SAND2017-7706C

Gaussian Mixture Models for Information Integration: Toward Gaze-Informed Information Foraging Models for Imagery Analysis

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

Sandia National Laboratories, Albuquerque, NM

Sandia National Laboratories

Problem

- As eyetracking data moves from laboratory to naturalistic domains, researchers have the opportunity to develop rich ecological models of human-information interaction.
- Doing so requires developing new data collection and analysis frameworks that facilitate reliable integration of eyetracking data with complementary indicators of human work behaviors, in the context of computer-supported visual workflows.
- Uncertainty quantification for gaze events is needed to understand quality of detection of such events. Current detection methods do not provide uncertainty information.
- Little research has been done with probabilistic clustering models that factor in temporal correlation between observations.

Solution

- Apply probabilistic clustering models, specifically *Gaussian mixture models* (GMMs), to the analysis of eyetracking data collected during a dynamic search task, in which participants were directed to look for specific features in a Synthetic Aperture Radar (SAR) image.
- Parameterize the GMM so that the temporal correlation between observations is utilized.
- Using probabilistic clustering models, such as the GMM, allows quantification of classification uncertainty for gaze events.
- Provide an efficient way to *associate* gaze events with geospatial content in dynamic, user driven workflows *under uncertainty*.

.

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 in a constrained visual search task.
- 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.
- See also Divis, Chen, McNamara, Morrow, & Perkins poster

.

Approach

1. Gaussian Mixture Model

- Density:

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

  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

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

  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 $Y_{t-1}, \dots, Y_1$, can be written as

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

  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.

Main Finding

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.

Future Work

- Create R package that can visualize clustering performance and uncertainty for multivariate longitudinal data. This capability exists for i.i.d data with the mclust package, but it does not exist for longitudinal data with the longclust package.
- Integrate clustering and uncertainty results across tasks and subjects.
- Factor in velocity of eyetracking points and time in between observations into clustering models.
- Determine spatial and temporal sources of uncertainty.

Results

Mclust (i.i.d. data):

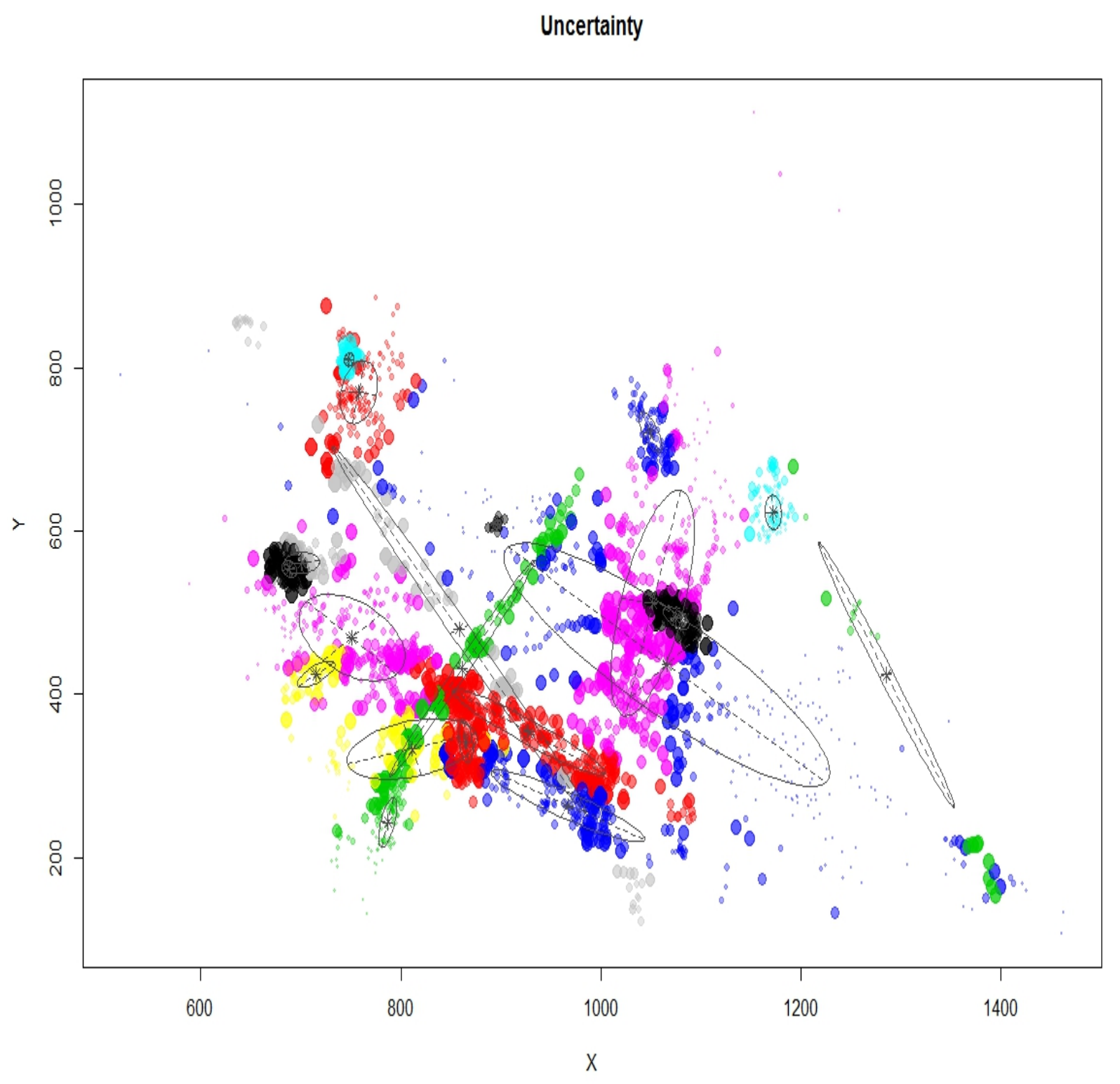

Figure: Clustering analysis of eyetracking data using a GMM fit to i.i.d. data. The spatial locations of the subject's eye fixation location is divided into 20 clusters, based on the BIC values of the models tested. The ellipses represent the uncertainty of the clustering performance.

