Curve Clustering with Random Effects Regression Mixtures

Scott J. Gaffney
Information and Computer Science
University of California, Irvine
sgaffney@ics.uci.edu

Padhraic Smyth
Information and Computer Science
University of California, Irvine
smyth@ics.uci.edu

Abstract

In this paper we address the problem of clustering sets of curve or trajectory data generated by groups of objects or individuals. The focus is to model curve data directly using a set of model-based curve clustering algorithms referred to as mixtures of regressions or regression mixtures. The proposed methodology is based on extension to regression mixtures that we call random effects regression mixtures which combines linear random effects models with standard regression mixtures. We develop a general expectation-maximization (EM) algorithm using maximum a posteriori (MAP) estimation for random effects regression mixtures and demonstrate how this technique can be applied to the problem of clustering cyclone data.

1 Introduction

Clustering is often used as a general tool for understanding and exploring large data sets. Most clustering algorithms operate on *feature vectors* of fixed dimension. In contrast in this paper we address the problem of clustering sets of curves or trajectories of variable length, generated by groups of objects or individuals. The curves $\mathbf{y}$ are sequences of observations measured over time (or some notion of time), functionally dependent on an independent variable or set of variables $\mathbf{x}$. Typically, $\mathbf{x}$ is itself time, but in general it can be any number of variables measured over the same interval as $\mathbf{y}$. Unlike fixed-dimensional feature vectors, the $\mathbf{y}$ curves can have variable lengths, can be observed at different measurement intervals, as well as contain missing observations.

This type of data is quite common in various scientific contexts. For example, Figure 1 shows a set of trajectory data from the atmospheric sciences. In this figure

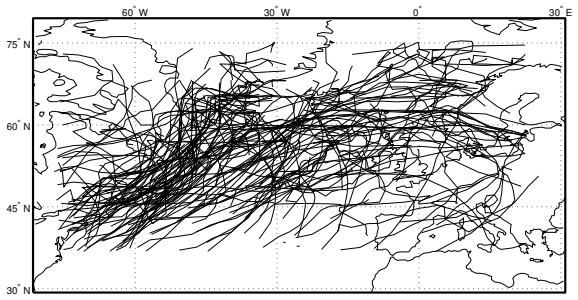

Figure 1: Some cyclone trajectories tracked over the North Atlantic.

we see a number of Extra Tropical Cyclones (ETCs) that were tracked over the North Atlantic in the winter months (November to April) from 1980 to 1995 (Gaffney, Robertson, & Smyth, 2001). The x-axis is longitude and the y-axis is latitude. Although the direction of movement is not explicitly shown, most ETCs tend to move from west to east. The curve data here is multidimensional with respect to time (i.e., there is a two dimensional lat-lon observation vector at each time). Research questions of interest to atmospheric scientists include the finding of evidence for the existence of clusters of cyclones, characterizing the component behaviors within these clusters, determining the relation of these clusters to other climate phenomena such as precipitation or global pressure patterns, and prediction of the most likely trajectories for new cyclones given both their trajectory up to some time point and the learned model of cyclone behavior (e.g., see Blender et al., 1997).

This type of curve data cannot be clustered with standard vector-based clustering algorithms without resorting to some *ad hoc* preprocessing procedure to reduce the data to a set of fixed-dimensional feature vectors. Our focus in this paper is to model curve data directly using a set of model-based *curve* clustering algorithms referred to as *mixtures of regressions* or *regression mixtures* (DeSarbo and Cron, 1988; Gaffney

and Smyth, 1999).

Model-based curve clustering has some inherent advantages compared to non-probabilistic clustering techniques. As well as providing a generative model for the data, it can easily handle missing and irregularly spaced measurements, it can be directly generalized to two-, three-, or higher-dimensional curves, it can explicitly model trajectory smoothness as a function of time, and the model can be evaluated out-of-sample in terms of predictive power.

Our proposed methodology concerns an extension to regression mixtures that we shall call *random effects* regression mixtures which integrates linear random effects models (Laird and Ware, 1982) with standard regression mixtures. Random effects mixture models were first introduced in the fully Bayesian setting (Lenk and DeSarbo, 2000) in which the mixture components were allowed to be generalized linear models (McCullagh and Nelder, 1983). In contrast, we propose the development of an explicit maximum a posteriori or MAP-based EM algorithm for random effects *regression* mixtures. This approach avoids the use of Markov Chain Monte Carlo (MCMC) techniques (Gelfand and Smith, 1990) used by Lenk and DeSarbo (2000) in order to perform inference.

After we present our new methodology, we show its application to the clustering of cyclones from an atmospheric dataset and show that it outperforms standard regression mixtures.

