

Multi-level mixture modeling with subset selection – with applications to clustering of gene expression data.

Rebecka Jorntsen, Department of Statistics, Rutgers University.

<http://www.stat.rutgers.edu> rebecka@stat.rutgers.edu

Abstract

Mixture models are popular tools to summarize high-dimensional and complex data structures. Applications of mixture models in the literature range from the social sciences (network modeling), to engineering (image processing), and functional genomics (gene/sample clustering).

We present statistical methodologies that can further enhance the utility of mixture models:

- (1) Multi-level mixture models allow for clustering, or the grouping of data objects, in multi-factor experiments. Potential applications include cross-species comparisons, and time-course expression studies. In addition, multi-level mixture models allow for the grouping of data using different similarity metrics simultaneously;
- (2) Simultaneous subset selection provides sparse model representations of each cluster component, e.g. a "flat" or "increasing" profile, and facilitates objective interpretation of the clustering outcome. We show that an efficient representation of each cluster profile can in fact enable the detection of more clusters, compared with standard clustering approaches.

Mixture models with multiple levels

These days, functional genomics studies have gone well beyond comparisons between two experimental conditions, e.g. cancer/not-cancer. Most microarray studies now cover multiple experimental factors, e.g. dose, time, animal-model.

Yet, when it comes to clustering such data, it is common to simply combine all arrays into one data set, and cluster as if only one experimental factor was present. Clearly, we should be taking the experimental design into account when we cluster, to enable direct interpretation of cluster profiles in the context of the study.

We propose the MIXL model to cluster multi-factor experiments. Take as an example a two-factor study of two proliferating stem cell lines (factors: cell line and time). One cell line tends to produce neurons, the other glia. Both cell lines are observed over a course of 3 days, during which cell differentiation takes place. In figure 1C below, we see that many of the glia cell line clusters overlap, whereas the same clusters are distinct in neurons. MIXL allow us to directly model clusters that overlap for one experimental factor level (glia), facilitating objective interpretation of clusters that are neuron-specific.

In figure 1B we show that MIXL allows for more clusters to be detected. The BIC curve (circles) with the standard method identifies 8 clusters, whereas MIXL (triangles) detects 9. MIXL combines clusters (2,5) and (6,9) for the glia cell-line (figure 1C). Looking at the gene ontology associated with these clusters we identify genes that are specific to neuron formation [2].

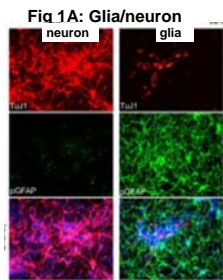


Fig 1A: Glia/neuron

Fig 1C: Cluster means. Note that clusters (6,9), (2,5) appear to overlap in the glia cell-line

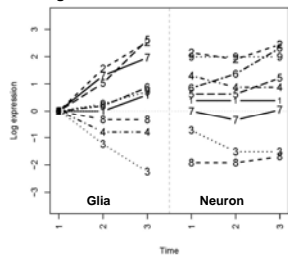
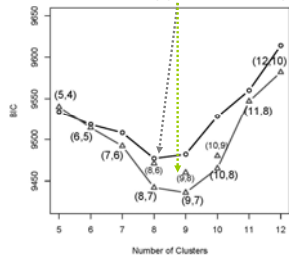


Fig 1B: Number of clusters detected by the standard method (o) and with MIXL (Δ).



Mixture models with multiple similarity metrics

A challenge in clustering is deciding on an appropriate (dis)similarity metric, e.g. correlation, absolute correlation, or euclidean distance. Euclidean distance groups genes with a similar level of expression in all conditions. Correlation groups genes with similar shape of expression changes, although the scale of expression can differ. While most experimentalists prefer correlation, few actually believe that the scale of expression is meaningless.

Multi-level mixture models can be used to incorporate different similarity metrics. Each level corresponds to a data transform, such as "sign-flip" (S), and "standardization" (T).

Take as an example a 3-level mixture model. At the 1st level we describe cluster shapes of expression profiles (e.g. "increasing", "transient"). At the 2nd level we allow for sign-flips (e.g. taking "increasing" to "decreasing"). At the 3rd level we allow for all shapes (and sign-flip shapes) to be scaled by a cluster specific factors (e.g. expression shape*10, expression shape*2).

We can choose to make inferences about gene expression at any of the levels of the modeling hierarchy. For example, if we only care about shape, we aggregate the clusters from the 2nd and 3rd levels. If we care about the absolute expression changes, we make inferences at the 3rd level.

