Research Directions in Adaptive Mixtures and Model-Based Clustering

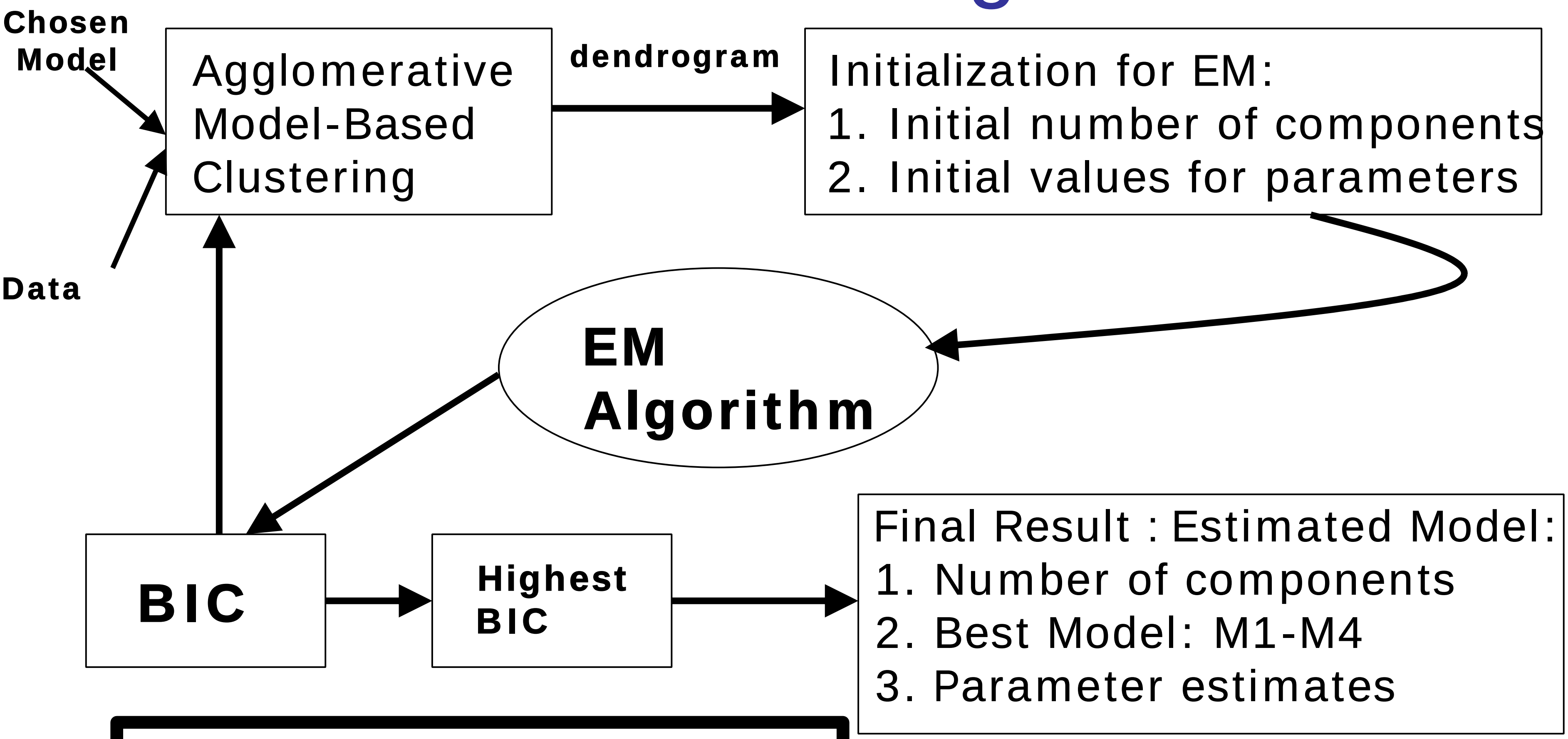
Wendy L. Martinez
Office of Naval Research
Jeffrey L. Solka
NSWCDD/GMU

ACAS 2004

Outline

- Model-based Clustering (MBC).
  - Mixture models and the EM algorithm.
  - The agglomerative step.
  - The model types.
- Adaptive Mixtures Density Estimation
- Their Synthesis
  - Initialization for MB agglomerative clustering
  - MB Adaptive Mixtures Density Estimation
- Preliminary Results.

Model-Based Clustering



Standard MBC performs hierarchical clustering starting with the full dataset, this is computationally intensive.

MODEL-BASED CLUSTERING

- This technique takes a density function approach.
- Uses finite mixture densities as models for cluster analysis.
- Each component density characterizes a cluster.

FINITE MIXTURES REVIEW

$$\hat{f}(x) = \sum_{i=1}^{g} \pi_i f_i(x,\theta)$$

$$f_i(x,\theta) = N(\mu_i,\Sigma_i)$$

- Model the density as a sum of *g* weighted densities.
- Expectation-maximization method used to estimate parameters.
- Must assume distribution for components - usually normal distribution.
- Each component characterizes a cluster.

