SAND2017-7948C

Visualizing Clustering and Uncertainty Analysis with Multivariate Longitudinal Data

Maximillian G. Chen* Kristin M. Divis* Laura A. McNamara* J. Dan 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 (GMM) 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 GMM 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 INTRODUCTION AND MOTIVATION

Longitudinal data, or panel data, refers to multi-dimensional data involving measurements over time. Observations of multiple phenomena over multiple time periods are taken for the sample subjects of interest. One motivating example is data collected from eye tracking systems, which are widely used to map human visual interactions with data and information. Eye trackers can generate voluminous spatio-temporal datasets comprising thousands of individual gaze samples that represent the calculated location of an individual's gaze against the display space. These gaze samples are aggregated using spatiotemporal thresholding algorithms into recognized behavioral indicators, such as saccades and fixations, that describe visual interaction with a stimulus.

Recently, eyetracking data analysis is extending to include probabilistic approaches that provide a more nuanced analysis of human-information interactions. We explore the application of finite mixture models, in which each component probability distribution corresponds to a cluster, for probabilistic clustering of gaze samples into compound measures, such as fixations and clusters of fixations (or "dwells"). 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. Most existing probabilistic clustering models that assume observations are independent and identically distributed (i.i.d.), meaning these datasets have one observation for each subject. Models for longitudinal data must account for the temporal correlation, or dependence, between observations. Furthermore, current visualization tools are inadequate for assessing the performance of finite mixture models with eyetracking datasets, which are both spatially *and* temporally distributed. We aim to develop visualizations that will enable researchers to assess the performance of finite mixture models for spatial clustering of longitudinal data. These visualizations allow us to see how well a clustering model performs at grouping observations and helps an analyst determine whether the current clustering model is appropriate or an alternative model is required.

* email: mgchen@sandia,gov
* email: kmdivis@sandia.gov
* email: lamcnam@sandia.gov
* email: jdmorr@sandia.gov

2 CLUSTERING WITH THE GAUSSIAN MIXTURE MODEL(GMM)

2.1 Model-Fitting Methods

We consider the prominent probabilistic clustering model, Gaussian mixture model (GMM), where the density of a random vector **y** can be written as a mixture of G components (or clusters) as follows:

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

where the *g*th component density is a multivariate normal distribution with mean μ_g and covariance matrix $\sum_g$, and $\pi_g > 0$ such that $\sum_{g=1}^G \pi = 1$ are called mixing proportions. The parameters of the GMM are estimated using an expectation-maximization (EM) algorithm [2], providing a closed form estimate of the probability that a sample *i* belongs to group *g*, $\hat{z}_{ig}$. The value z_{ig}^* of $\hat{z}_{ig}$ at a maximum of the complete-data likelihood 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] fits the model in the case where the data consists of *n* i.i.d. *p*-dimensional data vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$, and all *n* clusters are unlabelled or treated as unlabelled. It also computes the classification uncertainties and produces plots of the cluster assignments for each data point and the uncertainty ellipsoids for the clustering performance of the GMM.

[4] propose an EM algorithm for fitting a GMM to longitudinal data and create the R package longclust to implement the algorithm. The temporal correlation between observations is accounted for by the modified Cholesky decomposition [5,6] of the inverse covariance matrix. After fitting this model and obtaining the estimates for the parameters of the GMM, the classification uncertainty can be computed the same way as above. However, the longclust package does not compute and visualize the uncertainty of each classification.

2.2 Visualization Method

After fitting the GMM to a longitudinal dataset and obtaining the cluster assignments for each cluster, we compute the uncertainties of each observation belonging to each cluster and produce a plot of the cluster assignments of each data point (different colors and/or symbols indicating different clusters) and uncertainty ellipsoids surrounding the clustering performance of the GMM. If the data points are close together and the uncertainty ellipsoids are narrow for all clusters, then the model does a good job at clustering the data.

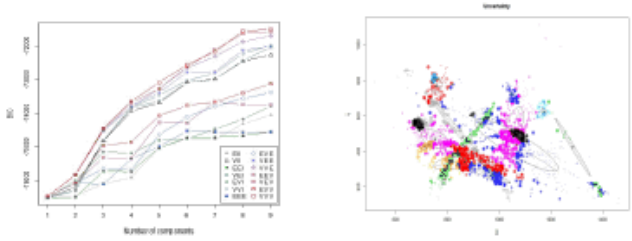
3 Visualization and Application to Eyetracking Dataset

3.1 Eyetracking Dataset

We focus on enhanced exploitation of an eyetracking dataset collected by Sandia National Laboratories that consists of 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 pre-determined 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.

3.2 Existing Methods

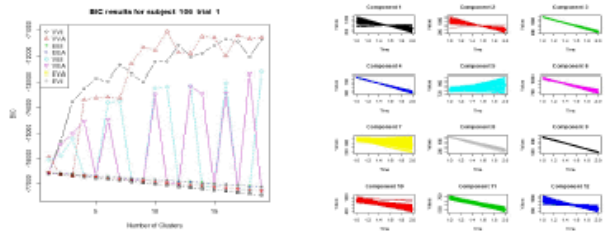
The R package longclust currently does not have the capability to plot the clustering results and uncertainty for fitting a GMM to longitudinal multivariate data that the mclust package does for i.i.d. data. Below in Figure 1(b) is a clustering and uncertainty plot for eyetracking data described in section 3.1 for one trial taken by one subject using the mclust package. In this plot, the clustering uncertainty ellipses drawn do not match up well with the observed data because it does not factor in the temporal correlation between observations. The plot indicates this clustering method is not appropriate for our data.



(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 what the values in the time plots in Figure 2(b) represent. Furthermore, we cannot assess the quality of the clustering from these plots, as we cannot visualize the clustering results and uncertainty.



(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. However, it is unclear what that parameter is.

3.3 Proposed Clustering and Uncertainty Plots for Longitudinal Data

We create an analogous 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 $\sum$, which is used to compute the uncertainty ellipse for each cluster. Below in Figure 3 is the resulting clustering and uncertainty plot. 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.

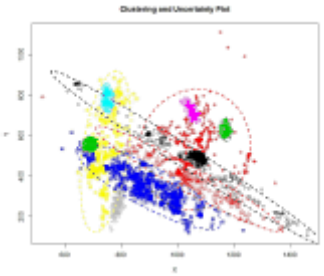


Figure 3: Clustering analysis of eyetracking data using a GMM fit to longitudinal data.

4 Conclusion

Utilizing recently developed methods for clustering multivariate longitudinal data via the Gaussian mixture model, we create and demonstrate novel visualization methods for the clustering results and uncertainty. The visualization methods allow us to gauge the significant improvement in clustering performance when the temporal correlation between observations is accounted for. 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 its associated uncertainty greatly enhances the ability to fully exploit all of the information available in any dataset.

Acknowledgement

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

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.