Replicating: EM Mixtures of Regression Models (Gaffney & Smyth 1999)[1]

...[1]

Abstract

...

1 Background

The Expectation-Maximization (EM) algorithm is a widely used iterative approach to estimate parameters in statistical models with latent variables. In this project, I apply EM to group-based trajectory modeling (GBTM), where the goal is to identify clusters of polynomial trajectories from noisy observations.

Each cluster $k \in \{1, 2, 3\}$ is modeled as a second-order polynomial:

$$y = \beta_{k0} + \beta_{k1}x + \beta_{k2}x^2 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_k^2)$$

Given observed data $\{(x_{ij}, y_{ij})\}$ for person j, the EM algorithm maximizes the expected complete data log-likelihood where h_{jk} is the responsibility (soft assignment) of person j to cluster k, and X_j is the design matrix of time points for that individual.[1]

1.1 E-Step

For each trajectory j, compute the responsibility h_{jk}:

$$h_{jk} = \frac{w_k \cdot \prod_{i=1}^{n_j} \mathcal{N}(y_j(i) \mid x_j(i)^\top \beta_k, \sigma_k^2)}{\sum_{\ell=1}^{K} w_\ell \cdot \prod_{i=1}^{n_j} \mathcal{N}(y_j(i) \mid x_j(i)^\top \beta_\ell, \sigma_\ell^2)}$$

1.2 M-Step

Given the responsibilities h_{jk}, update parameters for each cluster k:

$$\beta_k = (X^\top H_k X)^{-1} X^\top H_k Y$$

$$\sigma_k^2 = \frac{(Y - X\beta_k)^\top H_k (Y - X\beta_k)}{\sum_j h_{jk}}$$

$$w_k = \frac{1}{M} \sum_{j=1}^{M} h_{jk}$$

1.3 Log-likelihood

Used for convergence checking:

$$\log L = \sum_{j=1}^{M} \log \left(\sum_{k=1}^{K} w_k \cdot \prod_{i=1}^{n_j} \mathcal{N}(y_j(i) \mid x_j(i)^\top \beta_k, \sigma_k^2) \right)$$

Symbol	Description	Dimensions
X_j	Time matrix for person j	$n_j \times K$
X	Stacked matrix of all X_j	$N \times K$
Y_j	Output for person j	$n_j \times 1$
Y	Stacked output matrix	$N \times 1$
H_k	Diagonal matrix of weights for cluster k	$N \times N$

2 Methodology

This implementation closely follows the original procedure. The dataset consists of 12 individuals, each with a trajectory generated from one of three polynomials with added Gaussian noise. The algorithm begins with randomly initialized membership probabilities h_{jk}, followed by alternating M-steps and E-steps until convergence.

In the M-step, cluster-specific regression parameters are estimated via weighted least squares:

$$\hat{\beta}_k = (X^\top H_k X)^{-1} X^\top H_k Y$$

where H_k is a diagonal matrix of weights (responsibilities) for cluster k.

To address sensitivity to initialization, I performed multiple random restarts. Small floating point values (e.g., in `np.prod`) sometimes caused instability, and rounding errors may impact reproducibility despite fixing random seeds.

3 Experiment

3.1 Results

Two experiments are shown with different noise levels:

Low noise (std = 1):

- Converged in 31 iterations with log-likelihood $\mathcal{L} = -262.11$
- Final parameters closely matched true polynomials

Higher noise (std = 5):

- Converged with log-likelihood $\mathcal{L} = -440.79$
- Parameters still approximated true functions, but with larger variances

Due to time constraints, the higher-noise experiment used a lower standard deviation than in the original paper (which used std = 10). Longer training time and hyperparameter tuning are needed for more stable results.

3.2 Discussion

This replication confirms the EM algorithm's utility in estimating parameters of polynomial trajectory mixtures, despite challenges with initialization and floating-point precision. For educational purposes, the implementation was successful.

Future improvements include:

- Handling higher noise levels more robustly
- Refactoring the code using object-oriented design

References

[1] S. Gaffney and P. Smyth., “Trajectory clustering with mixtures of regression models.” *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 1999. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/312129.312198

