Big Data Visual Analytics (CS 661)

Instructor: Soumya Dutta

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

email: soumyad@cse.iitk.ac.in

Study Materials for Lecture 14

- A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models
- EM Algorithm: https://stephens999.github.io/fiveMinuteStats/intro_to_em.html
- SLIC Superpixels Compared to State-of-the-Art Superpixel Methods, Achanta et al.
- Homogeneity guided probabilistic data summaries for analysis and visualization of large-scale data sets, Dutta et al.
- Statistical visualization and analysis of large data using a value-based spatial distribution, Wang et al.

Mid Sem Exam

- Saturday 22nd Feb 8-10am
- Location: L18 and L19 (OROS)
- **Please bring your institute id card (mandatory)**
- No classes during Mid Sem week
- Syllabus:
  - Everything up to today's class

Final Project Group Formation

- Form your project team by March 2nd and update the google sheet with details of project members
  - https://docs.google.com/spreadsheets/d/1ZsWmnCRK4XEZV6YezgwM0CNk1nLMaZXSIPpv1d3WvAs/edit?usp=sharing
  - **Group size: 8**
  - Those who will not be part of a team, I will help them get assigned to groups

Parametric Distribution: GMM

- Gaussian Mixture Model (GMM): Represent a probability distribution function as a convex combination of multiple Gaussian functions

$$p(X) = \sum_{i=1}^{K} \omega_i * N(X \mid \mu_i, \sigma_i)$$

ω = Weights of the Gaussian components
K = Number of Gaussian components in the mixture model

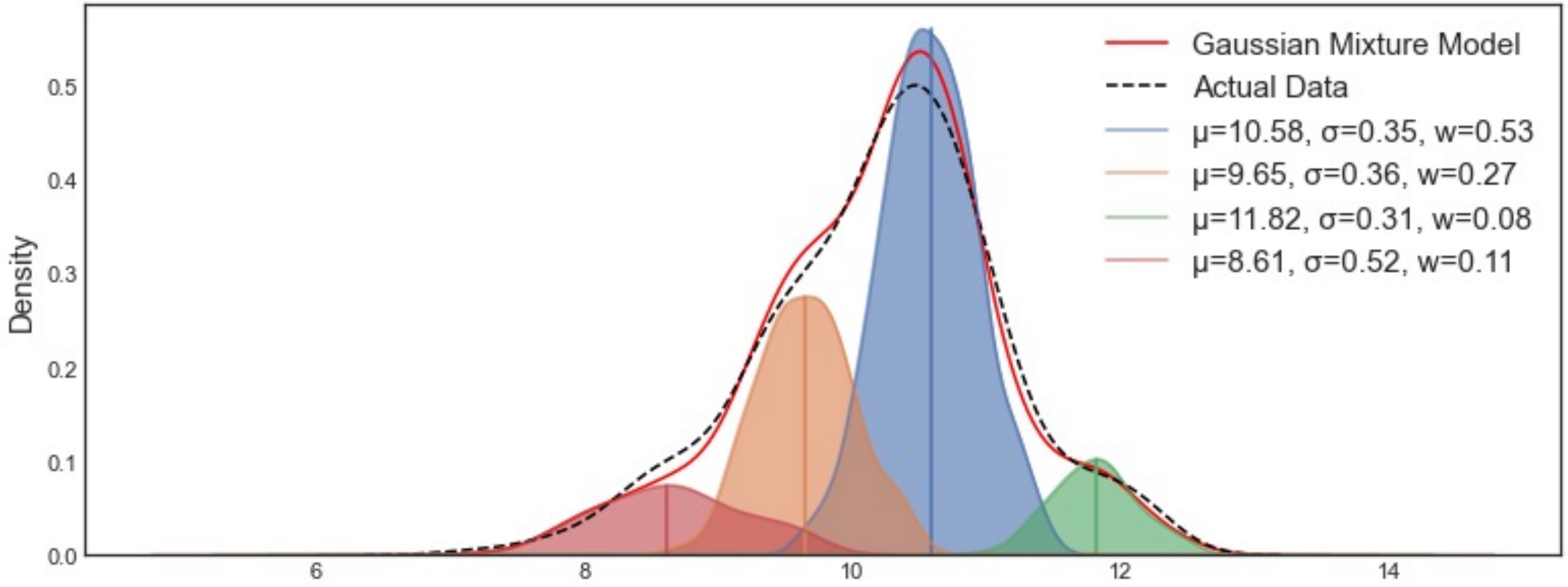


Fig. source: https://stats.stackexchange.com

Parameter Estimation Techniques

- Estimation of Gaussian distribution parameters are trivial
  - Maximum Likelihood Estimate (MLE)
  - Same as computing mean and variance

- Estimation of GMM parameters require Expectation Maximization (EM) algorithm
  - Iterative technique to fit GMM parameters

- Incremental schemes for GMM parameter estimation
  - Fast and approximate method to estimate GMM parameters
  - Can model streaming time-varying data