Longclust (longitudinal data):

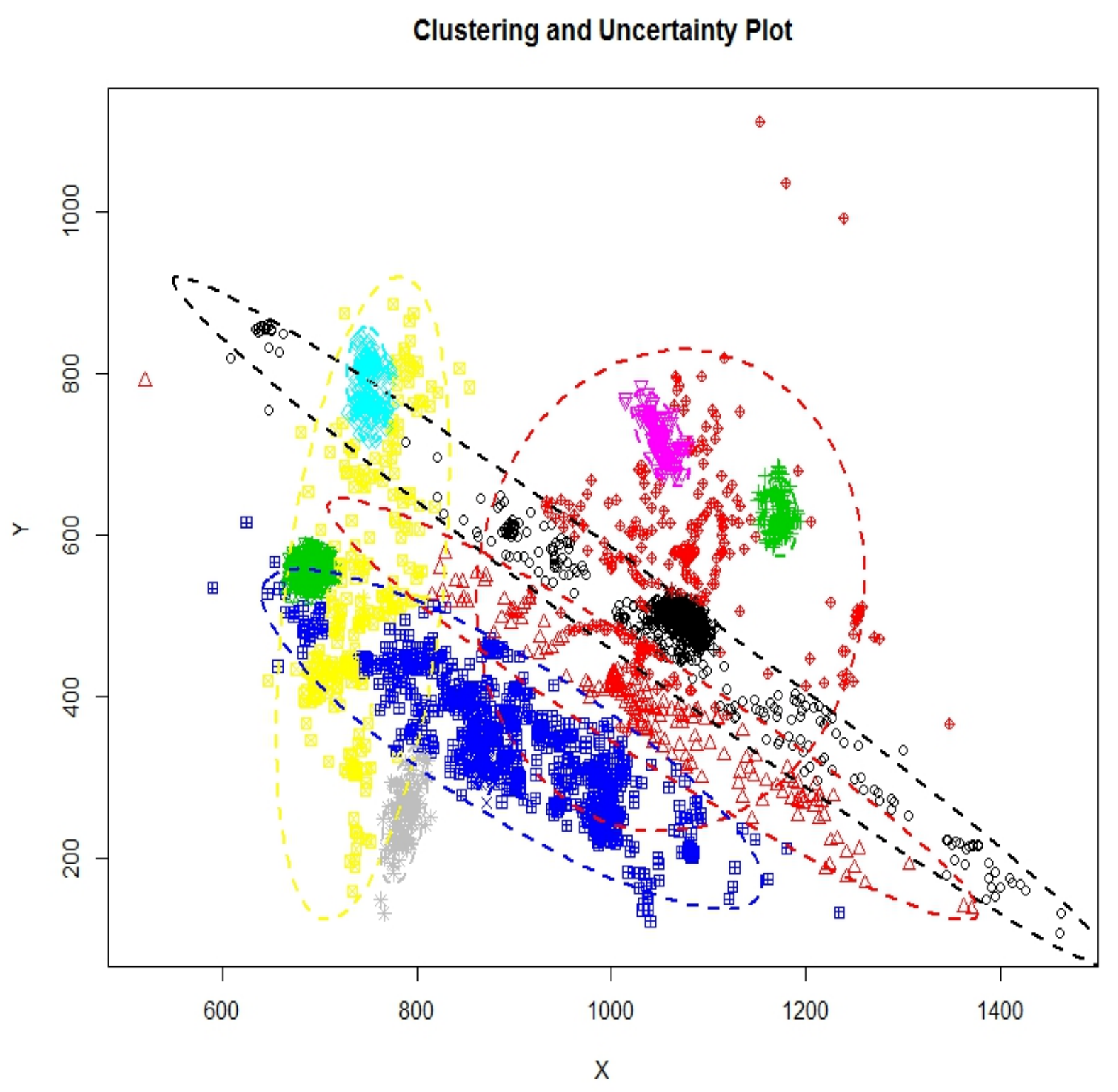

Figure: Clustering analysis of eyetracking data using a GMM fit to longitudinal data. The spatial locations of the subject's eye fixation location is divided into 12 clusters, based on the BIC values of the models tested. The ellipses represent the uncertainty of the clustering performance.

Benefits

- Allow us to gauge the significant improvement in clustering performance and uncertainty that correctly factoring in the temporal correlation between observations can bring.
- R package allows for methods to 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.
- This capability can be extended to other types of probabilistic clustering models. It is possible that alternate models are needed to better fit other types of distributions of data.
- Lead to the development of *gaze-informed foraging models* to understand how imagery analysts become efficient in navigating and detecting key event signatures in large, noisy geospatial datasets.

Contact: mgchen@sandia.gov

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

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.