2 Prior Work

The earliest works on regression-based mixtures focused on the definition of a simple two-component mixture likelihood using various methods (e.g., conjugate gradient descent) to estimate parameter values (Quandt, 1972; Quandt and Ramsey, 1978; Hosmer, 1974). Of particular note is the work of Hosmer (1974) who, in retrospect, developed an EM algorithm (Dempster et al., 1977) for the simple two-component case. Späth (1979), on the other hand, designed a non-probabilistic algorithm called *clusterwise linear regression* similar to K-means (Hartigan and Wong, 1978) that found the solution to K different regression equations over a single dataset.

Later the general K-cluster case employing EM was developed (DeSarbo and Cron, 1988; Jones and McLachlan, 1992; Gaffney and Smyth, 1999) and extended to various situations including the use of curve data and to non-parametric kernel regression models. General random effects mixtures in the fully Bayesian context were introduced by Lenk and DeSarbo (2000). They focused on full Bayesian inference for mixtures of generalized linear models with random effects, using MCMC techniques.

A closely related model family is that of Hierarchical Mixtures of Experts (HME; Jordan and Jacobs, 1994), which share some similar features to mixtures of regressions in that they define gating networks that contain mixtures of generalized linear models. Although an HME can be mathematically similar to some regression mixture models, the focus of HMEs is on supervised learning and not on unsupervised learning.

3 Generative Model-Based Curve Clustering

3.1 Model-Based Clustering

Model-based clustering is a probabilistic technique that assumes the data set can be explained by a finite mixture of group-specific density components (Banfield and Raftery, 1993; Fraley and Raftery, 1998). In the standard setup, we model multivariate $\mathbf{x}$ by the mixture density

$$p(\mathbf{x}|\theta) = \sum_{k}^{K} \alpha_k p_k(\mathbf{x}|\theta_k), \tag{1}$$

where p_k is the component density with parameters θ_k for the kth cluster, and α_k is the unconditional probability that $\mathbf{x}$ was generated by cluster k. It is possible to learn the parameters θ of many mixture models using standard EM-based algorithms (Dempster et al., 1977; McLachlan and Krishnan, 1997).

Model-based *curve* clustering is based on a mixture model similar to the above mixture; however, the component models $p_k(\mathbf{x}|\theta_k)$ are replaced by conditional densities $p_k(\mathbf{y}|\mathbf{x}, \theta_k)$ so that the resulting mixture density is conditional. In general, the curves $\mathbf{y}$ are sequences of observations measured over time (or some notion of time), functionally dependent on an independent variable or set of variables $\mathbf{x}$. The curve clustering model

$$p(\mathbf{y}|\mathbf{x}, \theta) = \sum_{k}^{K} \alpha_k p_k(\mathbf{y}|\mathbf{x}, \theta_k), \tag{2}$$

is completed with the specification of the conditional density components.

3.2 Standard Regression Mixtures

Standard regression mixtures results from placing regression components into the curve clustering model. This allows the modelling of a set of potentially heterogeneous curves by a mixture of K regression equations. For example, assume we have a p-th order regression

relationship between $\mathbf{y}$ and $\mathbf{x}$ as

$$\mathbf{y} = \mathbf{X}\beta_k + \epsilon_i, \qquad \epsilon_i \sim N(\mathbf{0}, \sigma_k^2\mathbf{I}), \tag{3}$$

in which $\mathbf{X}$ is the usual $n \times (p+1)$ regression matrix containing an initial column of ones, β_k are the regression coefficients for the k-th cluster, and $\sigma_k^2\mathbf{I}$ is the k-th covariance matrix. This is equivalent to setting $p_k(\mathbf{y}|\mathbf{x}, \theta_k)$ in the curve clustering model equal to $N(\mathbf{y}|\mathbf{X}\beta_k, \sigma_k^2\mathbf{I})$, where N is the conditional multivariate normal density with mean $\mathbf{X}\beta_k$ and covariance $\sigma_k^2\mathbf{I}$. A detailed description of this model and the accompanying EM algorithm for trajectory data is given in (Gaffney and Smyth, 1999).

4 Random Effects Regression Mixtures

In the standard regression mixtures framework it is assumed that each cluster is modelled with cluster-specific parameters θ_k and that the set of curves are defined as a mixture over these clusters. This can be used to effectively account for subpopulations of *homogeneous* behavior. However, more care should be taken when considerable *heterogeneity* exists within each subpopulation or group.

For example suppose we have a set of individuals from K groups. Using the standard regression mixtures framework, we would hypothesize that within each group all individuals are sufficiently homogeneous to appropriately fit the common group component model. However, in the presence of significant heterogeneity within any group, one would have to resort to fitting more groups than the known K to be able to sufficiently describe the data.