Expectation Maximization (EM) for GMM

- Initialize: means (μ), covariances (Σ), and weights (ω)
- Iterate until convergence:
  - **E-step:** Evaluate posterior probabilities given current parameters

$$\gamma_k^n = \frac{\omega_k \mathcal{N}(x^n | \mu_k, \Sigma_k)}{\sum_{j=1}^{k} \omega_k \mathcal{N}(x^n | \mu_j, \Sigma_j)}$$

  - **M-step:** Update the parameters to maximize the expected log-likelihood of the observed data

$$\omega_k = \frac{N_k}{N} \quad \text{and} \quad N_k = \sum_{n=1}^{N} \gamma_k^n$$

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k^n \; x^n$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k^n \, (x^n - \mu_k)\,(x^n - \mu_k)^{\mathrm{T}}$$

- Evaluate log likelihood at the end of each iteration and check for convergence

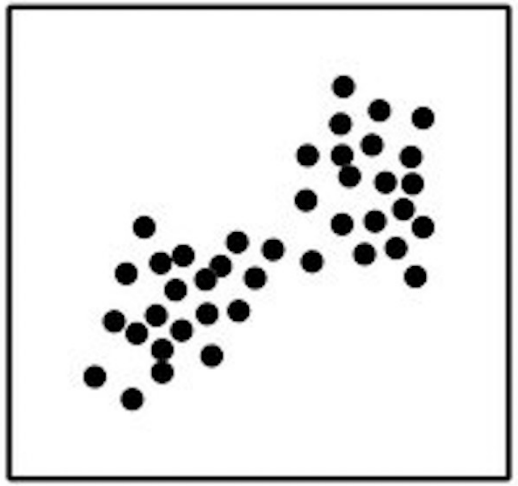
The theory of the method is much more involved!

<u>For a detailed derivation:</u>

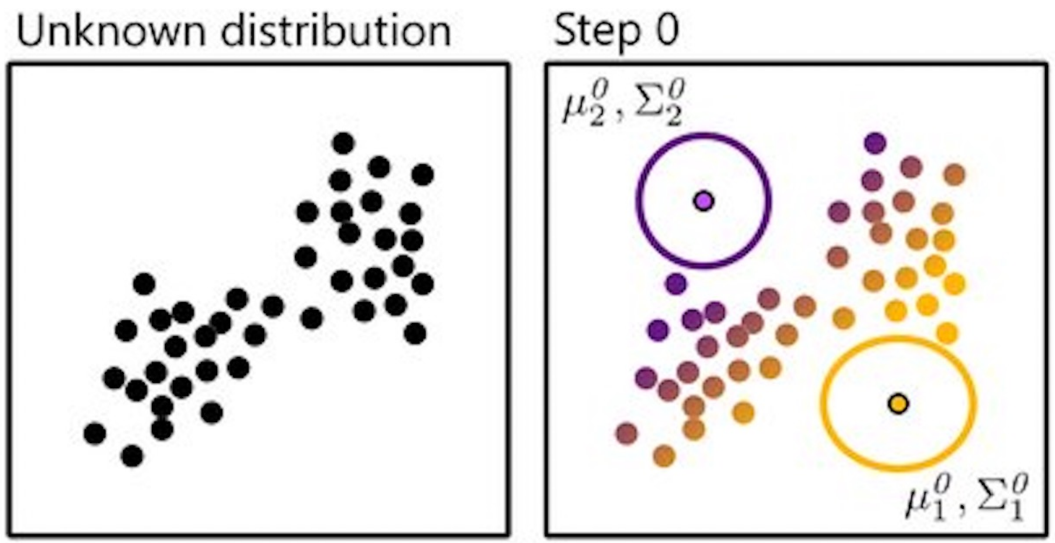
- A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models
- https://stephens999.github.io/fiveMinuteStats/intro_to_em.html

