\Big(\sum_{i=1}^{m}\boldsymbol{b}_{gi}\boldsymbol{b}_{gi}^T|\boldsymbol{y}_g,u_{gk}=1;\boldsymbol{\theta}^{(t)}\Big)\\
&= \sum_{i=1}^{m}\hat{\boldsymbol{b}}_{gik}\hat{\boldsymbol{b}}_{gik}^T + \boldsymbol{D}_k^{(t)} - \boldsymbol{D}_k^{(t)}\Big(\sum_{i=1}^{m}\boldsymbol{Z}_{gi}^T\boldsymbol{V}_{gik}^{-1(t)}\boldsymbol{Z}_{gi}\Big)\boldsymbol{D}_k^{(t)},\\
\hat{\boldsymbol{e}}_{gik} &= E(\boldsymbol{e}_{gi}|\boldsymbol{y}_g,u_{gk}=1;\boldsymbol{\theta}^{(t)}) = \boldsymbol{y}_{gi}-\boldsymbol{X}_{gi}\boldsymbol{\beta}_k^{(t)}-\boldsymbol{Z}_{gi}\hat{\boldsymbol{b}}_{gik},\\
\hat{e2}_{gk} &= E\Big(\sum_{i=1}^{m}\boldsymbol{e}_{gi}^T\boldsymbol{e}_{gi}|\boldsymbol{y}_g,u_{gk}=1;\boldsymbol{\theta}^{(t)}\Big)\\
&= \sum_{i=1}^{m}\big\{\hat{\boldsymbol{e}}_{gik}^T\hat{\boldsymbol{e}}_{gik} + \sigma_k^{2(t)}tr(\boldsymbol{I}-\sigma_k^{2(t)}\boldsymbol{V}_{gik}^{-1(t)})\big\}.
\end{aligned}$$

The M-step updates the parameters as

$$\begin{aligned}
\hat{\pi}_k^{(t+1)} &= \sum_{g=1}^{G}\hat{u}_{gk},\\
\hat{\boldsymbol{D}}_k^{(t+1)} &= \frac{\sum_{g=1}^{G}\hat{u}_{gk}\hat{\boldsymbol{b2}}_{gk}}{m\sum_{g=1}^{G}\hat{u}_{gk}},\\
\hat{\sigma}_k^{2(t+1)} &= \frac{\sum_{g=1}^{G}\hat{u}_{gk}\hat{e2}_{gk}}{nm\sum_{g=1}^{G}\hat{u}_{gk}},\\
\hat{\boldsymbol{\beta}}_k^{(t+1)} &= \Big(\sum_{g=1}^{G}\sum_{i=1}^{m}\hat{u}_{gk}\boldsymbol{X}_{gi}^T\boldsymbol{X}_{gi}\Big)^{-1}\sum_{g=1}^{G}\sum_{i=1}^{m}\hat{u}_{gk}\boldsymbol{X}_{gi}^T(\boldsymbol{y}_g-\boldsymbol{Z}_{gi}\hat{\boldsymbol{b}}_{gik}).
\end{aligned}$$

For both the CLM method and the CLMM method, convergence is considered to be achieved when the increase of the log-likelihood from one iteration to the next is less than 0.01.