EXPECTATION-MAXIMIZATION (EM) METHOD

- Method for building or estimating the model.
- Solution of likelihood functions requires iterative procedure.
- E Step - Expectation:
  - Find probability that observations belong to each component density - the posteriors (τ_{ij}'s).
- M Step - Maximization:
  - Update all parameters based on posteriors (π_i, μ_i, Σ_i).

EXPECTATION-MAXIMIZATION (EM) METHOD

- Issues:
  - Can converge to a local optimum.
  - Can diverge.
  - *Requires initial guess at the parameters of the component densities.*
    - *Requires initial guess at the weights (or priors).*
    - *Need an estimate of the number of components.*
  - Requires an assumed distribution for the component densities.
- *Model-based clustering addresses these issues.*

AGGLOMERATIVE MBC

- Regular agglomerative clustering:
  - Each point is in a cluster.
  - Two closest clusters are merged at each step.
  - Closeness is determined by distance and linkage.
- Agglomerative model-based clustering:
  - At each step, two clusters are merged such that the likelihood for the given model is maximized.
- We propose using Adaptive Mixtures to initialize MB agglomerative clustering.

MODEL-BASED CLUSTERING

- Best model is chosen using the Bayesian Information Criterion (m_M is # parameters, L_M is loglikelihood):

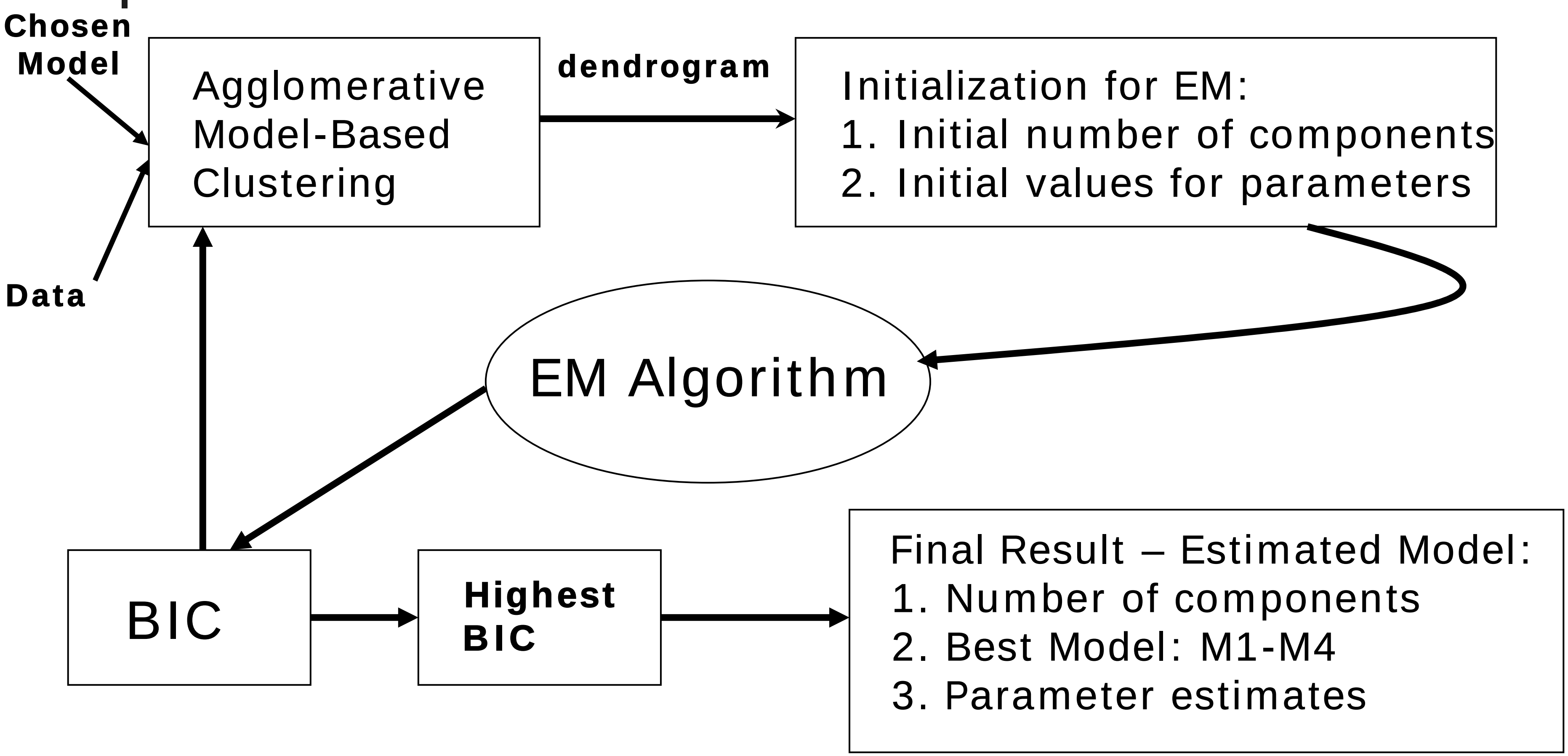
$$BIC \equiv 2L_M(\mathbf{x},\hat{\boldsymbol{\theta}}) - m_M \log(n)$$

