Journal of the Indian Statistical Association
Vol. 50 No.1 & 2, 2012, 143-150.

A growth curve approach to analyzing multiple-valued expression data

Arvind K. Jammalamadaka

Massachusetts Institute of Technology, USA

and

S. Rao Jammalamadaka

University of California, Santa Barbara, USA

Abstract

There is significant literature which explores methods for clustering time-series gene-expression data sets, such as the classical data set due to Spellman et al. (1998). For instance James and Hastie (2001) use linear or quadratic discriminant functions on fitted curves, while Bar-Joseph et al. (2003) using a similar approach, do the clustering based on the coefficients of the fitted splines. In a series of papers, Liu et al. (2006), medvedovic et al. (2004) and Medvedovic and Sivaganesan (2002) present methods of clustering gene-profiles, by treating them as multivariate vectors. In this work we take a very different approach. Our goal is not exploratory as when one does clustering, but confirmatory viz. to verify if the mean profiles of the obtained clusters are significantly different. We treat the observed vector on each gene as multivariate Gaussian, and fit a mean curve to each group, based on the "growth curve" analysis. This approach coming from linear models for multivariate data, allows us to do proper statistical significance tests for checking if a mean profile fits to the data, and if these profiles differ for the different groups.

1 Introduction

Experiments where multiple measurements are obtained on the same unit are referred to as repeated-measures models. A typical situation is one where the same child is observed at different time-points to note the growth, say in weight or height. Such measurements are clearly correlated, so that the observation vectors at say p different time-points can be considered as coming from a p-variate Gaussian. The

data might be from different groups corresponding to different treatment conditions and the goal is to compare the mean growth curves for these different groups. In what follows, we treat the gene expression data over the different time-points of a time-series data set as being multivariate Gaussian, to which we fit growth curves of appropriate degree. We perform statistical tests to verify if the different groups are significantly different in their mean profiles. In order to make the paper self-contained as well as to explain this rather non-standard topic and notations, we provide a somewhat detailed description of growth curve models in the next section, before applying it to a real data set in the final section.

2 Growth curve modeling

Although linear models are a classical topic, extensions to the multivariate case including the multivariate analysis of variance and in particular the idea of "growth curves" are a somewhat specialized topic. In view of this and to make the treatment self-contained, we give a brief review of the basic ideas. Further details can be found in books by Kshirsagar and Smith (1995) and Pan and Fang (2002).

Suppose that there are r different groups and y denotes the real valued (growth) variable measured at p different time points: $t_1, t_2, ..., t_p$ for n_j individuals chosen at random from the j^{th} group, $(j = 1, \ldots, r)$. We specify the following polynomial regression model of degree $(q-1)$ for the expression values y on the time variable t,

$$\begin{aligned} E(y_t) &= \psi_{j0}t^0 + \psi_{j1}t^1 + ... + \psi_{jq-1}t^{q-1}; \\ &\quad (t = t_1, ..., t_p; \quad p > q-1; \quad j = 1, 2, ..., r). \end{aligned} \tag{2.1}$$

Let

$$N = n_1 + n_2 + ... + n_r$$

denote the total number of observations in all the groups put together. Let

$$\psi_j' = [\psi_{j0}\psi_{j1}...\psi_{jq-1}] \tag{2.2}$$

denote the vector of the curve coefficients for the j^{th} group. Since the observations $y_{t_1}, ..., y_{t_p}$ are on the same item and hence correlated, we denote their variance-covariance matrix by $\mathbf{\Sigma}$. For simplicity and convenience, we assume $\mathbf{\Sigma}$ to be the same for all the r groups.

Let $\mathbf{Y}_j$ denote the $p \times n_j$ matrix of the observations for the j^{th} group, with each column of dimension p representing one gene. Let the $p \times N$ matrix

$$\mathbf{Y} = [\mathbf{Y}_1 \ \mathbf{Y}_2 \ ... \ \mathbf{Y}_r]. \tag{2.3}$$

denote the combined set of observations in all the r groups. Then from (2.1) we get,

$$\begin{aligned} E(\mathbf{Y}_j) &= [\mathbf{B}\psi_j \mathbf{B}\psi_j ... \mathbf{B}\psi_j] \\ &= \mathbf{B}\psi_j \mathbf{E}_{1n_j} \quad (j = 1, 2, ..., r), \end{aligned} \tag{2.4}$$