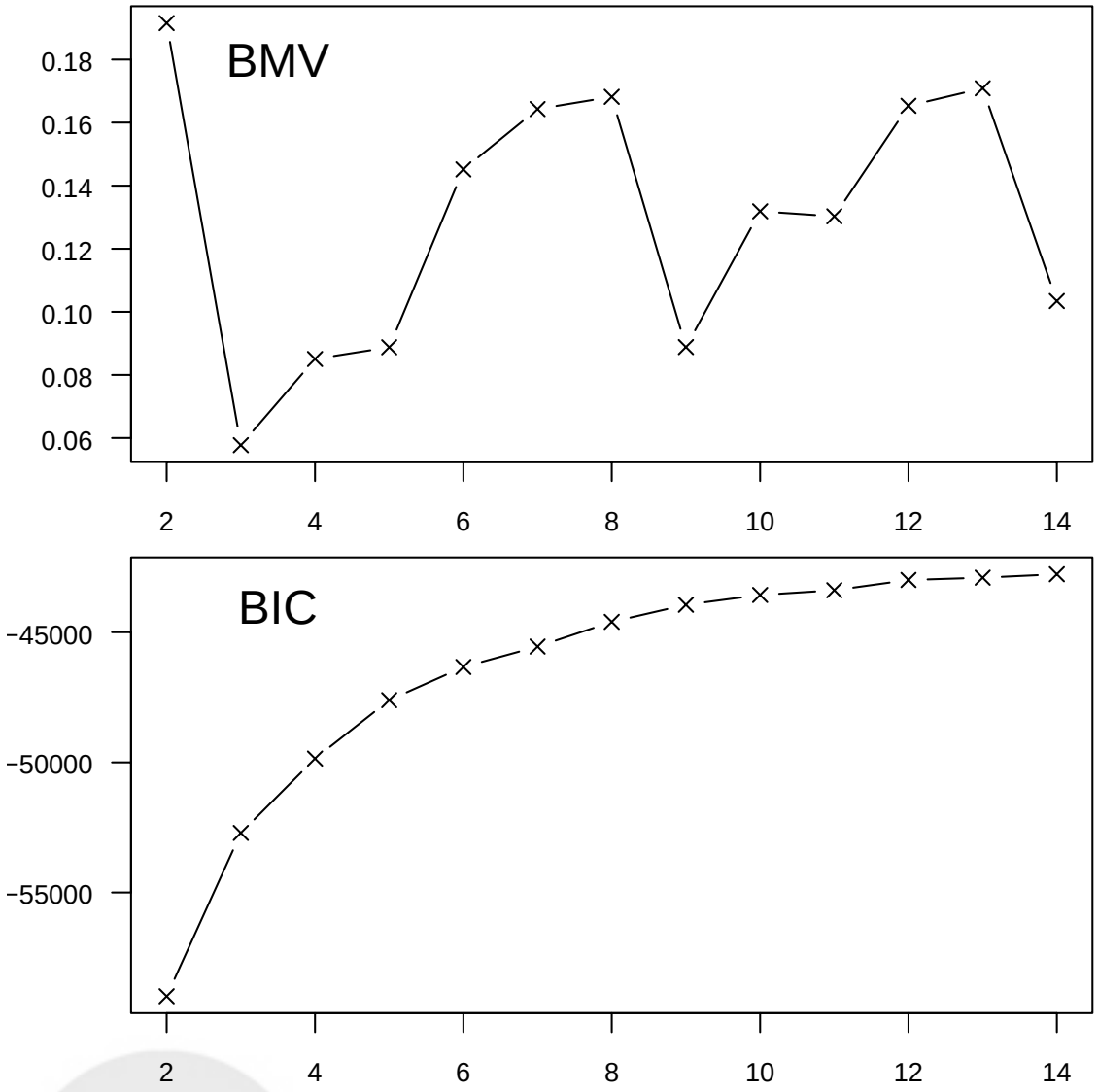

Figure 1. BMV and BIC values vs. candidate values for K.