How EM Algorithm Works

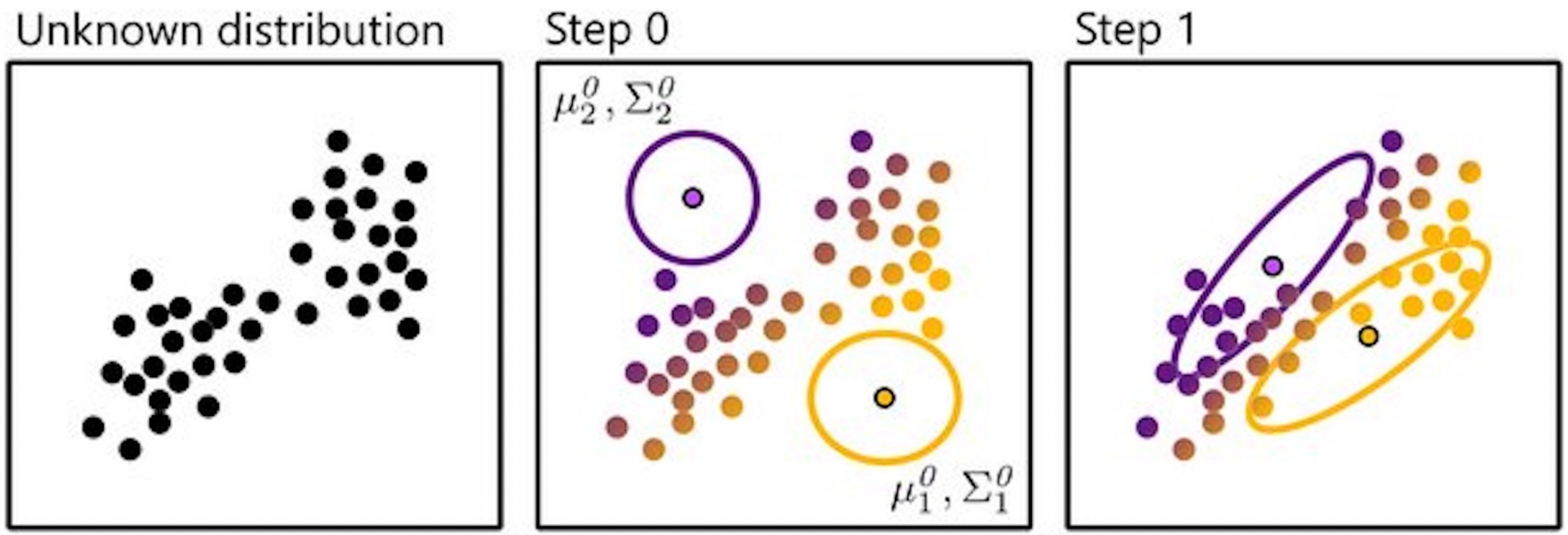
Unknown distribution



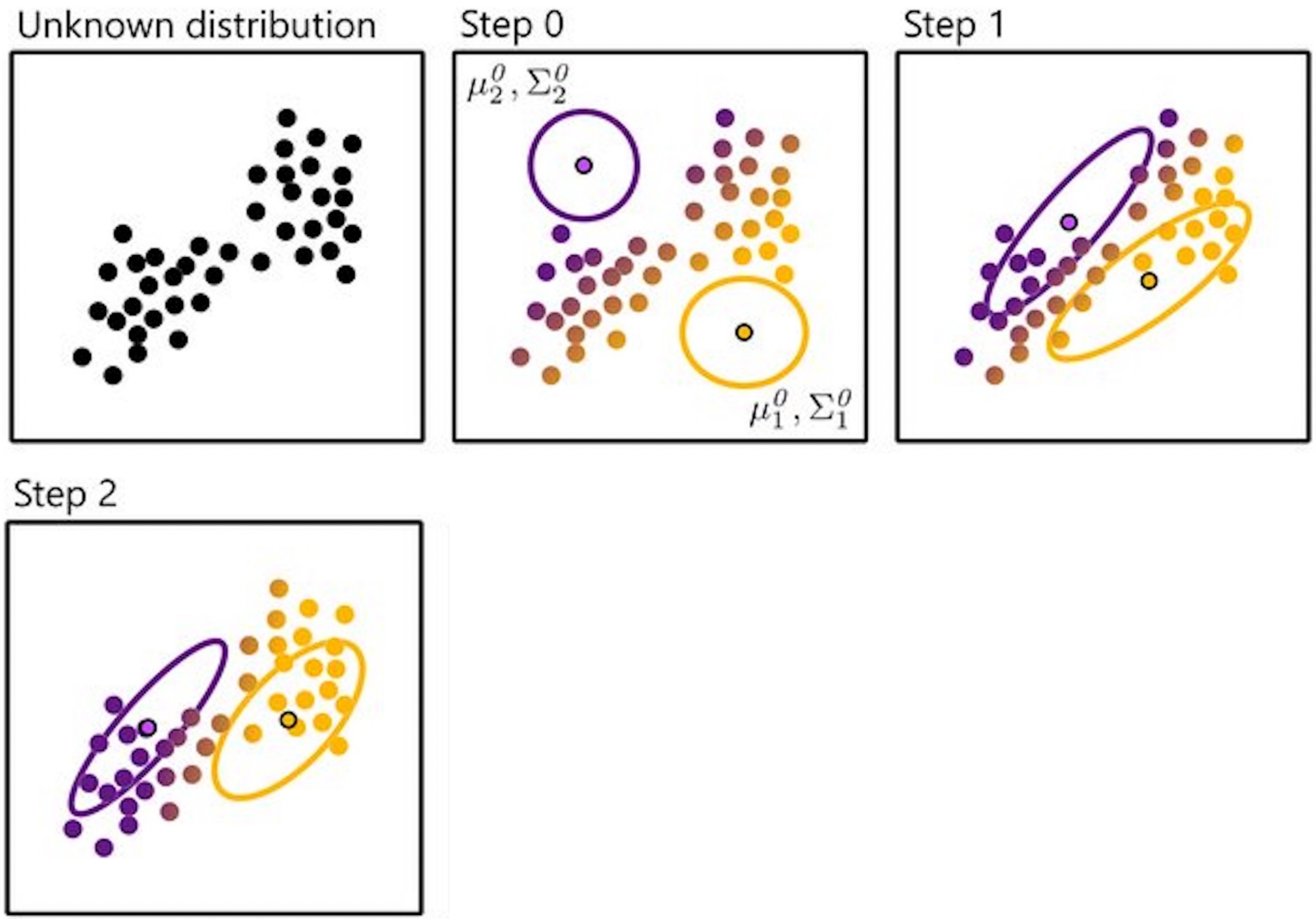
How EM Algorithm Works



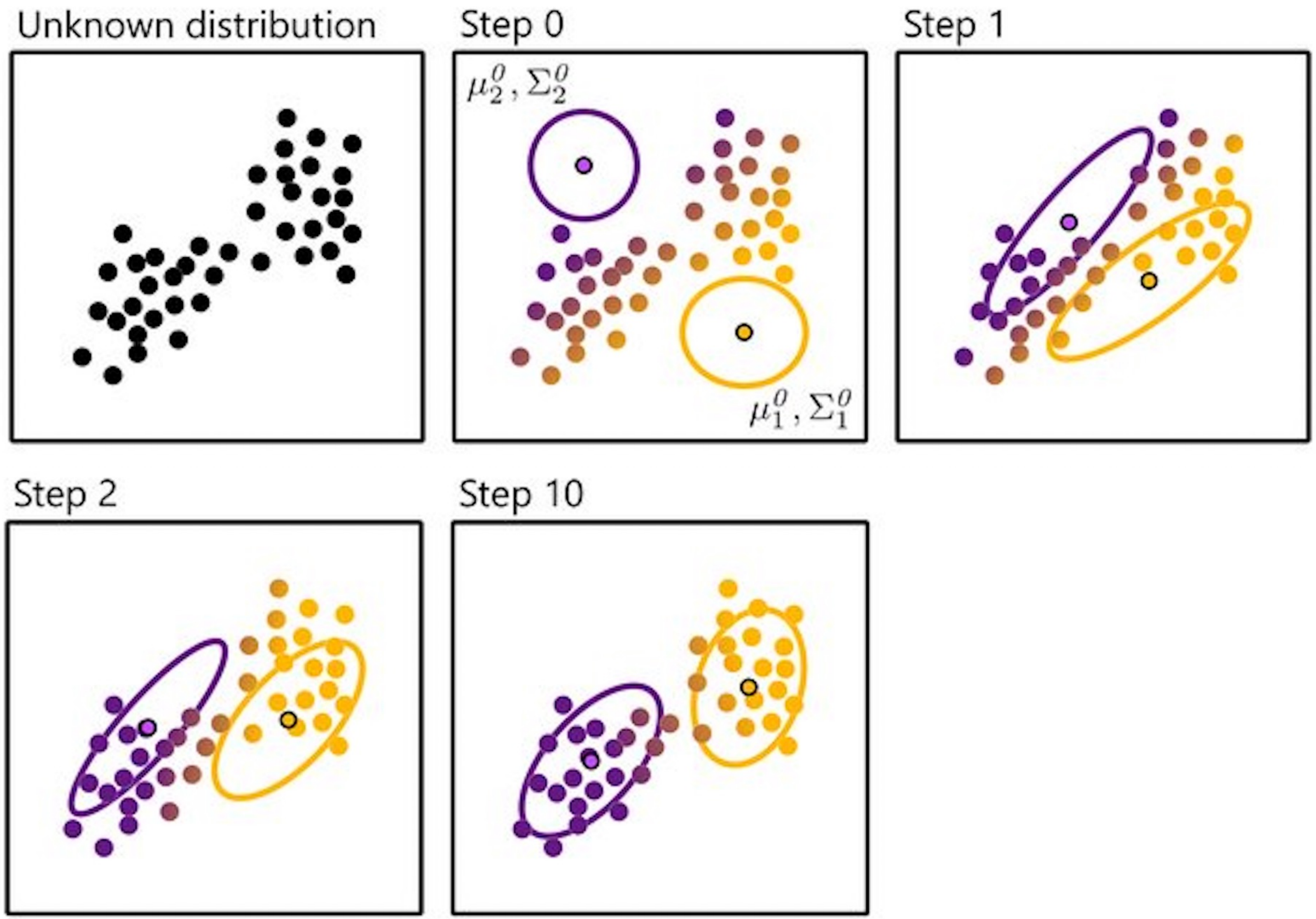
How EM Algorithm Works



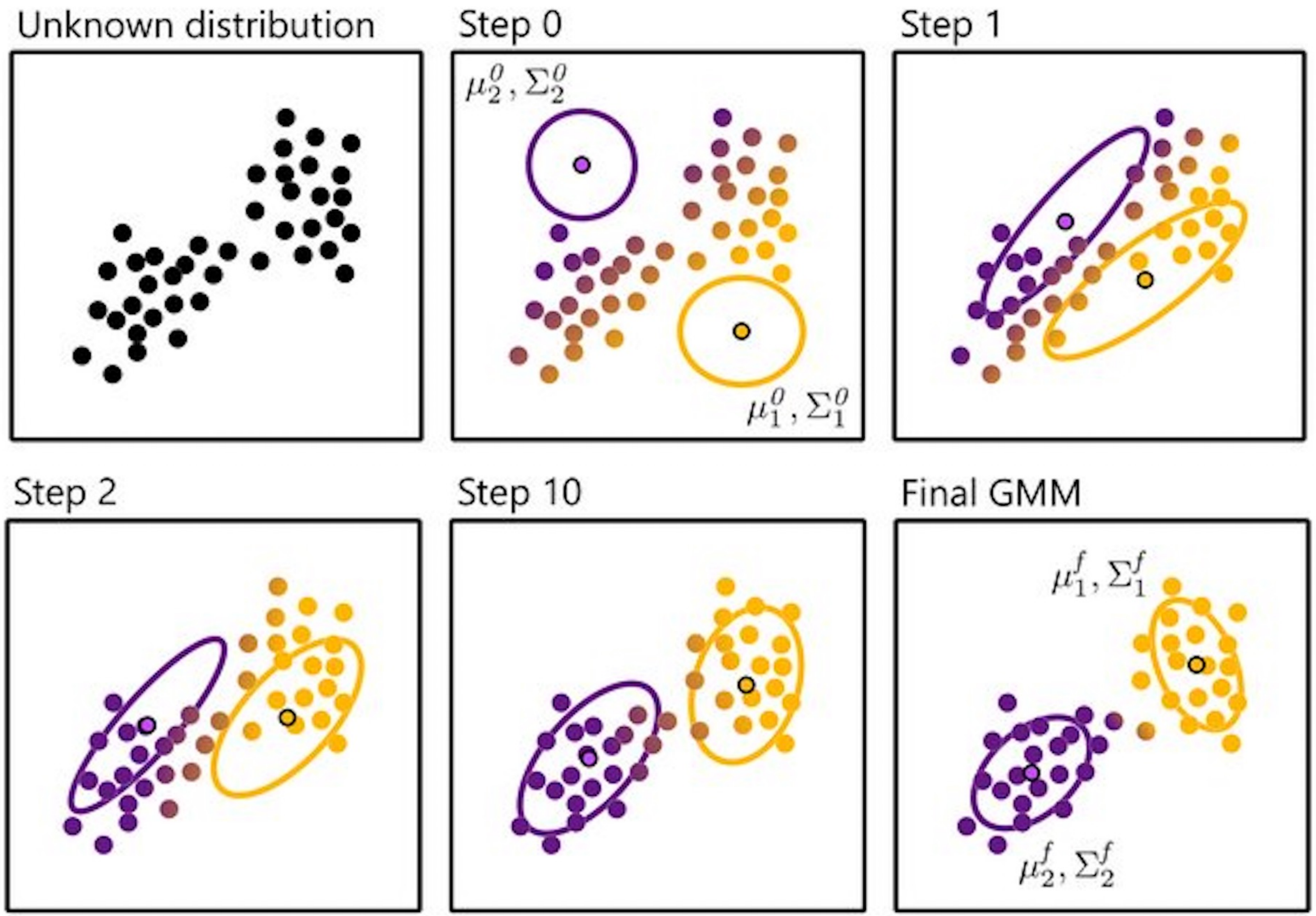
How EM Algorithm Works