Appendix

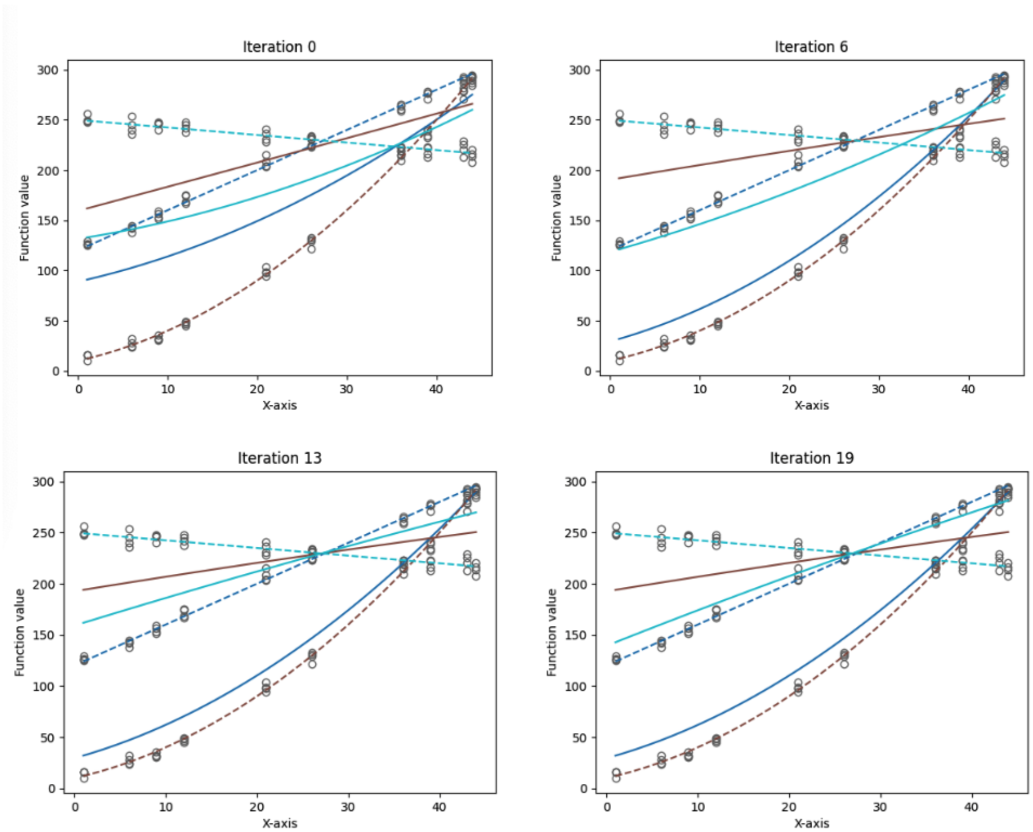

Figure 1: EM algorithm as applied to a linear regression mixture model. Both estimated trajectories (solid) and true data-generating trajectories (dotted) shown.

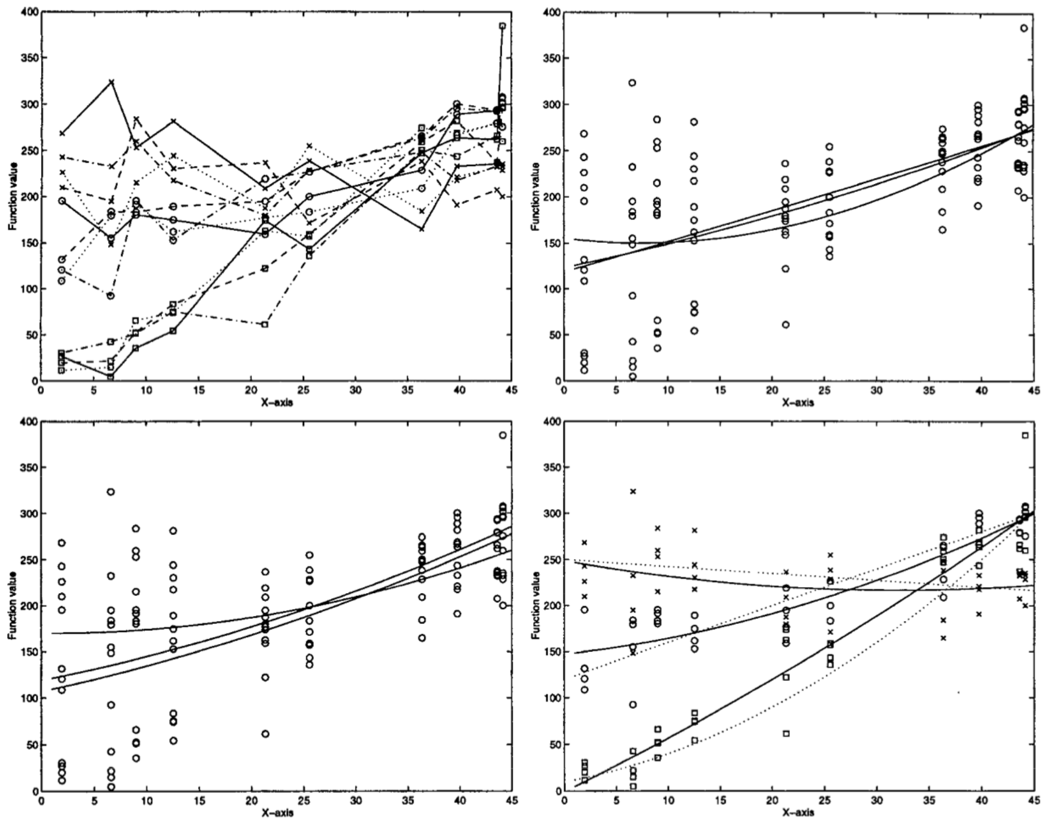

Figure 2: Original plots to replicate: "Trace of the EM algorithm as applied to a linear regression mixture model at various iterations. The upper left plot shows all of the original trajectories, the upper right shows the initial locations of the 3 cluster trajectories for EM, lower left shows the locations after 1 iteration of EM, and lower right shows the cluster locations (solid) after EM convergence (iteration 4), as well as the locations of the true data-generating trajectories (dotted)."[1]