What we want is the ability to let an individual vary from the template for its group, yet still exhibit the underlying behavior that distinguishes this group from the rest. Linear random effects models (Laird and Ware, 1982) allow individuals to vary from the population mean by an individual-specific random effects term. Lenk and DeSarbo (2000) proposed the integration of this idea with mixtures of generalized linear models in which each individual varied from each cluster mean.

We focus on the specific development of *random effects regression mixtures* in which we define a hierarchical model structure with a mixture on parameters at the top level (*parameter-level*) and a simple individual-specific regression model at the bottom level (*data-level*). We then define an EM algorithm using MAP estimation to enable inference in the hierarchy.

4.1 Hierarchical Model Structure

Let us assume we have a set of n individuals from K groups and that each individual i generates a trajectory of measurements $\mathbf{y}_i$ of length n_i according to the normal regression model

$$\mathbf{y}_i = \mathbf{X}_i\beta_i + \epsilon_i, \qquad \epsilon_i \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{4}$$

which gives conditional density

$$p(\mathbf{y}_i|\mathbf{X}_i, \beta_i, \sigma^2) = f(\mathbf{y}_i|\mathbf{X}_i\beta_i, \sigma^2\mathbf{I}), \tag{5}$$

where $\mathbf{X}_i$ and β_i are as in (3), with f the conditional multivariate normal density with mean $\mathbf{X}_i\beta_i$ and covariance $\sigma^2\mathbf{I}$. Notice that each individual has its own regression model through the parameter β_i (i.e., we do not have β_k here). This is the random effect. In fact, there is no dependence on group membership at all at this level, the bottom-level (or *data-level*) of the hierarchy. Instead at this level we allow for individual-specific heterogeneity.

At the top-level of the hierarchy we have a probabilistic model that describes the distribution of the parameters β_i for each individual. Suppose we let k_i give the group membership for the ith individual. Knowledge of membership allows us to define a distribution on β_i according to the group template as

$$p(\beta_i|k_i, \phi_{k_i}) = g_{k_i}(\beta_i|\mu_{k_i}, \mathbf{R}_{k_i}), \qquad \phi_{k_i} = \{\mu_{k_i}, \mathbf{R}_{k_i}\},$$

where g_{k_i} is the multivariate normal density with mean μ_{k_i} and covariance $\mathbf{R}_{k_i}$. Unconditional of class membership, the prior for β_i,

$$p(\beta_i|\Phi) = \sum_{k_i=1}^{K} \alpha_{k_i} g_{k_i}(\beta_i|\mu_{k_i}, \mathbf{R}_{k_i}), \tag{6}$$

is a finite mixture with $\Phi = \{\alpha_1, \ldots, \alpha_K, \phi_1, \ldots, \phi_K\}$. At this level of the hierarchy we allow for the clustering of homogeneous group behavior. As a result we now have a finite mixture model allowing for homogeneous group behavior at the top-level, and a simple regression model allowing for individual heterogeneity at the bottom-level.

One issue with this model is that we are now trying to estimate K distinct covariance matrices which may be problematic. One solution would be to pool the K covariance matrices into a single representative matrix $\mathbf{R}$. Banfield and Raftery (1993) introduce a number of methods to reparameterize covariance matrices so that instead of all clusters sharing a single $\mathbf{R}$, they only share certain chosen characteristics (e.g., orientation, size, or shape).

In addition, one can also introduce a Bayesian regularization methodology to the framework to curb problematic estimations. We define hyperpriors for $\mathbf{R}_k$ and

α_k in this regard. The standard conjugate priors for $\mathbf{R}_k^{-1}$ and $\alpha = (\alpha_1, \ldots, \alpha_K)'$ are the Wishart density $W(\mathbf{R}_k^{-1}|\mathbf{R}_0, \nu)$ and the Dirichlet density $D(\alpha|\eta)$ (Buntine, 1994; Gelman et al., 1995; Ormoneit and Tresp, 1995). We complete the model by assuming a simple non-informative prior for both σ^2 and μ_k.