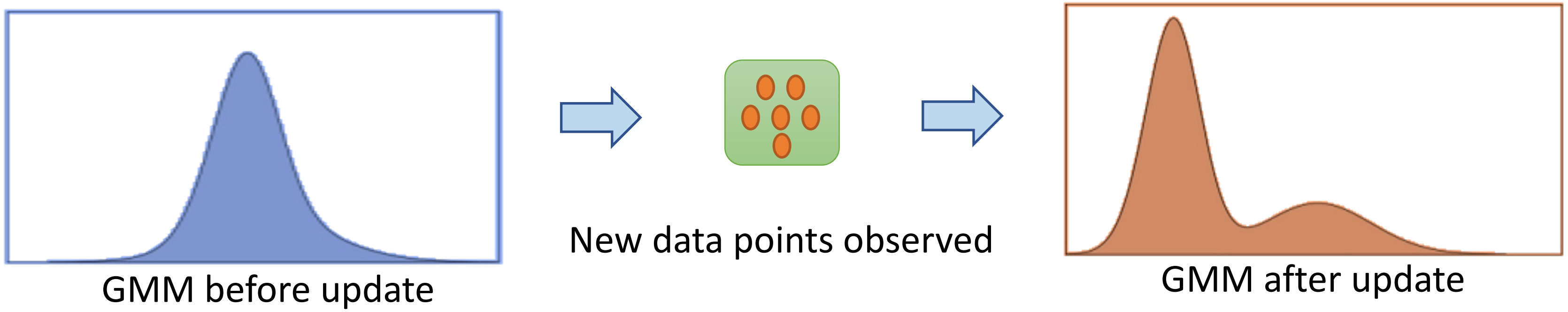
How EM Algorithm Works



How EM Algorithm Works



Incremental GMM Modeling for Time-varying Data



- Update weights as:

$$\omega_{k,t} = (1-\alpha)\omega_{k,t} + \alpha * M_{k,t}\,, \quad M_{k,t} = 1 \text{ for matched dist., 0 for others}$$

- Update means and covariances for the matched distribution as:

$$\mu_t = (1-\rho)\,\mu_{t\text{-}1} + \rho x_t$$

$$\sigma_t^2 = (1-\rho)\,\sigma_{t-1}^2 + \rho(x_t - \mu_t)^T(x_t - \mu_t)\,, \quad \rho = \alpha * N(x_t | \mu_{\mathrm{k}}, \sigma_{\mathrm{k}})$$