- The four models are (***more models are possible***):
  - Spherical/equal (M1): $\boldsymbol{\Sigma}_K = \sigma^2\mathbf{I}$
  - Spherical/unequal (M2): $\boldsymbol{\Sigma}_K = \sigma_K^2\mathbf{I}$
  - Ellipsoidal/equal (M3): $\boldsymbol{\Sigma}_K = \boldsymbol{\Sigma}$
  - Ellipsoidal/unequal (unconstrained) (M4): $\boldsymbol{\Sigma}_K = \boldsymbol{\Sigma}_K$

MODEL-BASED CLUSTERING in a Nutshell

1. Apply the unconstrained agglomerative MBC procedure.
2. Choose number of clusters/densities, g.
3. Choose model: M1 – M4.
4. Find the partition given by step 1 for the specified g.
5. Using this partition, find the weights, means and covariances for each term, based on the model in step 3.
6. Using the chosen g (step 2) and the initial values (step 5), apply the EM algorithm.
7. Calculate the BIC for this value of g and M.
8. Go to step 3 to choose another value of M and repeat.
9. Go to sep 2 and choose another model g and repeat.

MODEL-BASED CLUSTERING



ADAPTIVE MIXTURES DENSITY ESTIMATION (AMDE)

- Priebe and Marchette; 1990s.
- Hybrid of Kernel Estimator and Mixture Model.
- Number of Terms Driven by the Data.
- L1 Consistent.

AMDE ALGORITHM

1 - Given a New Observation.

2 - Update Existing Model Using the Recursive EM.

or

3 - Add a New Term to “Explain” This Data Point.

Recursive EM Update Equations

$$\hat{\tau}_{n+1}^{(i)} = \frac{\pi_n^{(i)} \hat{f}^{(i)}(\vec{x}_{n+1}; \hat{\theta}_n)}{\sum_{t=1}^{g} \pi_n^t \hat{f}^{(t)}(\vec{x}_{n+1}; \hat{\theta}_n)}$$

$$\hat{\pi}_{n+1}^{(i)} = \hat{\pi}_n^{(i)} + \frac{1}{n}(\hat{\tau}_{n+1}^{(i)} - \hat{\pi}_n^{(i)})$$

$$\hat{\mu}_{n+1}^{(i)} = \hat{\mu}_n^{(i)} + \frac{\hat{\tau}_{n+1}^{(i)}}{n\hat{\pi}_n^{(i)}}[(\vec{x}_{n+1} - \hat{\mu}_n^{(i)})A - \hat{\Sigma}_n^{(i)}]$$