4.2 MAP Estimation

This hierarchical model specification naturally leads to MAP instead of ML (maximum likelihood) estimation. That is, it is natural to define the posterior of the parameters given the data as being proportional to the likelihood of the bottom-level times the prior of the top-level. Let $\Theta = \{\beta_1, \ldots, \beta_n, \sigma^2\}$ be the parameters at the bottom-level and let Φ be the parameters at the top-level. Then we define our MAP objective function $\mathcal{M}$ to be proportional to the posterior $p(\Theta, \Phi|\mathbf{Y})$ of the parameters (note that $\mathbf{X}$ is left out for simplicity). Thus, we define

$$\begin{aligned}\mathcal{M}(\Theta, \Phi) &= \log\left[p(\boldsymbol{Y}|\boldsymbol{X}, \Theta, \Phi)p(\Theta, \Phi)\right] \\ &= \log\left[p(\boldsymbol{Y}|\boldsymbol{X}, \Theta)p(\Theta|\Phi)p(\Phi)\right],\end{aligned}$$

as our MAP objective function for the parameters Θ and Φ, where

$$p(\boldsymbol{Y}|\boldsymbol{X}, \Theta) = \prod_i f(\mathbf{y}_i|\mathbf{X}_i\beta_i, \sigma^2\mathbf{I}),$$

$$p(\Theta|\Phi) = \prod_i \sum_k^K \alpha_k g_k(\beta_i|\mu_k, \mathbf{R}_k),$$

and

$$p(\Phi) = D(\alpha|\eta)\prod_k W(\mathbf{R}_k^{-1}|\mathbf{R}_0, \nu).$$

4.3 MAP-based EM Algorithm

Analysis of $\mathcal{M}$ in our setting leads to the conclusion that direct maximization is not feasible. However, we can develop a MAP-based EM algorithm that will produce consistent parameter estimates.

In the *random effects regression mixture* model there is a notion that each individual is chosen from, or is generated from, one of K different groups. Usually we are not given the group memberships along with the data $\boldsymbol{X}$,$\boldsymbol{Y}$. The group memberships are instead *hidden* or *missing*. The EM algorithm is an iterative method that deals with these so-called *missing data problems*. In fact there are two sources of hidden data in a random effects regression model. First, the memberships are hidden but also the individual-specific regression coefficients β_i are also considered hidden and not observable.

4.4 Complete-Data Function

In the EM framework, the function $\mathcal{M}$ is referred to as the *incomplete-data function* since it does not contain all the missing data. It is the missing data that makes the problem complex. Therefore, to make the problem easier we simply define another function that *does* contain the missing data. Suppose we define $\boldsymbol{Z}$ to be the set consisting of memberships k_i for all individuals i and notate the set of all unobservable β_i as β. Then we define the *complete-data* MAP objective function of σ^2 and Φ as

$$\begin{aligned}\mathcal{M}_C(\sigma^2, \Phi) &= \log\left[p(\boldsymbol{Y}, \boldsymbol{Z}, \beta|\boldsymbol{X}, \sigma^2, \Phi)p(\Phi)\right] \\ &= \log\left[p(\boldsymbol{Y}|\boldsymbol{X}, \Theta)p(\Theta, \boldsymbol{Z}|\Phi)p(\Phi),\right]\end{aligned}$$

with

$$p(\boldsymbol{Y}|\boldsymbol{X}, \Theta) = \prod_i f(\mathbf{y}_i|\mathbf{X}_i\beta_i, \sigma^2\mathbf{I}),$$

$$p(\Theta, \boldsymbol{Z}|\Phi) = \prod_i \alpha_{k_i} g_{k_i}(\beta_i|\mu_{k_i}, \mathbf{R}_{k_i}),$$

and

$$p(\Phi) = D(\alpha|\eta)\prod_k W(\mathbf{R}_k^{-1}|\mathbf{R}_0, \nu).$$

Notice that we have gotten rid of the logarithm of the summation due to the move from $p(\Theta|\Phi)$ to $p(\Theta, \boldsymbol{Z}|\Phi)$. Of course, we don't know the true values for $\boldsymbol{Z}$ and β so we take expectations with respect to their joint posterior distribution as we shall see next.

4.5 EM Solutions

The EM algorithm consists of two steps: (1) the expected value of $\mathcal{M}_C$ is taken with respect to the posterior *hidden distribution* $p(\boldsymbol{Z}, \beta|\boldsymbol{Y})$, and (2) this expectation is maximized over the parameters σ and Φ to yield the new parameter values.