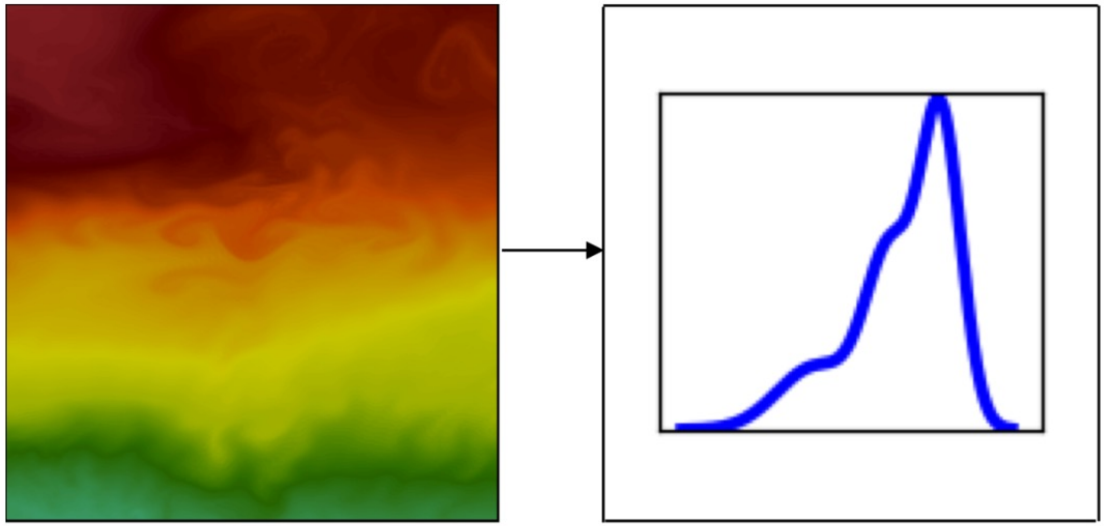
$\Sigma_{k,t} = \sigma_k^2 \mathrm{I}$, where I = Identity matrix, α = learning rate

Distribution-based Large Data Summarization and Visualization

Distribution Models for Big Data Summarization

- Distribution models that can be estimated efficiently and has a compact memory footprint is preferred over other models
- For Univariate Data:
  - Histograms: Fast but takes more space
  - KDE: Computationally expensive and takes more space
  - Gaussian: Fast but often the model is too simple
  - Gaussian mixture model: Parameter estimation can be a little expensive, but representation is compact
- For Multivariate Data:
  - Many of the standard multivariate models become either slow or space inefficient
  - Statistical Copula functions can be used effectively

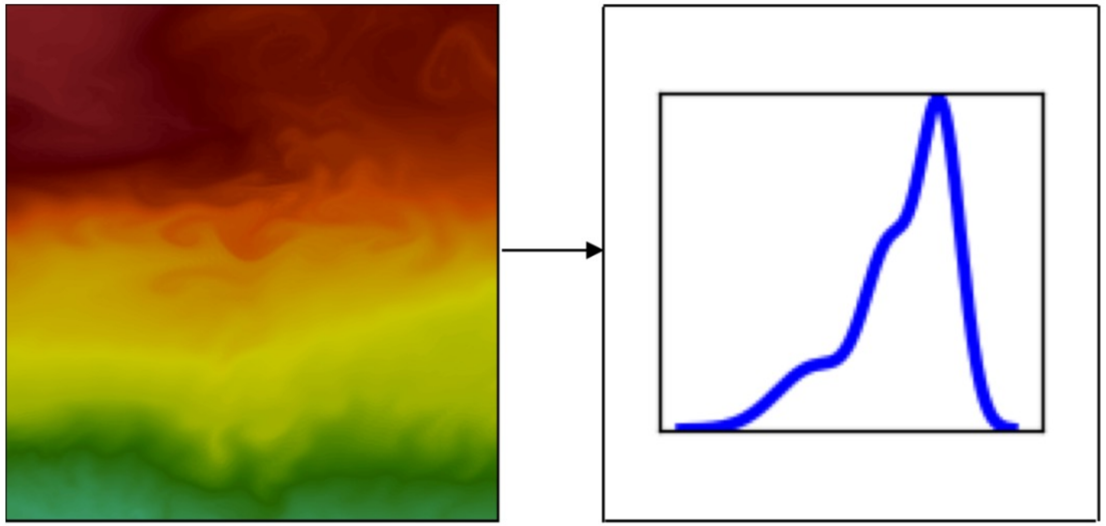
Distribution-based Data Summarization Strategies