$$A = (\vec{x}_{n+1} - \hat{\mu}_n^{(i)})^T$$

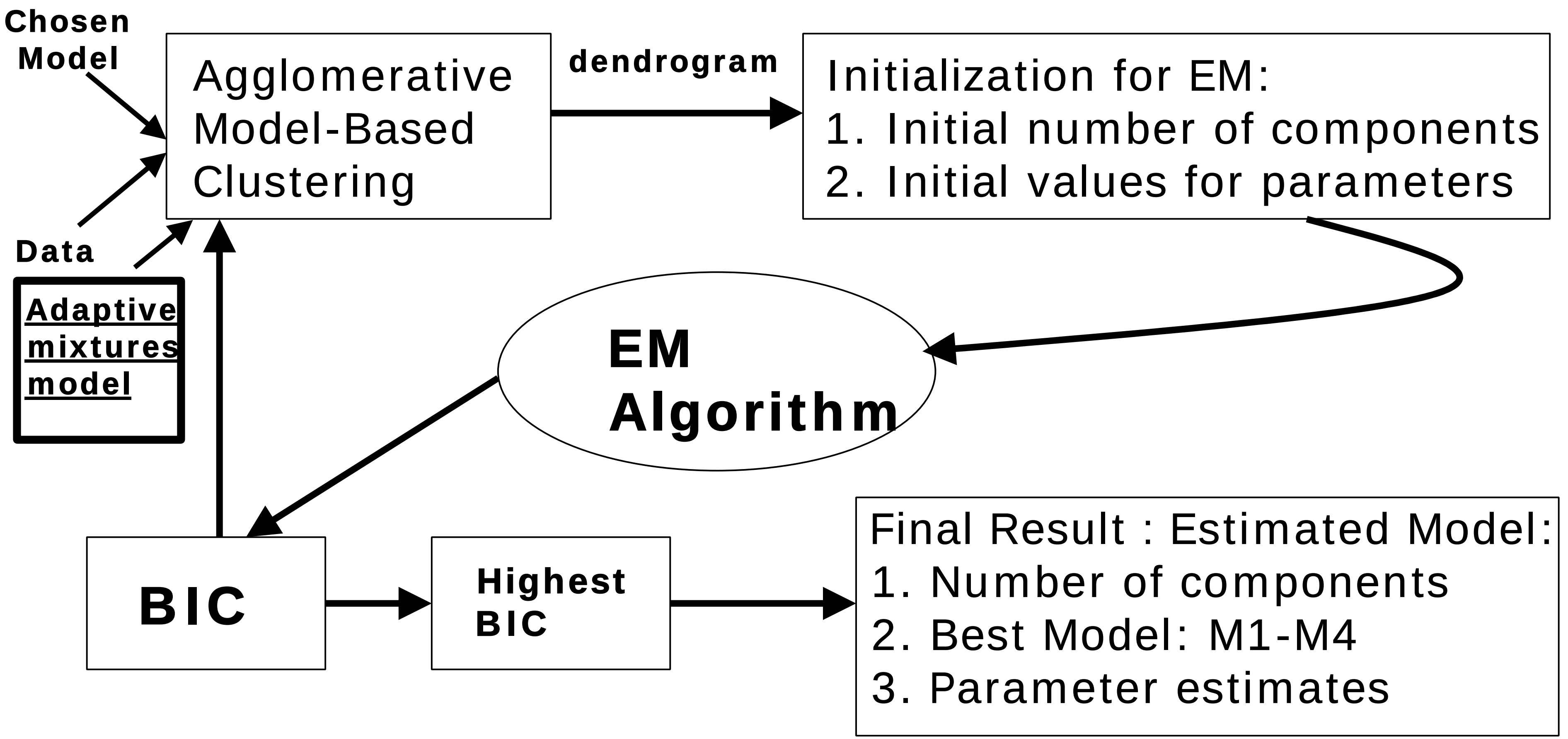
Similarly for Σ

CREATE RULE - AMDE

- Test the Mahalanobis distance from current data point to each mixture term in the existing model.

- Add in a new term when this distance exceeds a certain “create threshold”
  - Location given by current data point.
  - Covariance given by weighted average of the existing covariances.
  - Mixing coefficient set to 1/n.

MBC with an MADE Start



MBC With AMDE Smart Start

1. Form an adaptive mixtures model of the dataset. (Set create threshold in order to guarantee an over determined model.)
2. Partition the data based on the AMDE model using τ_{ij}. (Note some of the original AMDE mixture terms "die" due to insufficient support.)
3. Utilize this partition as a start to the usual MBC procedure. (Instead of starting with as many terms as points we start with approximately log(n) number of points.)