4.5.1 E-Step

In the E-step, the expectation of $\mathcal{M}_C$ is taken with respect to $p(\boldsymbol{Z}, \beta|\boldsymbol{Y})$ which factors into $p(\boldsymbol{Z}|\boldsymbol{Y})p(\beta|\boldsymbol{Z}, \boldsymbol{Y})$. In this step we simply need to calculate two things: the *membership* probability $p(z_i = k|\mathbf{y}_i)$,and the expected value of the posterior $p(\beta_i|z_i, \mathbf{y}_i)$. First we calculate the *membership* probability

$$\begin{aligned}w_{ik} &= p(z_i = k|\mathbf{y}_i) \\ &\propto \alpha_k p(\mathbf{y}_i|\sigma^2, \phi_k) \qquad (7)\end{aligned}$$

that curve i was generated from cluster k. Note that we are not given β_i in $p(\mathbf{y}_i|\sigma^2, \phi_k)$; this is the marginal

model of $\mathbf{y}_i$. And second we set the expected value of β_i given $\mathbf{y}_i$ and k_i to

$$\hat{\beta}_{ik} = (1/\sigma^2 \mathbf{X}_i' \mathbf{X}_i + \mathbf{R}_k^{-1})^{-1} (1/\sigma^2 \mathbf{X}_i' \mathbf{y}_i + \mathbf{R}_k^{-1} \mu_k)$$

which is the mean of the posterior $p(\beta_i | \mathbf{y}_i, k_i)$. Note that $\hat{\beta}_{ik}$ is simply the result of Bayesian regression with prior μ_k. Also, for simplicity, we set

$$\mathbf{V}_{\hat{\beta}_{ik}} = (1/\sigma^2 \mathbf{X}_i' \mathbf{X}_i + \mathbf{R}_k^{-1})^{-1}$$

which gives the posterior covariance.

4.5.2 M-Step

In the M-step we use w_{ik}, $\hat{\beta}_{ik}$, and $\mathbf{V}_{\hat{\beta}_{ik}}$ from the E-step to update the model parameters. First we maximize the top-level (the mixture model on parameters), and then we maximize the bottom-level (the regression model on y and x data). For the top-level we update the parameters

$$\hat{\alpha}_k = \frac{\sum_i^n w_{ik} + (\eta_k - 1)}{n + (\sum_k \eta_k - K)},$$

$$\hat{\mu}_k = \frac{\sum_i^n w_{ik} \hat{\beta}_{ik}}{\sum_i^n w_{ik}},$$

and

$$\hat{\mathbf{R}}_k = \frac{\sum_i^n w_{ik} \left[\| \hat{\beta}_{ik} - \hat{\mu}_k \|^2 + \mathbf{V}_{\hat{\beta}_{ik}} \right] + \mathbf{R}_0^{-1}}{\sum_i^n w_{ik} + (\nu - (p+1))}, \quad (8)$$

while on the bottom-level we update the parameter

$$\hat{\sigma^2} = \frac{\sum_{ik} w_{ik} \left[\| \mathbf{y}_i - \mathbf{X}_i \hat{\beta}_{ik}) \|^2 + \mathbf{V}_{\hat{\beta}_{ik}} \right]}{N}.$$

There is a small issue of setting the hyperparameters for the hyperprior $p(\Phi)$. One can set ν to the neutral value of $p+1$ which then cancels in the denominator of (8) as well as set $\mathbf{R}_0^{-1}$ to $\omega \mathbf{I}$ for some positive ω. In this way, ω acts as a type of smoothing parameter. One can also set the Dirichlet to neutral values (e.g., $\eta_1 = \cdots = \eta_k = 1$), or it can be used to deal with issues such as background clusters. In this case, you may want to enforce a rule that for every 100 curves, there "should" be at least one in the background.

5 Cyclone Clustering

The primary application of random effects regression mixtures that we have investigated up to this point is the clustering of ETC (Extra-Tropical Cyclone) tracks from meteorological data. Atmospheric scientists are interested in the spatio-temporal patterns of evolution of ETCs for a number of reasons. For example, it is not well-understood how long-term climate changes (such as global warming) may influence ETC frequency, strength, occurrence and spatial distribution. Similarly, changes in ETC patterns may provide clues of long-term changes in the climatic processes that drive ETCs. The links between ETCs and local weather phenomena are also of interest: clearly ETCs have significant influence on local precipitation, and in this context, a better understanding of their dynamics could provide better forecasting techniques both on local and seasonal time-scales.

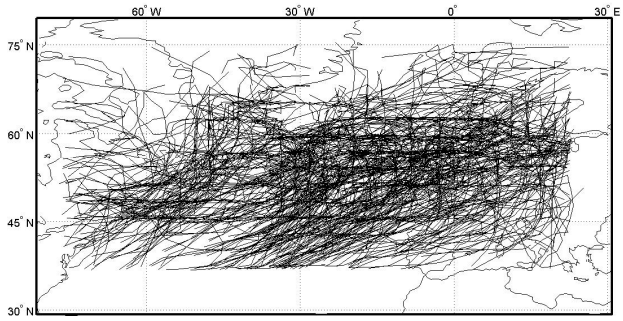

Figure 2: The full set of cyclone trajectories.