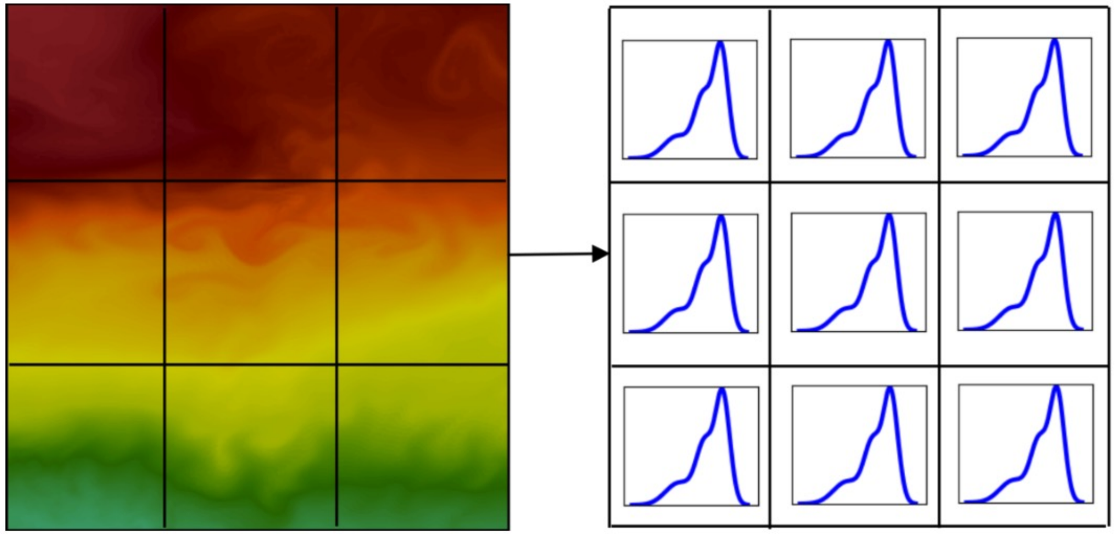


<u>Global distribution model:</u> A single distribution model to represent the entire data set

- Significant data reduction is possible
- Coarse representation of the data
- Not suitable for fine grained visual analysis

Distribution-based Data Summarization Strategies





Global distribution model: A single distribution model to represent the entire data set

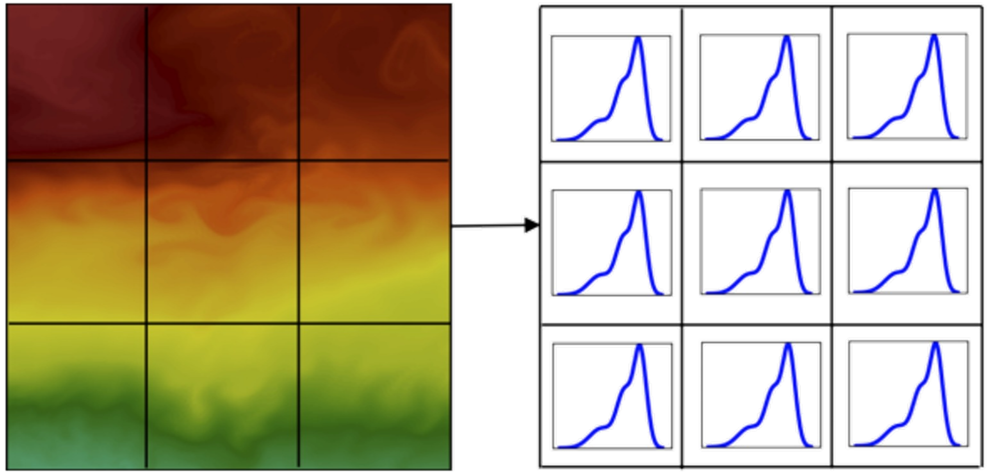
- Significant data reduction is possible
- Coarse representation of the data
- Not suitable for fine grained visual analysis

Local distribution model: Data is divided into small blocks and then each block is summarized using a separate distribution model

- Data reduction at an acceptable range is possible
- Fine details of the data and statistical properties are preserved
- Preferred over global model for scientific data summarization

Local/Region-wise Distribution-based Data Modeling

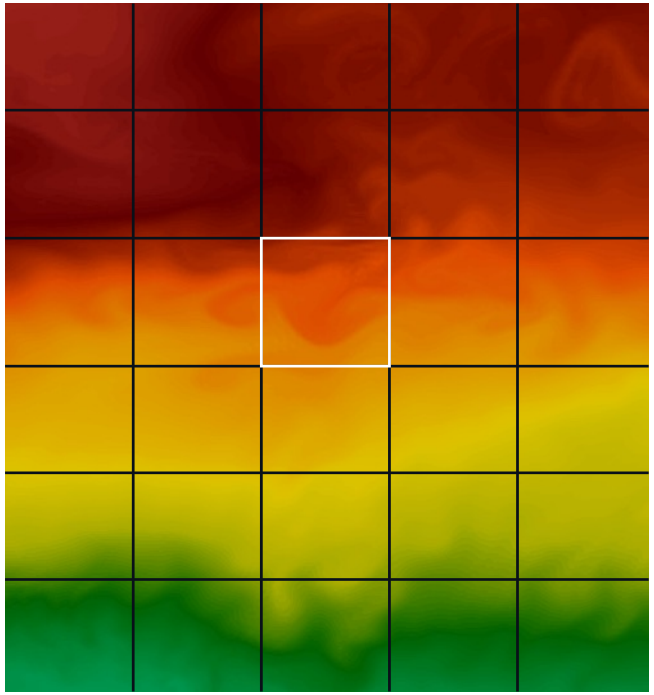
- Local Statistical distribution-based data modeling
  - Partition data into local regions
  - Summarize each region with a statistical distribution model
  - Benefits:
    - Distributions preserve local statistical data properties
    - Reduce data size significantly
    - Enables sampling-based analysis and reconstruction
    - Allows uncertainty quantification