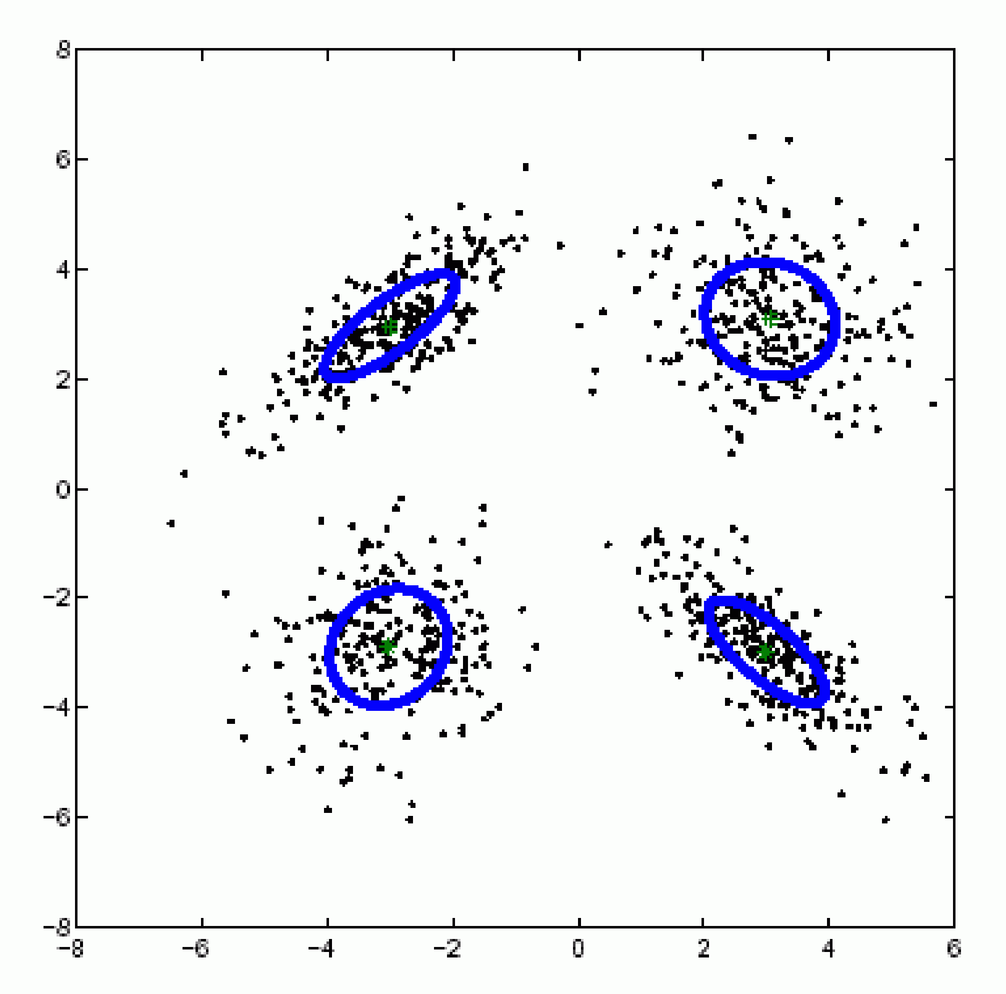
Other Possibilities

- Other types of initialization:
  - Posse (JCGS) used initial partitions based on minimal spanning tree.
  - K-means
- Benefits of AMDE initialization:
  - Do not have to specify number of clusters as in k-means.
  - Methods like k-means impose a certain structure.
  - In most cases, initial clusters are not singletons.

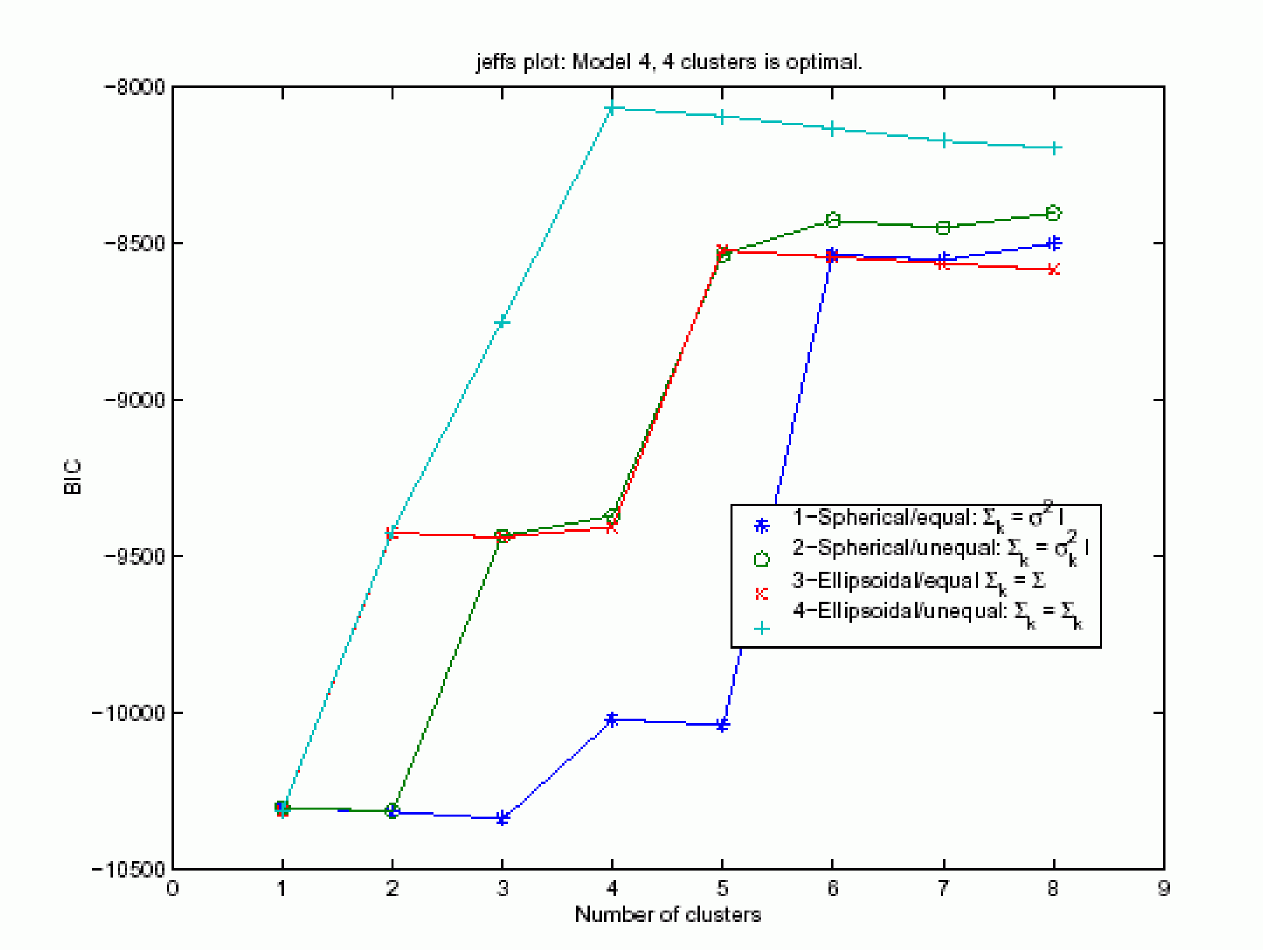
Why Do This?

