Visualizing Clustering and Uncertainty Analysis with Multivariate Longitudinal Data

Maximillian G. Chen, Kristin M. Divis, Laura A. McNamara, and J. Dan Morrow

Sandia National Laboratories, Albuquerque, NM

SAND2017-0574C

Sandia National Laboratories

Introduction and Motivation

Background:

- Longitudinal data (panel data): multi-dimensional data where observations of multiple phenomena over multiple time periods are taken for the same subjects of interest
- Longitudinal, multivariate data are intrinsic to the study of dynamic, naturalistic behavior.
- Probabilistic clustering models, such as the Gaussian mixture model (GMM), allow for identifying patterns in data under conditions of uncertainty.
- Most existing probabilistic clustering models assume observations are independent and identically distributed (i.i.d.), meaning these datasets have one observation for each subject.
- Probabilistic clustering models have only recently been extended to longitudinal data, accounting for the temporal correlation between observations.

Eyetracking Data:

- Eyetrackers 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.
- Gaze samples are aggregated using spatiotemporal thresholding algorithms into recognized behavioral indicators, such as saccades and fixations, that describe visual interaction with a stimulus.
- Current visualization tools are inadequate for assessing the performance of finite mixture models with eyetracking datasets, which are both spatially and temporally distributed.
- Question: **Can we use GMMs for probabilistic clustering of spatio-temporal eyetracking data?**

Approach

1. Gaussian Mixture Model

- Density:

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

where μ_g is the mean vector and Σ_g is the covariance matrix of component g.

- Complete-Data Likelihood:

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

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.

- EM algorithm estimates all parameters
- Classification MLE: $\{j|z_{ij}^* = \max_g z_{ig}^*\}$
- Classification Uncertainty: $(1 - \max_g z_{ig}^*)$

2. Independent and Identically Distributed (i.i.d.) Data

- R package: mclust
- Geometric cross-cluster constraints in multivariate normal mixtures by parameterizing covariance matrices through eigenvalue decomposition in the form

$$\Sigma_g = \lambda_g D_g A_g D_g^T, \quad (3)$$

where D_g is the orthogonal matrix of eigenvectors, A_g is a diagonal matrix whose elements are proportional to the eigenvalues, and λ_g is an associated constant of proportionality.

3. Longitudinal Data

- R package: longclust
- The temporal correlation between observations is accounted by the modified Cholesky decomposition 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 values of T and D have interpretations as generalized autoregressive parameters and innovation variances, respectively, so that the linear least-squares predictor of Y_t, based on $Y_{t-1}, \ldots, Y_1$, can be written as

$$\hat{Y}_t = \mu_t + \sum_{s=1}^{t-1}(-\phi_{ts})(Y_s - \mu_s) + \sqrt{d_t}\epsilon_t, \quad (4)$$

where $\epsilon_t \sim N(0, 1)$, the ϕ_{ts} are the (sub-diagonal) elements of T and the d_t are the diagonal elements of D.

Existing Visualization

Mclust:

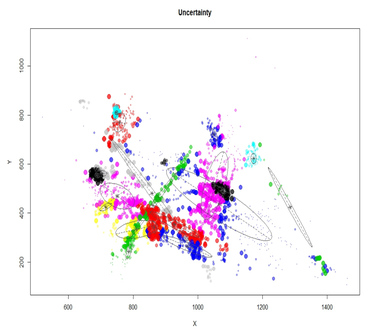


Figure: Clustering analysis of eyetracking data. The spatial locations of the subject's eye fixation location is divided into 20 clusters, based on the BIC values of the models tested. The ellipses represent the uncertainty of the clustering performance.

Issues: The clustering uncertainty ellipses do not match up well with the observed data because it does not factor in the temporal correlation between observations. This plot does not give us useful information about the clustering of the data and the accuracy of the clustering.

Longclust:

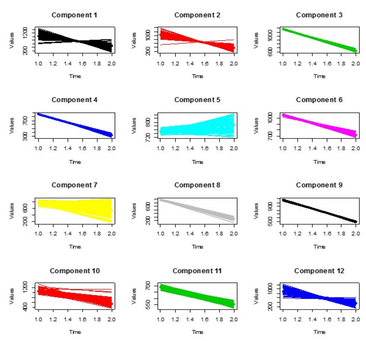


Figure: Time plots for a longitudinal mixture model with 12 clusters, based on the BIC values of the models tested, of what appears to be the values for a parameter associated with the 12 clusters over the running of the EM algorithm until convergence.

Issues: It is unclear how to read these plots. While these plots appear to be the values of a parameter associated with the 12 clusters, we do not know what this parameter is and what these values represent, In addition, the behavior of the lines in these plots is confusing. Without any plots to determine the clustering uncertainty, we have no way to gauge the clustering performance of the GMM.

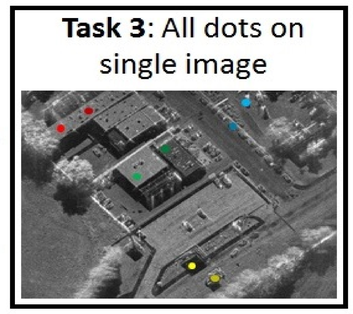
Eyetracking Dataset

- 16 human subjects
- Each subject looks at various points in an image, and the locations that the subject looks at are tracked in a one-hour long experiment consisting of four constrained visual search tasks.
- A datapoint containing the spatial location of the subject's eye target is recorded every 17 milliseconds.

Simplified Task Example

- Constrained visual search task where participants pan through the image and switch between images freely to find dots, sometimes making comparisons between dots.

Find **pairs of dots** in set order. Look back & forth between dots for set amount of time to make relative luminance judgment before finding next pair of dots.



Proposed New Visualization

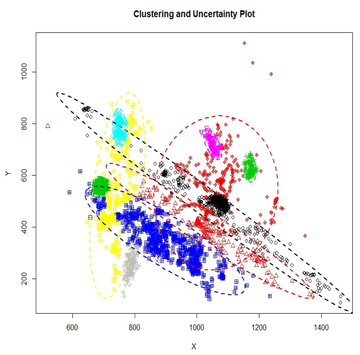


Figure: Clustering analysis of eyetracking data using a GMM fit to longitudinal data.

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.

Conclusions

- Utilize recently developed methods for clustering multivariate longitudinal data via the Gaussian mixture model.
- Create and demonstrate novel visualization methods for the clustering performance and assessing the clustering uncertainty.
- Allow us to gauge the significant improvement in clustering performance and uncertainty that correctly factoring in the temporal correlation between observations can bring.
- 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.

Contact: mgchen@sandia.gov