where

$$\mathbf{B} = \begin{bmatrix} t_1^0 & t_1^1 & ... & t_1^{q-1} \\ t_2^0 & t_2^1 & ... & t_2^{q-1} \\ ... & ... & ... & ... \\ t_p^0 & t_p^1 & ... & t_p^{q-1} \end{bmatrix} \tag{2.5}$$

and $\mathbf{E}_{ab}$ denotes, a matrix of order $a \times b$ with all elements equal to 1. $\mathbf{B}_{p \times q}$ is called the "design matrix," whose elements depend on the basis we use to represent the mean function. Let

$$\mathbf{A}_{r \times N} = \text{diag}[\mathbf{E}_{1n_1}, \mathbf{E}_{1n_2}, ..., \mathbf{E}_{1n_r}], \tag{2.6}$$

a block diagonal matrix with $\mathbf{E}_{1n_j} (j = 1, 2, ..., r)$ along the diagonal blocks and zeros elsewhere. From Equation (2.4), we get

$$\begin{aligned} E(\mathbf{Y}) &= [\mathbf{B}\psi_1 \mathbf{E}_{1n_1} | \mathbf{B}\psi_2 \mathbf{E}_{1n_2} | ... | \mathbf{B}\psi_r \mathbf{E}_{1n_r}] \\ &= \mathbf{B}\mathbf{\Psi}\mathbf{A}, \end{aligned} \tag{2.7}$$

where

$$\mathbf{\Psi} = [\psi_1 \ ... \ \psi_r] \tag{2.8}$$

is the $q \times r$ matrix of the curve-coefficients.

Let Vec$\mathbf{Y}$, be defined as the column vector obtained by stacking the columns of $\mathbf{Y}$ one below the other. Denoting Var(Vec$\mathbf{Y}$) by Var($\mathbf{Y}$) we see that,

$$\text{Var}(\mathbf{Y}) = \mathbf{I}_N \otimes \mathbf{\Sigma}, \tag{2.9}$$

where $\otimes$ denotes the Kronecker product of two matrices. Equation (2.7) together with Equation (2.9) is called the "growth curve model," introduced by Patthoff and Roy (1964) and later analyzed by Khatri (1966), among others. See also Rao (1973) (Section 8c) and Kshirsagar and Smith (1995) for exposition.

2.1 Fitting growth curves and testing

To fit the growth curve model and to test various hypotheses of interest, we need to perform the following computations (see Kshirsagar and Smith (1995), Chapter 2 for details). Obtain a matrix $\mathbf{B}_2$ of order $p \times (p-q)$ such that

$$\mathbf{B}_2'\mathbf{B} = 0, \tag{2.10}$$

where $\mathbf{B}$ is as in Equation (2.5). This is accomplished for instance by choosing $(p-q)$ linearly independent columns of the matrix $(\mathbf{I}_p - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}')$. Next we compute,

$$\mathbf{S} = \mathbf{Y}(\mathbf{I} - \mathbf{A}'(\mathbf{A}\mathbf{A})^{-1}\mathbf{A})\mathbf{Y}', \tag{2.11}$$

where $\mathbf{A}$ is defined in Equation (2.6). The growth curve coefficient matrix is then obtained as

$$\hat{\psi} = (\mathbf{B}'\mathbf{S}^{-1}\mathbf{B})^{-1}(\mathbf{B}'\mathbf{S}^{-1}\mathbf{Y})\mathbf{A}'(\mathbf{A}\mathbf{A})^{-1}, \tag{2.12}$$

which reduces to

$$(\mathbf{B}'\mathbf{S}^{-1}\mathbf{B})^{-1}(\mathbf{B}'\mathbf{S}^{-1}\mathbf{Y})$$

for the special matrix $\mathbf{A}$ defined in Equation (2.6). We now discuss two basic tests of interest, namely, if the given model provides an adequate fit, and if so, based on this model, if the data indicate significant differences between the r groups. To test the first of these hypotheses,