- Computational tradeoff of the AMDE procedure vs. the agglomerative procedure on the full dataset.
  - Advantages as the size of the dataset grows.
  - Non-singleton clusters
  - Save on storage
- AMDE is data order dependent.
  - Multiple mixture models/clustering can be obtained by merely reordering the dataset.
  - Could get a distribution of models (number of clusters/BICs)

4 Term Test Case



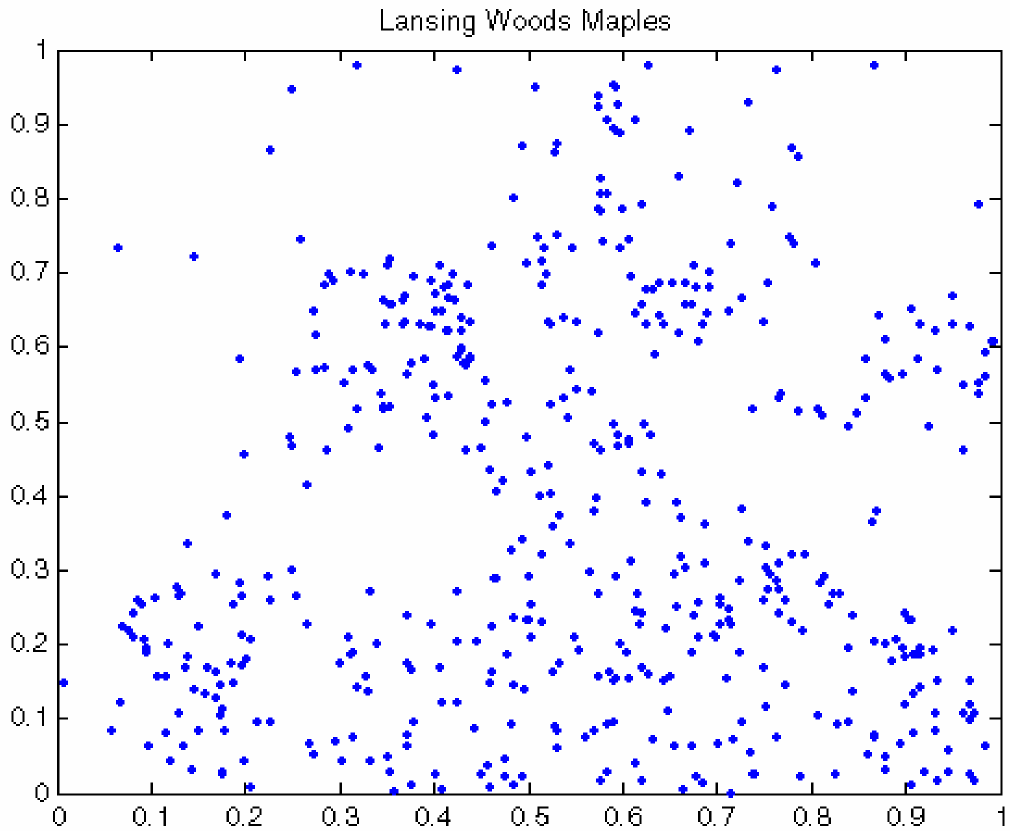
4 Term BIC Curves



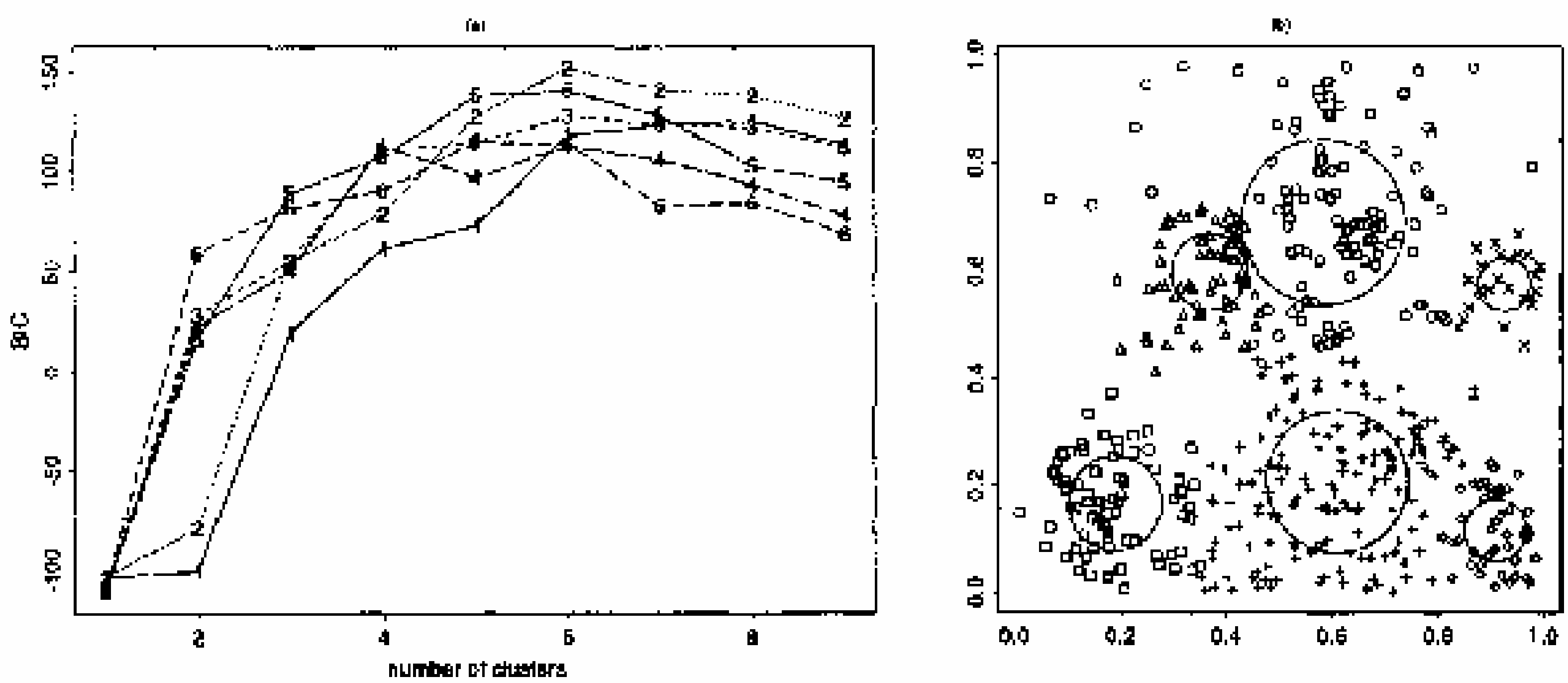
Experiment – Real Data

- Model-based clustering was applied to Lansing Woods maples.
- Ran 20 trials with AMDE initialization.
- Re-ordered data each time.
- Maximum BIC model is 6 component non-uniform spherical mixture.
- This is model 2:
  - Covariances are diagonal – equal variances.
  - Covariances are not equal across terms.

