SAND2017-6232C

Visualizing Clustering and Uncertainty Analysis with Multivariate Longitudinal Data

Maximillian G. Chen* Kristin M. Divis† Laura A. McNamara‡ James D. Morrow§

Sandia National Laboratories

ABSTRACT

Longitudinal, multivariate datasets are intrinsic to the study of dynamic, naturalistic behavior. Statistical models provide the ability to identify event patterns in these data under conditions of uncertainty. To make use of statistical models, however, researchers must be able to evaluate how well a model uses available information in a dataset for clustering decisions and for uncertainty estimation. The Gaussian mixture model is a prominently used model for clustering multivariate data. However, it has only been recently extended to longitudinal data, and useful visualization tools have yet to be developed in this context. In this paper, we develop novel methods for visualizing the clustering performance and uncertainty of fitting a Gaussian mixture model to multivariate longitudinal data. We demonstrate our methods on eyetracking data and explain the usefulness of uncertainty quantification and visualization with evaluating the performance of clustering models.

1 MODEL-BASED CLUSTERING WITH THE GAUSSIAN MIXTURE MODEL (GMM)

Cluster analysis is the automated search for groups of related observations in a dataset. Groups of observations that are cohesive and separated from other groups are identified. Finite mixture models, in which each component probability distribution corresponds to a cluster, provide a principled statistical approach to determining the number of clusters and choosing an appropriate clustering method. Models that differ in number of components and/or in component distributions can be compared. By using a probabilistic model to represent the clustering problem, we can use uncertainty quantification to assess the variability in the performance of a clustering model.

1.1 GMM for Independent and Identically Distributed (i.i.d.) Data

We refer to "model-based clustering" as the use of (finite) mixture models to perform clustering. We consider the Gaussian mixture model (GMM), where the density of a random vector $\mathbf{y}$ can be written as a mixture of G components as follows:

$$f(\mathbf{y}|\theta) = \sum_{g=1}^{G} \pi_g \frac{\exp\{-\frac{1}{2}(\mathbf{y}-\mu_g)^T \Sigma_g^{-1}(\mathbf{y}-\mu_g)\}}{\sqrt{\det(2\pi\Sigma_g)}}, \qquad (1)$$

where the gth component density is a multivariate normal distribution with mean μ_g and covariance matrix Σ_g, and $\pi_g > 0$ such that $\sum_{g=1}^{G} \pi = 1$ are called mixing proportions. Suppose n p-dimensional data vectors $\mathbf{y}_1, ..., \mathbf{y}_n$ are observed, independent and identically (i.i.d.) distributed, and all n are unlabelled or treated as unlabelled. The complete-data likelihood for the mixture model is

$$\mathscr{L}_C(\pi_g, \mu_g, \Sigma_g) = \prod_{i=1}^{n} \prod_{g=1}^{G} [\pi_g f(y_i|\mu_g, \Sigma)]^{z_{ig}}, \qquad (2)$$

where z_{ig} denotes the membership of observation i in component g so that $z_{ig} = 1$ if observation i belongs to component g and $z_{ig} = 0$ otherwise. The parameters of the GMM are estimated by an expectation-maximization (EM) algorithm [2], providing a closed form estimate of the probability that a sample i belongs to a group g, $\hat{z}_{ig}$. The value z^*_{ig} of $\hat{z}_{ig}$ at a maximum of (2) is the estimated conditional probability that observation i belongs to group g. The maximum likelihood classification of observation i is $\{j|z^*_{ij} = \max_g z^*_{ig}\}$, so that $(1 - \max_g z^*_{ig})$ is a measure of the uncertainty in the classification [1]. The R package mclust [3] performs this model fitting.

1.2 GMM for Longitudinal Data

[4] propose an EM algorithm for fitting a GMM to longitudinal data. The temporal correlation between observations is accounted by the modified Cholesky decomposition [5,6] of the inverse covariance matrix,

$$\Sigma^{-1} = T' D^{-1} T,$$

where T is a unique lower triangular matrix with diagonal elements 1 and D is a unique diagonal matrix with strictly positive diagonal entries. The linear least-squares predictor of Y_t, based on $Y_{t-1}, ..., Y_1$, can be written as

$$\hat{Y}_t = \mu_t + \sum_{s=1}^{t-1} (-\phi_{ts})(Y_s - \mu_s) + \sqrt{d_t}\varepsilon_t, \qquad (3)$$

where $\varepsilon_t \sim N(0,1)$, the ϕ_{ts} are the (sub-diagonal) elements of T and the d_t are the diagonal elements of D. The density of an observation y_i in group g is given by

$$f(y_i|\mu_g, T_g, D_g) = \frac{1}{\sqrt{(2\pi)^p |D_g|}} \exp\{-\frac{1}{2}(y_i - \mu_g)' T_g' D_g^{-1} T_g (y_i - \mu_g)\}. \qquad (4)$$

The complete-data likelihood for the mixture model is given by

$$\mathscr{L}_C(\pi_g, \mu_g, T_g, D_g) = \prod_{i=1}^{n} \prod_{g=1}^{G} [\pi_g f(x_i|\mu_g, T_g, D_g)]^{z_{ig}}. \qquad (5)$$

The R package longclust [4] implements the EM algorithm to fit this model. We can also take the value z^*_{ig} of $\hat{z}_{ig}$ at a maximum of (5) to be the maximum likelihood classification of observation i is $\{j|z^*_{ij} = \max_g z^*_{ig}\}$ and $(1 - \max_g z^*_{ig})$ as a measure of the uncertainty in the classification [1]. However, the longclust currently neither computes this uncertainty nor creates visualization plots.

