Mixture-Regression Cluster Model applied to Longitudinal Microarray Experiments

Emma Holian
Department of Mathematics and Statistics, University of Limerick, Ireland.

Keywords: Microarray; Longitudinal; Mixtures; Regression; Random effects.
AMS: 60J80

Abstract

The aim of this work is to explore various statistical techniques to identify genes which contribute to some change in phenotype level. For example, the response of fish kept under stressful conditions for various lengths of time. We aim to assess the level of differential expression of each gene in the tissue samples and also attempt to model the expression patterns of genes over time, not only to classify genes by similarities in expression patterns, but also to model these patterns as specified functions.

The proposed Mixture-Regression Cluster Model is developed to model and cluster the genes into groups according to their expressions measured over time. This model is similar to that of the multivariate normal mixture model in that clusters are identified by the EM algorithm but is adapted to incorporate the flexibility of regression curves to fit the trends. In this way, additional features such as covariates, random effects and correlation structures can be incorporated into the model while potentially offering a considerable saving on the number of parameters required to model the trends.

1. Introduction and Background

Microarray technology measures genetic expression in the cells of a tissue sample and is implemented to identify the function of genes in an organism. A cDNA microarray can measure the genetic expression exhibited in two tissue samples. The animal sources from which these tissue samples are taken are often chosen because they differ in phenotype for some particular trait, for example, trout fish displaying symptoms of stress versus unstressed trout. The source with the phenotype trait is often labelled as the treatment and the source not displaying the trait labelled the control.

Of course the genetic makeup of any one phenotype consists of many thousands of genes and so detecting which genes are relevant to that particular trait is no menial task. The microarray facilitates detection of the presence and abundance of the expression exhibited in the tissue samples of thousands of genes simultaneously since an array consists of thousands of probes of different genetic material spotted at key locations on a glass slide. The level of expression of a gene at a spot is measured by recording the levels of intensity of two fluorescent dye molecules, Cyan 5 and Cyan 3, when the array is excited by
a laser. Say the the Cyan 5 molecule is attached to the treatment genetic expression and the Cyan 3 attached to the control genetic expression then the ultimate aim is to examine the level of differential expression estimated between these two tissue sources.

In order to estimate and remove the effects of experimental and biological variation, the arrays are repeated with various combinations of technical and biological replicate samples. In this simple two tissue variety experiment, the analysis is theoretically a two sample comparison test for each gene, filtering for those genes that display significant differential expression. In practice however, in the context of microarrays, the researcher often finds numerous difficulties in this task. To name a few for example, the relatively low number of biological replicates available versus the high degree of experimental error seen in these experiments and the need for corrections for multiple testing.

If the level of differential expression for a gene is not significant the gene is said to be inactive with respect to the phenotype difference, that is, in the example given, the gene is inactive with respect to exposure to stress. If the level of measured expression for the gene is significantly higher in the treatment tissue than in the control tissue then the gene is said to be over-expressed meaning the gene becomes more active when the treatment phenotype trait is seen. Conversely, if the level of measured expression for the gene is significantly higher in the control tissue than in the treatment tissue, under-expression, then the gene becomes less active when the treatment phenotype trait is seen. Thus there is a natural classification of gene expression into groups or clusters of inactive, over-expressed and under-expressed genes.

As microarray experiments are increasing in popularity, geneticists have become more adventurous in the genetic questions they aim to answer. Thus experiments are increasing in complexity, requiring statistical consultation in first attaining the most efficient design for the experiment and later to analyse the data since these more adventurous experiments leads to unique computational and statistical problems in the fields of gene filtering and gene clustering.

In particular gene expression profiles obtained from time-course microarray experiments exhibit a unique opportunity to model the trends and correlations between longitudinal genetic expressions. In the following we outline details of the longitudinal trout fish stress experiment, and explore how the differential expression profiles can be clustered under a flexible parametric framework.

2. Outline of the Data Provided

Samples of liver tissue were extracted from rainbow trout fish exposed to confinement stress for varying lengths of time, at times 2, 6, 24, 168 and 504 hours of stress, these times represented by $t, t \in 1:5$, and labelled as tissue variety treatment. Samples of liver tissue from unstressed fish left for the same
period of time in a neighbouring tank are also included in the analysis, labelled as control samples.

Although a comparison is required between the expressions of treatment and control tissue samples at each time-point, these are not measured together directly on the same array. Instead each sample is paired with a common reference sample on an array. Let the measured expression for a probe at spot \( s \in [1 : 21168] \) be denoted by \( y_{sijkt} \) where \( i \in [1 : 80] \) indicates on which of the 80 arrays the expression was measured on, \( j \) is the dye indicator where \( j = 1 \) for Cyan 5 and \( j = 2 \) for Cyan 3, \( k \) denotes tissue variety, \( k = 0 \) for reference, \( k = 1 \) for treatment and \( k = 2 \) for control.