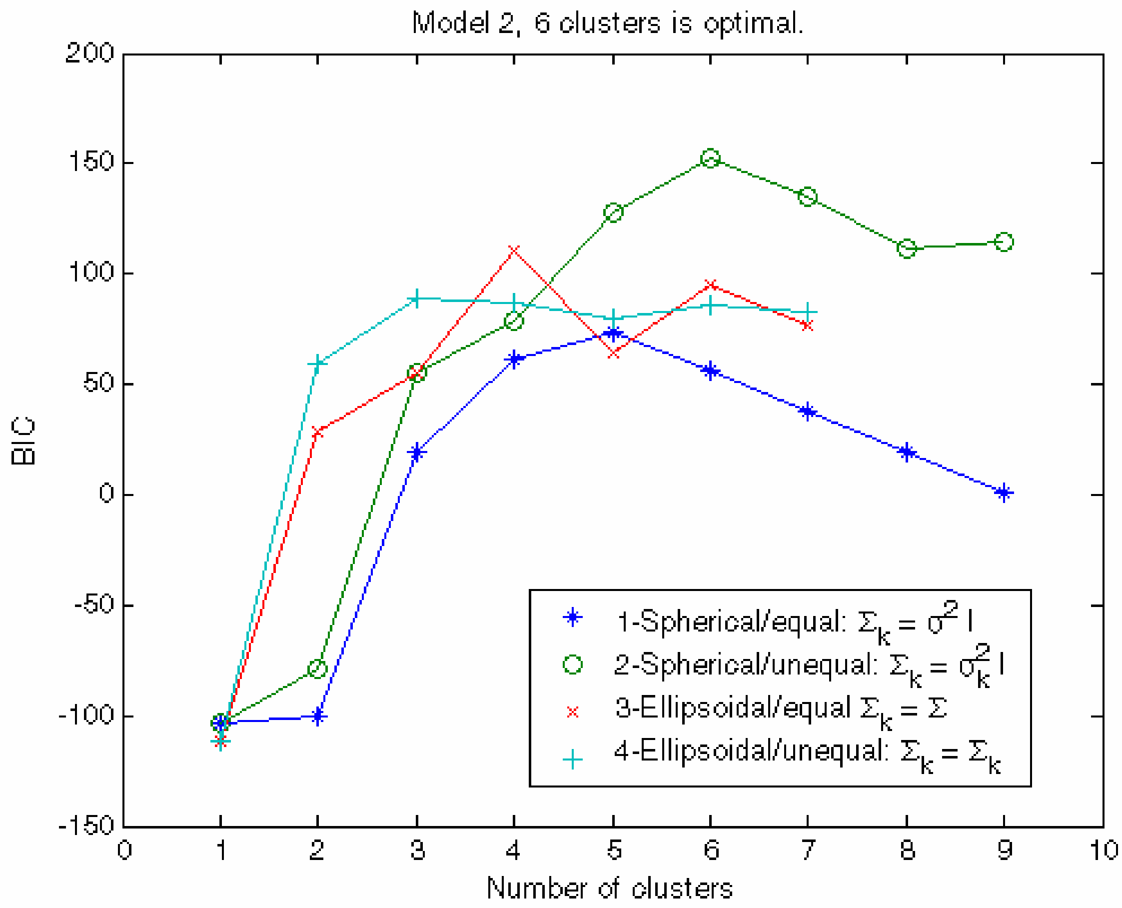
The Raw Data


Lansing Woods Maples
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

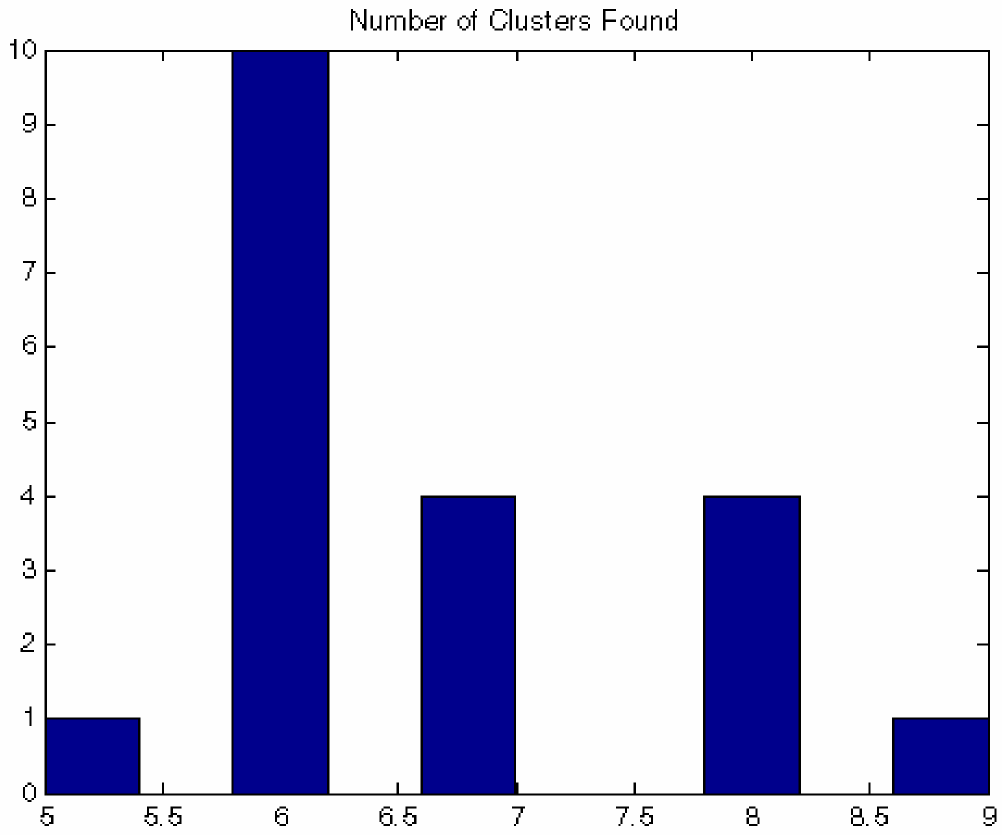
Original Configuration



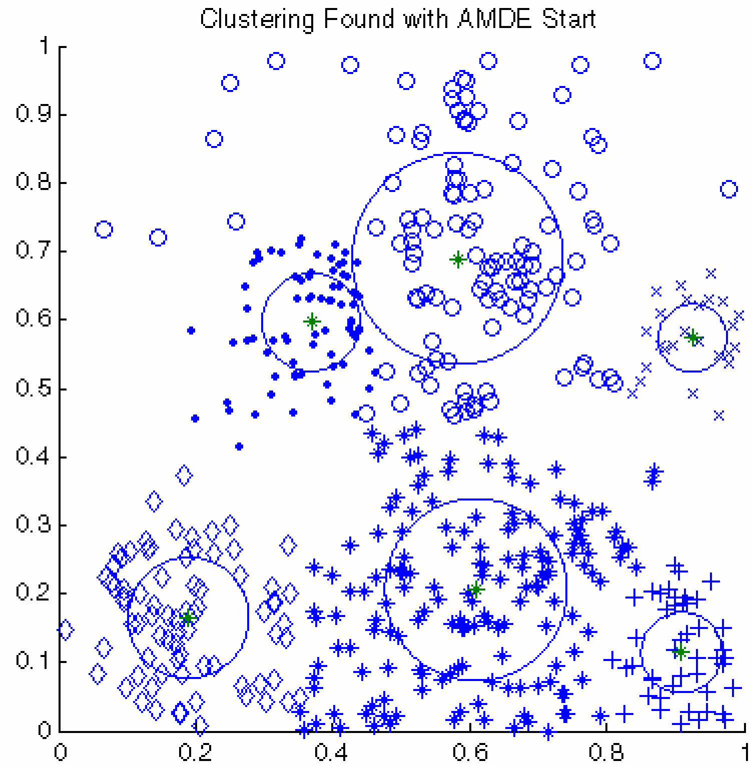
BICS for Best Trial



Number of Clusters



Configuration with AMDE



Conclusion

- Discussed an initialization procedure for the model-based agglomerative clustering.
- Showed applications to synthetic and real data.
- Possible advantages:
  - Savings in storage.
  - Possibly find other solutions – greedy algorithm
- Formulation of Model-Based AMDE.
- Use of MB-agglomerative clustering as a way of pruning terms.