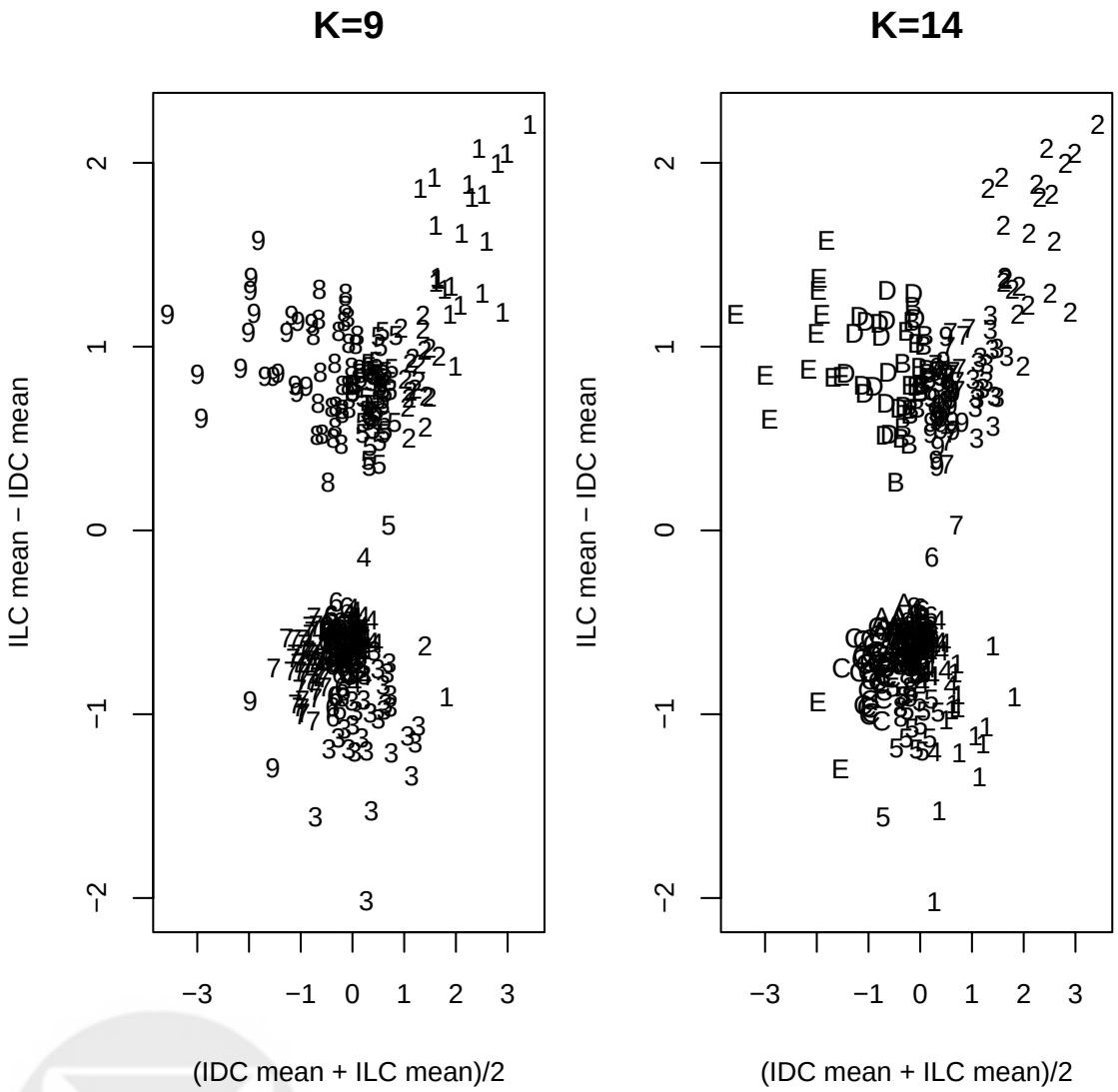

Figure 2. Scatter plots for the difference of gene-specific means among IDC samples and that among ILC samples *versus* their average. Fitted clusters for $K = 9$ are labeled with numbers $1 - 9$ and fitted clusters for $K = 14$ are labeled with $1 - 9$ and A−E.