In order to remove some experimental variation, a series of corrections are applied to the measured expressions, a process referred to as normalisation. As part of this process we assume the treatment and control intensities are calibrated for array effects using the reference intensities. The following probe or spot-specific model is then applied to normalised measured expressions \( y'_{sijkt} \):

\[
\log_2 y'_{sijkt} = \mu_s + \text{Ref}_s + D_{sj} + V_{sk} + T_{st} + V T_{skt} + \epsilon_{sijkt}
\]

where \( \mu_s \) is the overall mean and \( D_{sj} \) is the spot specific dye effect. Note that the reference variety, \( k = 0 \) is parameterised separately as \( \text{Ref}_s \) and since the calibration applied earlier assumes the reference values to be uniform over time, \( T_{st} \) represents the spot-specific time effect common to both varieties treatment and control. \( V_{sk} \) is the spot specific treatment and control effects, \( k \in [1 : 2] \), while \( V T_{skt} \) estimates the changes in treatment and control expressions over the time-course of the experiment. Using these estimates, the differential expression profile between treatment and control for each probe \( s \) can be calculated as \( Y_s \), with elements

\[
Y_{st} = (\hat{V}_{s1} - \hat{V}_{s2}) + (\hat{V} T_{s1t} - \hat{V} T_{s2t}) \text{ for } t = [1, 5].
\]

Clustering these probes into groups of similarity may give some indication as to genes that co-regulate in the production of proteins in response to exposure to stress.

3. Formulation of Cluster Model and Estimation

The proposed Mixture-Regression Cluster Model is developed to model and cluster the genes into groups according to their expressions measured over time. This model is similar to that of the multivariate normal mixture model in that clusters are identified by the EM algorithm but is adapted to incorporate the flexibility of regression curves to fit the trends. In this way, additional features such as covariates, random effects and correlation structures can be incorporated into the model while potentially offering a considerable saving on the number of parameters required to model the trends.
For a particular cluster \( i \in [1 : c] \) let the differential response vector be modelled by \( Y_s = X_s \beta_i + Z_s b_{is} + \epsilon_{is} \), for fixed effects \( \beta_i \), usually the regression curve in time, specified by the design matrix \( X_s \), and any optional random effects \( b_{is} \) specified by design matrix \( Z_s \). Where the errors have a normal density \( \epsilon_{is} \sim N(0, \Sigma_i) \), and the random effects \( b_{is} \) have a normal density function \( f_i(b_{is}) = \phi(0, D_i) \), then the marginal model for \( Y_s \) is normally distributed, \( f_i(Y_s) = \phi(X_s \beta_i, V_{is}) \) with \( V_{is} = Z_s D_i Z_s' + \Sigma_i \).

The full distribution \( f(Y_s; \Psi) \) is then a mixture of the clusters so that

\[
Y_s \sim \sum_{i=1}^{c} \pi_i N(X_s \beta_i, V_{is}).
\]

Let the set of parameters for each of the component densities be denoted by \( \theta_i = (\beta_i, V_{is}) \) then \( \Psi = (\theta_1, \ldots, \theta_c, \pi_1, \ldots, \pi_{(c-1)})' \) is the vector containing all unknown parameters.

Estimation of these parameters, for a pre-specified number of components, \( c \), can be done via the Expectation-Maximization (EM) algorithm, an iterative procedure which is initiated be a random allocation of probes into the clusters. The M-step estimates the parameters, \( \Psi^{(1)} \), using this initial allocation, by a weighted regression using R-subroutines GLS for a model with no random effects and LME fitting a model with a random intercept or slope. The E-step then updates the allocation ratios using the estimated set of parameters \( \Psi^{(1)} \) from the M-step. The iterations continue until there is little difference in the observed log likelihood as calculated in the E-step.

**4. Results and Remarks**

Simulations have shown that the mixture-regression model can recover clusters successfully and for each resulting cluster can provide a parametric model for the longitudinal trend followed by probes in the same cluster.

To find the model specification of optimal fit to the data, certain features of the model can be varied and the model refitted. For example re-specifying, the number of clusters \( c \), or re-specifying \( X_s \beta_i \) to be polynomials of varying degrees in time, re-specifying \( Z_s b_{is} \) to include a random intercept or slope and varying the correlation structure within each cluster \( \Sigma_i \). The optimal model is selected so that the log-likelihood is maximised, or if a penalisation for the number of parameters is more desirable aim to minimise Akaike’s Information Criterion \( AIC \), or the more prudent Bayesian Information Criterion \( BIC \).

We show how these procedures were applied to a filtered subset of the fish stress probes resulting in a 17-component mixture. Some discussion will also follow as to how a number of these interesting clusters have proven to be a very useful source of information in understanding the molecular processes.
tested in these experiments.

**Acknowledgements:** Special Thanks to colleagues working at the National Diagnostics Centre, and in the Mathematics Department of National University of Ireland, Galway, where the work was carried out. This research was supported by the Embark Initiative operated by the Irish Research Council for Science, Engineering and Technology (IRCSET).

5. Bibliography