Local distribution-based data model

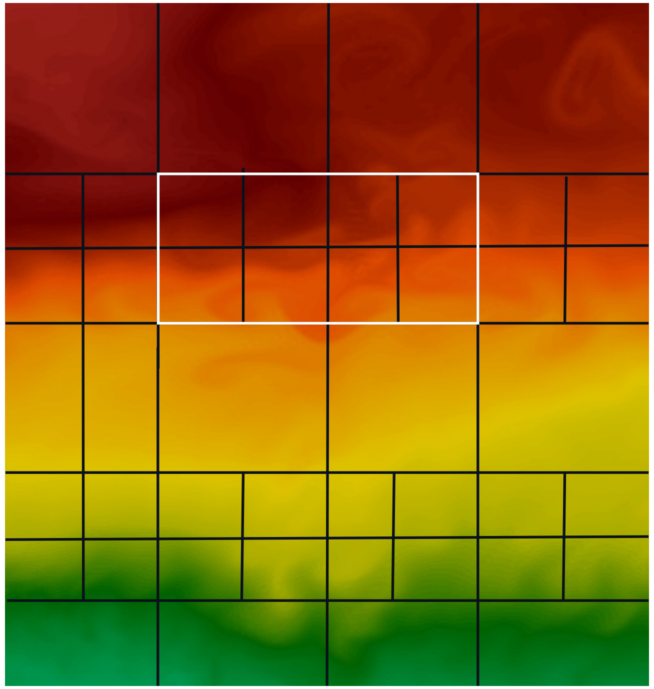
Goals of a Region-wise Statistical Summarization

- Produce coherent partitions
  - Similar data values are grouped together
  - Partitions are spatially contiguous
- Preserve the statistical properties of the data accurately
  - Minimize sampling errors
  - Efficient feature analysis
- Use appropriate distribution models for summarization
  - A compact storage representation



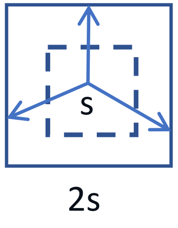
Goals of a Region-wise Statistical Summarization

- Produce coherent partitions
  - Similar data values are grouped together
  - Partitions are spatially contiguous
- Preserve the statistical properties of the data accurately
  - Minimize sampling errors
  - Efficient feature analysis
- Use appropriate distribution models for summarization
  - A compact storage representation

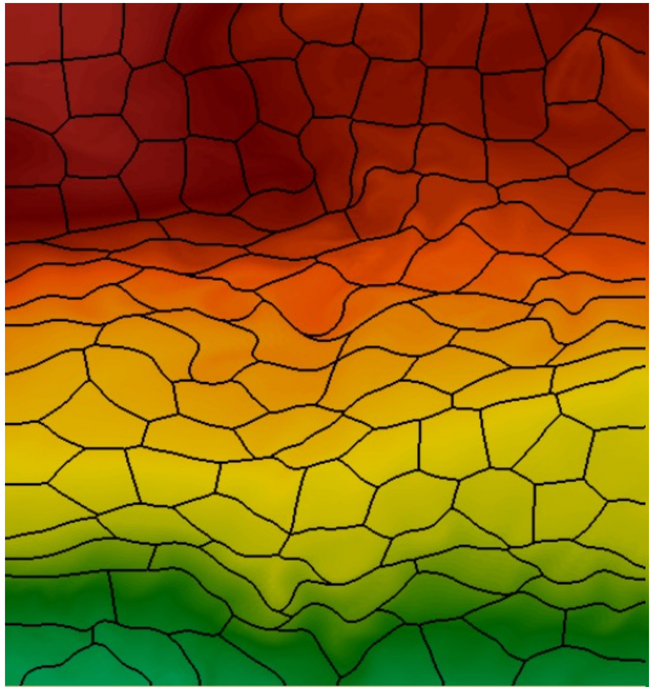


A Superior Solution for Region-wise Statistical Summarization

- Generate partitions based on data homogeneity
- Simple Linear Iterative Clustering (SLIC)
  - Produces irregular shaped partitions/clusters
  - Value variation inside partitions is minimized
  - Reduced sampling error



SLIC concept



$$dist(i,j) = \alpha . \| C_i - P_j \|_2 + (1-\alpha) . | val_i - val_j |$$

SLIC Algorithm Steps

Algorithm 1 Efficient superpixel segmentation

```
Initialize cluster centers C_k = [l_k, a_k, b_k, x_k, y_k]^T by sampling pixels at regular grid steps S.
Perturb cluster centers in an n × n neighborhood, to the lowest gradient position.
repeat
    for each cluster center C_k do
        Assign the best matching pixels from a 2S × 2S square neighborhood around the cluster center according to the distance measure (Eq. 1).
    end for
    Compute new cluster centers and residual error E {L1 distance between previous centers and recomputed centers}
until E ≤ threshold
Enforce connectivity.
```