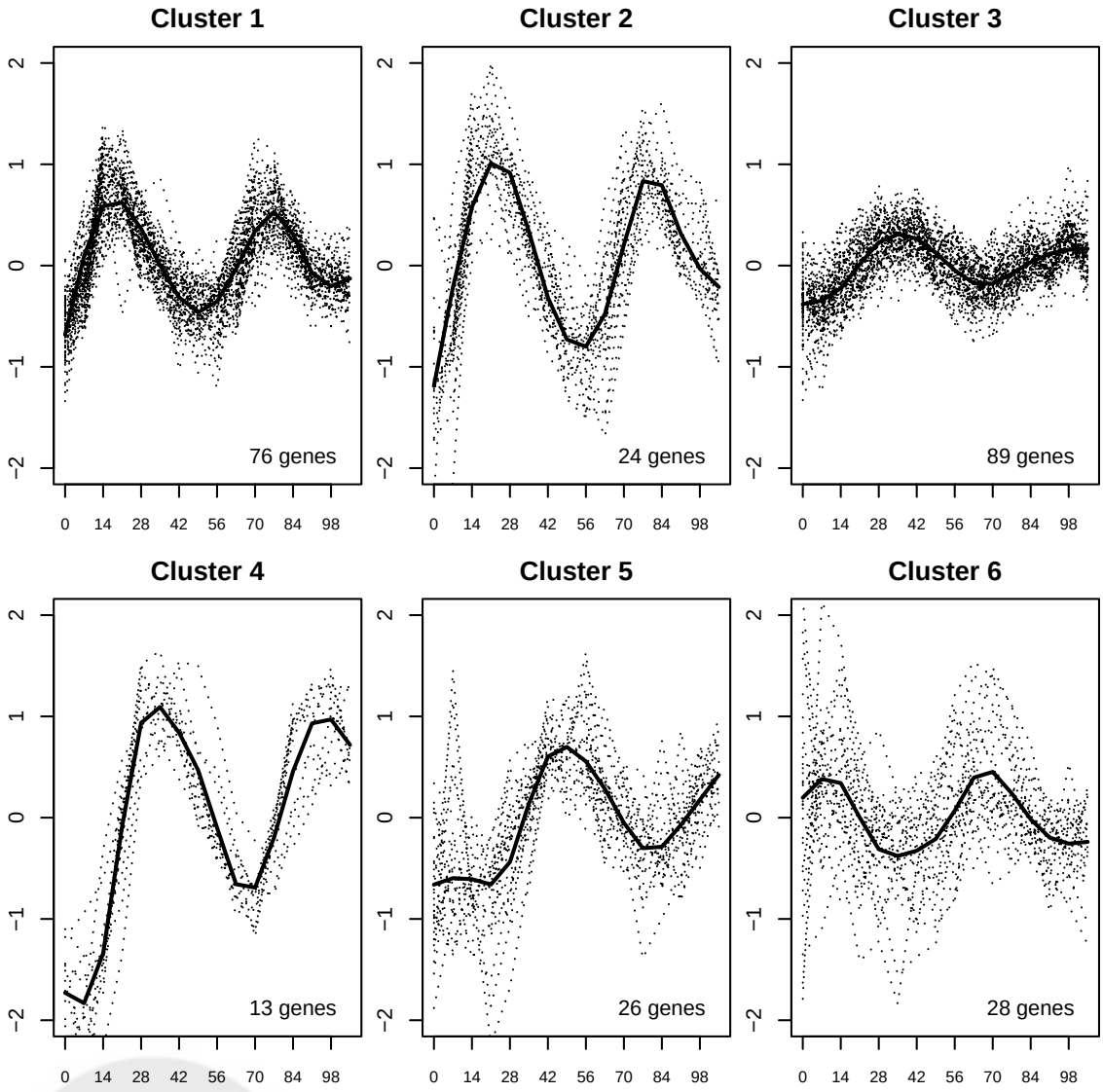

Figure 3. Fitted and observed expression profiles over time for the 256 genes from Spellman *et al.* (1998). Each of the six panels plots the fitted profile (solid line) of one cluster and the observed profiles (dotted line) of genes in that cluster *versus* time in minutes. The number of genes in each fitted cluster is labeled in the bottom right corner of each panel. Clusters are ordered by the estimated time to the first peak.

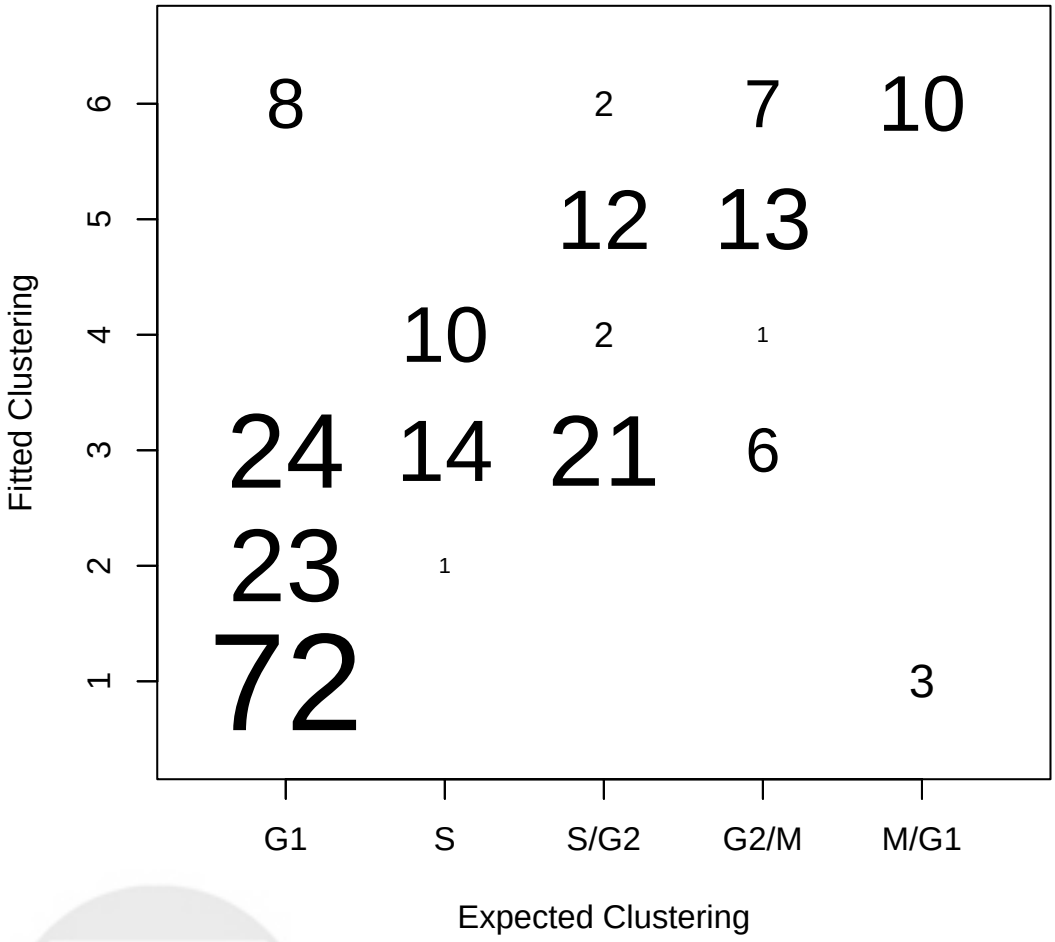

Figure 4. Compare the fitted clustering based on the CLMM model with Spellman *et al.*'s clustering for the 256 genes. The six fitted clusters seem to be shifting from cell cycle phase M/G1 to phases G1, S, S/G2, and G2/M.