*e-mail: mgchen@sandia.gov
†e-mail:kmdivis@sandia.gov
‡e-mail:lamcnam@sandia.gov
§e-mail:jdmorr@sandia.gov

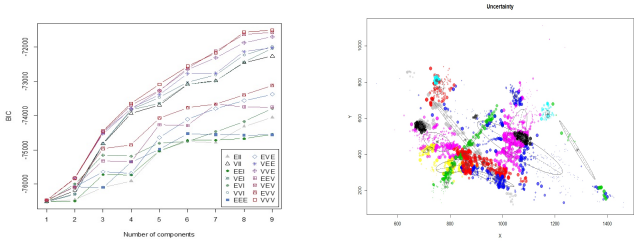
2 Visualization and Application to Eyetracking Dataset

2.1 Eyetracking Dataset

Our interest in this problem is motivated by an interest in enhanced exploitation of eyetracking data. We focus on a dataset collected by Sandia National Laboratories. The dataset consists of eyetracking data collected from 16 human subjects. Each subject looks at various points in an image, and the locations that the subject looks at is tracked in a one-hour long experiment, with predetermined locations popping up in the image over the course of the experiment. A datapoint containing the spatial location of the subject's eye target is recorded every 17 milliseconds, so there are 25,000 sample points for the one subject throughout the four trials.

2.2 Existing Methods

The R package mclust fits a GMM to i.i.d. multivariate data and clustering results and uncertainty. No such capability currently exists in the R package longclust. Below in Figure 1 is a clustering and uncertainty plot for eyetracking data described in the previous section for one trial taken by one subject. From the plot below, we see that the clusters drawn from the mclust package do not match up well with the observed data because it does not factor in the temporal correlation between observations.

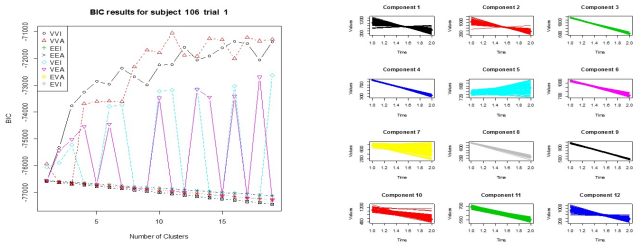


(a) BIC values of mclust models

(b) Clustering results and uncertainty

Figure 1: BIC and clustering results and uncertainty plots for R package mclust applied to eyetracking data. The model chosen by the highest BIC value (a) is a model with 20 clusters and parametrization VVV. (b) consists of the clustering results and associated uncertainty, which is represented by the ellipses.

Below in Figure 2 are the plots currently available in the longclust package for longitudinal data applied to the same dataset. It is unclear that the values in the time plots (Figure 2(b)) represent.



(a) BIC values of longclust models

(b) Time plots

Figure 2: Currently available plots for R package longclust applied to eyetracking data. The model chosen by the highest BIC value (a) is a model with 12 clusters. (b) consists of time plots for the 12 clusters and appears to be the values for a parameter associated with the 12 clusters over the running of the EM algorithm until convergence.

2.3 Clustering and Uncertainty Plots for Longitudinal Data

We create an analagous clustering and classification uncertainty plot as the one available in the mclust package for longitudinal data. We use the T and D matrices (computed by longclust) to estimate the covariance matrix Σ, which is used to compute the uncertainty ellipse for each cluster. Below in Figure 3 is the resulting clustering and uncertainty plot. We see that by factoring in the temporal correlation between observations, we get much better clustering results, as the uncertainty ellipses encompass the data better and the ellipses are thinner, which indicate lower classification uncertainty and the GMM is a reasonable fit for the data.

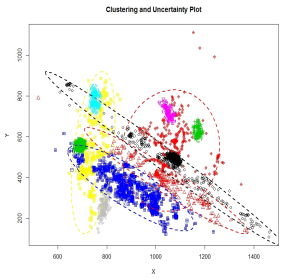


Figure 3: Clustering analysis of eyetracking data using a GMM fit to longitudinal data.

3 Conclusions

Utilizing recently developed methods for clustering multivariate longitudinal data via the Gaussian mixture model, we create and demonstrate novel visualization methods for the clustering performance and assessing the clustering uncertainty. We show how the visualization methods can allow us to gauge the significant improvement in clustering performance and uncertainty that correctly factoring in the temporal correlation between observations can bring. Furthermore, we argue for the usefulness of these visualization techniques to assess the performance of clustering models and the potential to try alternative clustering models for a particular dataset. We demonstrate our methods on an eyetracking dataset, but our methods can be applied to longitudinal datasets in a wide array of application areas, such as radar and surveillance, medicine, and finance. The capability to visualize clustering performance and uncertainty greatly enhances the ability to fully exploit all of the information available in any dataset.

References

[1] H. Bensmail, G. Celeux, A. E. Raftery, and C. P. Robert. Inference in model-based cluster analysis. *Statistics and Computing*, 7(1):1–10, Jan. 1997. doi: 10.1023/A:1018510926151

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

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

[4] P. D. McNicholas and T. B. Murphy. Model-based clustering of longitudinal data. *Canadian Journal of Statistics*, 38(1):153–168, 2010. doi: 10.1002/cjs.10047

[5] M. Pourahmadi. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

[6] M. Pourahmadi. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 87(2):425–435, 2000.