5.1 Prior Work on Clustering ETC Trajectories

The work of Blender et al. (1997) is illustrative of the use of conventional clustering techniques in atmospheric science. Using sea-level pressure data on a grid over the North Atlantic (measurements every 6 hours, available over several winters) they detect local minima in the pressure map and then use a nearest-neighbor tracking algorithm to connect up the minima in successive maps and determine trajectories. The trajectories are then converted into a fixed-dimensional vector for clustering by the K-means algorithm. Based on subjective analysis of the data, $K = 3$ clusters are chosen and fit in the resulting fixed-dimensional space. Despite the somewhat ad hoc nature of the approach the resulting clusters demonstrate that storms in the North Atlantic clearly cluster into different types of trajectory paths.

5.2 Experimental Results w/ Random Effects Regression Mixtures

In this section we describe some experimental results with cyclone clustering using our random effects regression mixtures framework. The datasets that we are working with are widely used simulation data sets known as general circulation model data (see Gaffney, Robertson, and Smyth, 2001, for further details). Specifically, we have data for the winter months (November to April) from 1980 to 1995 that give

mean sea-level pressure (MSLP) measurements on a $2.5° \times 2.5°$ grid over the earth every 6hrs. Since we are interested in ETCs over the North Atlantic we focus only on the area between 30°N-70°N and 80°W-10°E.

This "raw" data is taken and cyclones are detected and tracked over space and time in a similar manner as that described in Section 5.1. Essentially static detection of relative pressure minima at each time is followed by a simple nearest-neighbor-based tracking algorithm to associate the minima over time to form trajectories (representing cyclones). Further details about the raw data and the detection and tracking of cyclones can be found in (Gaffney, Robertson, & Smyth, 2001).

In Figure 2 we see the resulting set of all 614 tracked cyclones. The cyclones have varying durations but all have a minimum of 10 observations (this is due to the definition of cyclones in the tracking algorithm). We take this set of curves as input to our algorithm.

Looking at the bottom-level of our proposed hierarchical model in (4), we take $\mathbf{y}_i$ equal to the $n_i \times 2$ matrix of latitude-longitude positions for the i-th cyclone of length n_i. All cyclones are "zeroed" so that $\mathbf{y}_i$ begins at the *relative* latitude-longitude position of $(0, 0)$. This allows a clustering on the basis of shape and eliminates initial starting position as a source of variation. Furthermore, we employ a quadratic polynomial fit in the regression model and thus we set $\mathbf{X}_i$ to the $n_i \times 3$ matrix consisting of an initial column of ones followed by a column of the times at which $\mathbf{y}_i$ was measured, and ending with a column of the squared values in column 2.

In the top-level of the hierarchy we have a mixture model on the resulting regression coefficients from the bottom-level. Thus, the top-level employs a clustering of cyclones based on a notion of cyclone velocity as well as direction. This means that clustered cyclones will tend to share common component velocities and will tend to move in the same latitude-longitude direction. However direction is not unique to a cluster since two cyclones can move at rather different component velocities–thus resulting in different cluster assignments–and yet still move in the same lat-lon direction (which is determined by the ratio of velocities). Therefore, analysis of the resulting clustering is quite a bit more complex than allowed by the limited space here; however, the following figures are useful nonetheless.

5.2.1 Cluster Analysis

Figures 3-5 show the three returned clusters mapped onto a projection of the earth over the North Atlantic. We set the algorithm to find $K = 3$ clusters not only for simplicity but also because a three-cluster descrip-

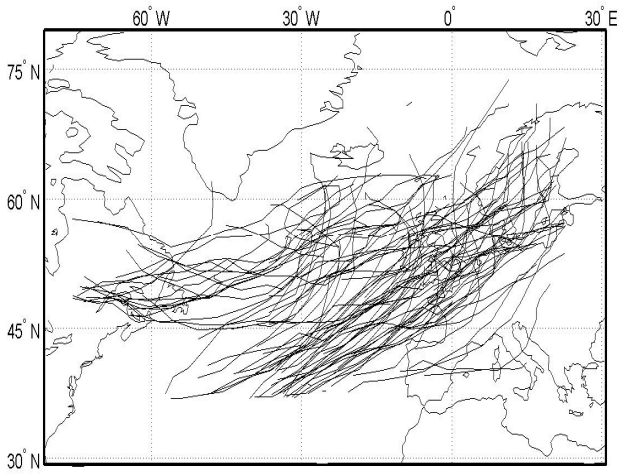

Figure 3: Diagonally oriented cluster.