$\mathbf{H}_0$: Degree $(q-1)$ provides an adequate fit for the curves,

To test H_0 we find the Wilks' $\mathbf{\Lambda}$ statistic defined by

$$\mathbf{\Lambda}_0 = \frac{|\ \mathbf{E}_0\ |}{|\ \mathbf{E}_0 + \mathbf{H}_0\ |}, \tag{2.13}$$

where $\mathbf{E}_0 = \mathbf{B}_2'\mathbf{S}\mathbf{B}_2$ and $\mathbf{H}_0 + \mathbf{E}_0 = \mathbf{B}_2'\mathbf{Y}\mathbf{Y}'\mathbf{B}_2$. The test statistic (see Rao (1973)):

$$F_0 = \frac{1 - \sqrt{\mathbf{\Lambda}_0}}{\sqrt{\mathbf{\Lambda}_0}} . \frac{ms - 2\lambda}{r(p-q)}, \tag{2.14}$$

which has an approximate distribution F_{df_1, df_2} with degrees of freedom $df_1 = (p-q)r$, and $df_2 = ms - 2\lambda$. Here $m = N - \frac{p-q+r+1}{2}$, $s = (\frac{((p-q)r)^2 - 4}{r^2 + (p-q)^2 - 5})^{\frac{1}{2}}$, and $\lambda = ((p-q)r - 2)/4$.

Next, we test if the groups are significantly different from each other, i.e. the hypothesis

$$\begin{aligned} \mathbf{H}_1 &: \psi_1 = \psi_2 = ... = \psi_r \\ &: \textit{or } \mathbf{\Psi}\mathbf{M} = 0; \end{aligned} \tag{2.15}$$

where

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ . & . & . & ... & . \\ -1 & -1 & -1 & ... & -1 \end{bmatrix}.$$

To test this hypothesis we need:

$$\mathbf{E}_1 = (\mathbf{B}'\mathbf{S}^{-1}\mathbf{B})^{-1}$$

and

$$\mathbf{H}_1 = (\hat{\psi}\mathbf{M})(\mathbf{M}'\mathbf{R}_{11}\mathbf{M})^{-1}(\hat{\psi}\mathbf{M})'$$

where

$$\mathbf{R}_{11} = (\mathbf{A}\mathbf{A}')^{-1}[\mathbf{I} + \mathbf{A}\mathbf{Y}' \times \left\{\mathbf{S}^{-1}\mathbf{B}\left(\mathbf{B}\mathbf{S}^{-1}\mathbf{B}\right)^{-1}\right\}\mathbf{Y}\mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}]. \tag{2.16}$$

Under the hypothesis H_1, the test statistic

$$F_1 = \frac{1 - \sqrt{\mathbf{\Lambda}_1}}{\sqrt{\mathbf{\Lambda}_1}} . \frac{(N - r - (-p-q) - 1)}{m}. \tag{2.17}$$

where

$$\mathbf{\Lambda}_1 = \frac{|\ \mathbf{E}_1\ |}{|\ \mathbf{E}_1 + \mathbf{H}_1\ |}, \tag{2.18}$$

is the Wilks' $\mathbf{\Lambda}$, has an F distribution with $df = 2m, 2(N - r - (p-q) - 1)$ corresponding to the hypothesis and error degrees of freedom.

3 Fitting growth curves to yeast cell-cycle data

In this section we fit growth curve models to a labeled portion of the yeast cell-cycle data of Spellman et al. (1998). This microarray data set is publicly available at http::/genome-www.stanford.edu/cellcycle/ and has been analyzed by many authors. The particular data set we analyze here, deals with 798 genes and their log-transformed expression ratios are given at 18 time periods, so that in the growth-curve context $p = 18$. In this data set, there are $r = 5$ biologically different groups, labeled $MG1, G1, S, S/G2$, and $G2/M$ with $n_1 = 113, n_2 = 299, n_3 = 71, n_4 = 121$, and $n_5 = 194$ respectively for a grand total of $N = 798$ observations. Mean curves were tried with polynomials of varying degrees and the AIC and BIC criteria appear to indicate that a q value of 6 (corresponding to a 5th degree polynomial) provides a good fit. Figure 1 demonstrates the fit, by plotting the model-mean against the observed (sample) average at each time point for one of the groups, namely Group 5.

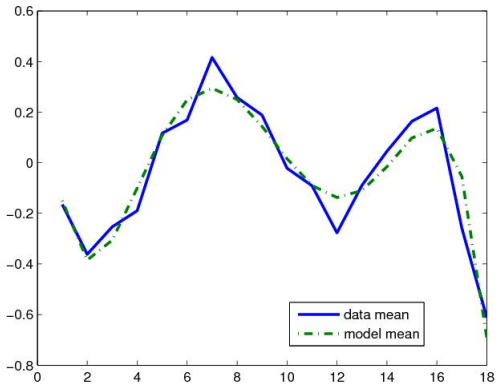

Figure 1. Observed and Model based Mean curves for Group 5

While this illustrates graphically that the fit is quite good, we do the test for the model fit, i.e., the hypothesis H_0. This gives $\mathbf{\Lambda}_0 = 0.1518$ (see Equation (2.13)) and a p-value close to zero, signifying a good fit.