A standard mixture modeling or clustering approach would have to describe all clusters with the same level of detail. In a multi-level approach, many of the parameters are shared. For example, the shapes are shared across levels and this constitutes a substantial savings in terms of the complexity of the model. As a result, the multi-level approach can afford many more clusters than a standard approach.

Fig 2A: cluster profiles from the scale-transform model. Clusters 5 and 6 have the same shape, but different scale.

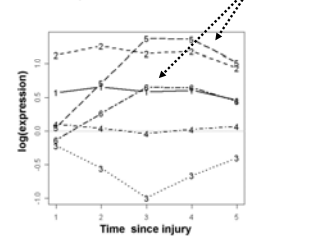
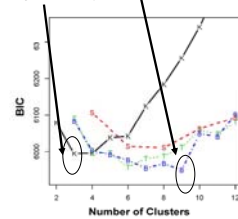


Fig 2B: BIC curves for the standard (K), sign-flip (S), scale-transform (T) and sign-flip+scale (B) models. The transform models detects more clusters than the standard model (e.g. 9 with B, 6 with R, 8 with S, and only 3 with K).



We apply the multi-level mixture model to a time course expression data set (time after spinal cord injury). In figure 2A we show the clustering outcome of a 2-level model with scale transforms, where cluster pairs (1,2), (3,4), and (5,6) have the same shape, but different expression scale. Cluster 5 represents a stronger response to injury compared with cluster 6. The ontology categories of cluster 5 are associated with cell-death, whereas cluster 6 is associated with "repair".

In figure 2B, we show that multi-level models (S=sign-flip, T=scale transform, and the 3-level model B=both S and T) allow for many more clusters than the standard model (K). This is because parameters are shared across levels. The standard model (K) can only detect 3 clusters, whereas the 3-level mixture model detects 9 clusters with 3 unique shapes.

Potential relevance to environmental toxicology

The mixture models outlined here are generally applicable to any high-dimensional complex data structure. In particular, the multi-level model (MIXL) with subset selection can be applied to cross-species studies, to facilitate the comparisons of gene expression between species, in a dose-response setting. If we replace the glia and neuron cell-lines in our example with e.g. mouse and rat, the MIXL model would identify rat-specific gene expression patterns as well as gene groups with a common expression profile for both species.

These methodologies are also applicable to e.g. protein or tissue array studies.

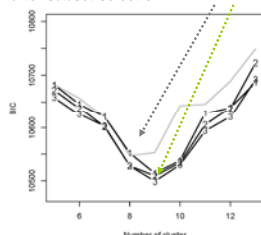
The transform mixture models are especially efficient for modeling time course data, and data where the metric is difficult to specify. Instead of making arbitrary choices of the metric to use for summarizing the high-dimensional data, the multi-level model allows for the analysis to incorporate several metrics of interest all at once.

Cluster model subset selection

The outcome of clustering is a set of cluster profiles which are commonly interpreted in a subjective fashion, e.g. "this cluster appears to represent increasing expression over time". We propose a subset selection scheme that provides sparse and easy-to-interpret representations of each cluster profile. To date, model selection in clustering has primarily focused on selecting variables (experimental conditions) that appear to be related to the clustering. However, this constitutes a very limited set of models for the cluster profiles. Instead, we propose a parameterization of the cluster profiles, and select parameters instead of variables (e.g. a linear increase over time instead of quadratic).

As an example, consider the cell-line time course data (above). We can parameterize each cluster in terms of cell-line differential expression (1), in terms of time course models for each cell-line (3), or as a time course of cell-line differences (2). In figure 3A we show BIC curves obtained under each parameterization, which indicates that parameterization 3 is the most efficient. We can now interpret the cluster directly in terms of the time course patterns corresponding to each model (e.g. static, increasing). Furthermore, by performing parameter selection we detect one more cluster compared with the standard mixture model [3].

Fig 3A: Parameterization (3) provides the smallest BIC. We detect one more cluster after subset selection.



References:
 [1] Jorntsen, R. (2006) Simultaneous subset selection via rate-distortion theory, with applications to gene clustering and significance analysis of differential expression. Technical report, Rutgers University, Department of Statistics.
 [2] Jorntsen, R. (2006) Multi-level mixture models – cluster profile transformations. Technical report, Rutgers University, Department of Statistics.
 [3] Jorntsen, R., Keles, S. (2006) Mixture models with multiple levels, with applications to the analysis of multi-factor gene expression experiments. Technical report, Rutgers University, Department of Statistics.



Acknowledgements

Support for this work has been provided primarily by the USEPA-funded Environmental Bioinformatics and Computational Toxicology Center (ebCTC), under STAR Grant number GAD R 832721-010. This work has not been reviewed by and does not represent the opinions of the funding agency.