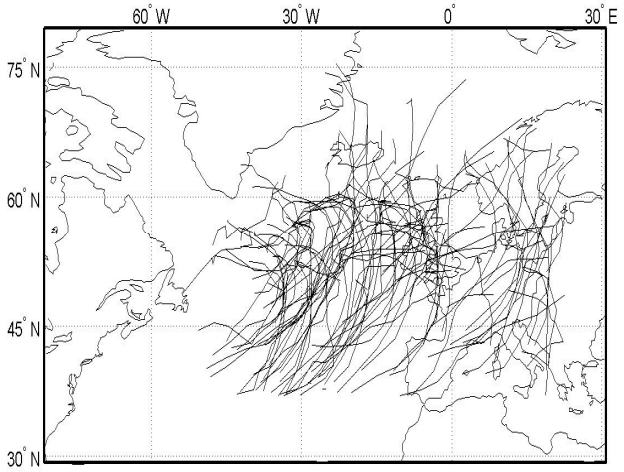

Figure 4: Vertically oriented cluster.

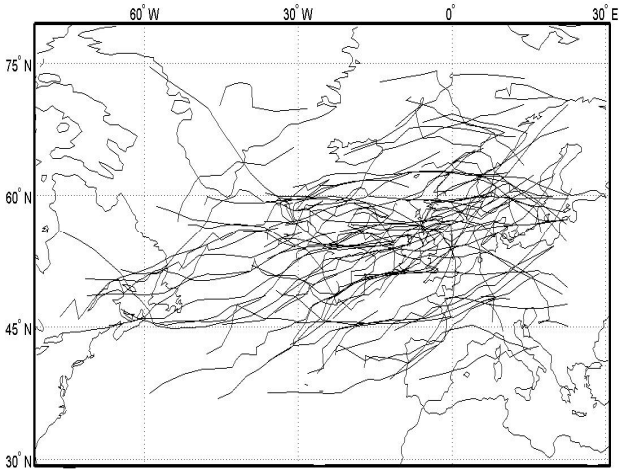

Figure 5: Horizontally oriented cluster.

tion of North Atlantic ETCs has been found useful in previous studies (e.g., Blender et al., 1997). The figures show a few randomly selected cyclones from each cluster so as to reduce the visual clutter.

In Figure 3 we see a large cluster of diagonally oriented cyclones. Further analysis also reveals that these cyclones have a larger average velocity (59 km/h) than the cyclones in the other clusters and also exhibit a larger variance around this value. In Figure 4 we see a cluster of vertically-moving and somewhat backbending cyclones. The cyclones in this cluster have an average velocity of 42 km/h and individual cyclones tend to exhibit erratic velocity change during their lifespan. Figure 5 shows the final cluster; it depicts cyclones that move horizontally into the coastline of Europe. This cluster seems to be somewhat more *directionally* noisy than the others. However, this cluster has a somewhat larger average velocity (44 km/h) than the vertical-moving cluster and consists of many smaller duration cyclones (approximately 40% have durations less than 3.25 days as opposed to only 26–30% for the other two clusters). This cluster is of particular interest to atmospheric scientists since this cluster contains many of the storms that head straight into the European coastline and thus have a potential to cause much damage.

Although we see a similar overall picture here as that found by Blender et al. (1997) using their K-means-based method, their approach is unable to handle cyclones of varying durations due to their fixed-dimensional vector space and is not able to use the smoothness information inherent in trajectories so as to better guide the clustering. Furthermore, it should be noted that the cyclone tracks are unregistered or misaligned. That is, the tracking algorithm may pick up one cyclone too early or too late, or the cyclone might have been merged into another larger cyclone that was separately tracked. In any case, the problem is the same: the curves are misaligned. By adding the random effect in the mixture model we can allow for this deficiency by letting individual parameters vary to some degree from the group mean behavior.

We used a randomized cross validation scheme known as Monte Carlo Cross Validation (MCCV; Shao, 1993) to evaluate how our proposed methodology compares to both the fixed-dimensional setup as in Blender et al. (1997) as well as the standard regression mixtures setup in which there is no random effect. Since K-means is not a probabilistic method, one can simply use multivariate Gaussian mixtures as a proxy for K-means and simply truncate all cyclones to a common minimum duration in conforming to the Blender et al. (1997) methodology.