Figure 2 provides a graphical comparison of the mean curves for the 5 groups.

The statistical test of hypothesis H_1 gives a $\mathbf{\Lambda}_1 = 0.5649$ (see Equation (2.18)) with a p-value again close to zero, indicating that the 5 groups do indeed have significantly different mean curves.

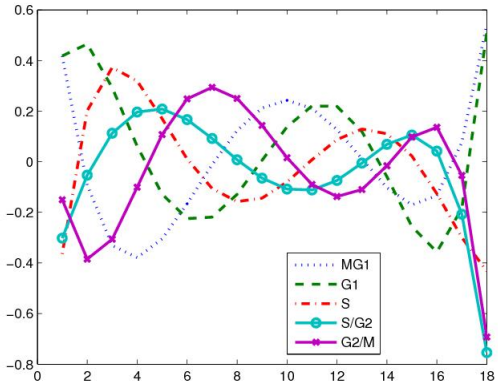

Figure 2. Model-based Mean curves for the 5 groups

4 Conclusions

Many authors who consider gene expression data such as the Spellman data, use supervised or unsupervised tools such as classification and clustering. Instead we consider here a statistical or confirmatory approach based on growth-curve models, to see what the mean curves of the biologically distinct groups look like, and whether they differ significantly. The statistical test concludes that these are indeed significantly different groups. Potential extensions of the current work include (i) estimating the common covariance matrix Σ by assuming that it depends on a smaller number of unknown parameters λ say $\Sigma(\lambda)$ as in time-series models, instead of the $\frac{p(p+1)}{2}$ unknowns, (ii) using cubic or B-splines instead of polynomials for modeling the mean functions.

References

[1] **Bar-Joseph, Z., Gerber, G., Jaakkola, T. S., Gifford, D. K. and Simon, I.**, (2003). Continuous representations of time series gene expression data, *Journal of Computational Biology*, **3**, 341-356.

[2] **James, G. and Hastie, T.**, (2001), Functional Linear Discriminant Analysis for Irregularly Sampled Curves, *J. Roy. Statist. Soc.*, **B63**, 533-550.

[3] **Khatri, C. G.**, (1966), A note on MANOVA model applied to problems in growth curves, *Ann. Inst. Stat. Math.*, **18**, 75-86.

[4] **Kshirsagar, A. M. and Smith, W. B.**, (1995), *Growth Curves*, Marcel Dekker, New York.

[5] **Liu, X., Sivaganesan, S., Yeung, K. Y., J. Guo, J., Baumgarner, R. E. and M. Medvedovic**. (2006), Context-specific infinite mixtures for clustering gene expression profiles across diverse microarray data set, *Bioinformatics*, **22**, 1737-1744.

[6] **Medvedovic, M. and S. Sivaganesan**, (2002), Bayesian infinite mixture model based clustering of gene expression data, *Bioinformatics*, **18**, 1194-1206.

[7] **Medvedovic, M., Yeung, K. Y. and Baumgarner, R. E.**, (2004), Bayesian mixture model based clustering of replicated microarray data, *Bioinformatics*, **20**, 1222-1232.

[8] **Pan, J-X. and Fang, K-T.**, (2002), *Growth Curve Models and Statistical Diagnostics*, Springer, New York

[9] **Potthoff, R. and Roy, S. N.**, (1964), A generalized multivariate analysis of variance models useful especially for growth curve problems, *Biometrika*, **51**, 313-326.

[10] **Rao, C. R.**, (1973), *Linear Statistical Inference 2nd Edn.*, Wiley, New York.

[11] **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, *Mol. Biology of the Cell*, **9**, 3273-3297.

Arvind K. Jammalamadaka
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology, USA
E-mail: *ajamma@gmail.com*

S. Rao Jammalamadaka
Department of Statistics and Applied Probability
University of California, Santa Barbara, USA
E-mail: *rao@pstat.ucsb.edu*