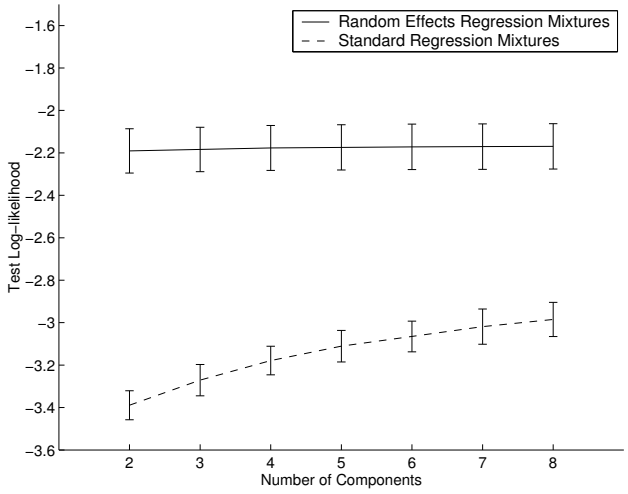

Figure 6: Cross Validation comparison of random effects regression mixtures and standard regression mixtures on the cyclone data. The error bars give two standard deviations on each side of the mean.

For this test we trained each clustering method on 70% of the cyclone tracks and then calculated the test log-likelihood on the other 30% of the data. Each regression method was run while fitting polynomials of order two (quadratic). In addition, each algorithm was allowed ten different random starts of EM. This whole procedure was then iterated fifty different times with a different random 70-30 split. The results are plotted in Figure 6.

We can see that the random effects mixture easily outperforms the non-random effects method on the cyclone dataset. The error bars give two standard deviations on each side of the mean. Not on the graph is the result for the proxy Gaussian mixtures since it is well below both of these methods in terms of test log-likelihood.

6 Conclusion

We have developed a random effects regression mixture framework and derived a MAP-based EM algorithm to perform inference. Our methodology has roots in the work of Lenk and DeSarbo (2000) in which they define mixtures of generalized linear models with random effects in the fully Bayesian setting. We depart from Lenk and DeSarbo by not requiring the use of MCMC techniques for parameter inference. This reduces the complexity of the programming required to implement the inference scheme and allows for simplified debugging due to the non-decreasing likelihood guarantee from EM theory.

An application to cyclone clustering was presented and

it was shown that the proposed methodology outperforms standard regression mixtures as well as a proxy method representing current clustering work in this area. However there are many avenues for future research with this work. For example, the problem of registration was pointed out with the cyclone data and this is a common problem with all curve-type datasets. Integration of registration into the clustering could play a key role in the final results. Also the automatic and objective identification of the number of clusters under this methodology is an important problem.

References

Banfield, J.D. and Raftery, A.E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, **49**, 803–821.

Blender, R., Fraedrich, K. & Lunkeit, F. (1997). Identification of cyclone-track regimes in the North Atlantic. *Quart J. Royal Meteor. Soc.*, **123**, 727–741.

Buntine, W. (1994). Operations for learning with graphical models. *Journal of the Artificial Intelligence Research*, **2**, 159–225.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. R. Stat. Soc.*, B, **39**, 1–38.

DeSarbo, W.S. and Cron, W.L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, **5**(1), 249–282.

Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis. *The Computer Journal*, **41**(8), 578–588

Gaffney, S., Robertson, A. & Smyth, P. (2001). Clustering of extra-tropical cyclone trajectories using mixtures of regression models. Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Fourth Workshop on Mining Scientific Datasets.

Gaffney, S. and Smyth, P. (1999). Trajectory clustering using mixtures of regression models. In *Proceedings of the ACM 1999 Conference on Knowledge Discovery and Data Mining*, S. Chaudhuri and D. Madigan (eds.), New York, NY: ACM, 63–72.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S & Rubin, D.B. (1995). *Bayesian Data Analysis*, New York, NY: Chapman & Hall.

Hartigan, J.A. and Wong, M.A. (1978). Algorithm AS 136: a K-means clustering algorithm. *Appl. Stat.*, **28**, 100–108.

Hosmer, D. (1974). Maximum likelihood estimates of the parameters of a mixture of two regression lines. *Communications in Statistics*, **3**(10), 995–1006.

Jones, P.N. and McLachlan, G.J. (1992). Fitting finite mixture models in a regression context. *Austral. J. Statist.*, **34**(2), 233–240.

Jordan, M.I. and Jacobs, R.A. (1994). Hiearchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**, 181–214.

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lenk, P.J. and DeSarbo, W.S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, **65**(1), 93–119.

McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, New York, NY: John Wiley and Sons.

McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models*, New York: Chapman and Hall.

Ormoneit, D. and Tresp, V. (1995). Improved gaussian mixture density estimates using bayesian penalty terms and network averaging. In *Advances in Neural Information Processing Systems*, D. Touretzky, M. Mozer and M. Hasselmo (eds.), New York: MIT Press, 542–548.

Quandt, R.E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association*, **67**, 306–310.

Quandt, R.E. and Ramsey, J.B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, **73**, 730–738.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**(422), 486–494.

Späth, H. (1979). Algorithm 39: clusterwise linear regression. *Computing*, **22**, 367–373.