Model-Based Methods for Clustering of Spatial Time Series Data—Lecture 1: Model-Based Clustering

Hien D. Nguyen[1,2]

[1]School of Mathematics and Physics, University of Queensland
[2]Centre for Advanced Imaging, University of Queensland

University of Western Australia, 23-06-2016

About Myself

- Postdoctoral Research Fellow:
  - School of Mathematics and Physics, UQ.
  - Centre for Advanced Imaging, UQ.
- PhD Graduate, UQ May/2015, in Computational Statistics and Image Analysis.
  - Thesis titled: Finite mixture models for regression problems.
  - Advisors: Prof. Geoffrey McLachlan, Dr. Ian Wood, and Dr. Andrew Janke.
- Statistics and Data Science Lecturer, La Trobe University Dec/2016.

About Myself

- Postdoctoral Research Fellow:
  - School of Mathematics and Physics, UQ.
  - Centre for Advanced Imaging, UQ.
- PhD Graduate, UQ May/2015, in Computational Statistics and Image Analysis.
  - Thesis titled: Finite mixture models for regression problems.
  - Advisors: Prof. Geoffrey McLachlan, Dr. Ian Wood, and Dr. Andrew Janke.
- Statistics and Data Science Lecturer, La Trobe University Dec/2016.

About Myself

- Postdoctoral Research Fellow:
  - School of Mathematics and Physics, UQ.
  - Centre for Advanced Imaging, UQ.
- PhD Graduate, UQ May/2015, in Computational Statistics and Image Analysis.
  - Thesis titled: Finite mixture models for regression problems.
  - Advisors: Prof. Geoffrey McLachlan, Dr. Ian Wood, and Dr. Andrew Janke.
- Statistics and Data Science Lecturer, La Trobe University Dec/2016.

Current Research

- Model-based clustering for unconventional regression data.
  - Nguyen, H.D. and McLachlan G.J. (2016). **Linear mixed models with marginally symmetric nonparametric random effects**. Computational Statistics & Data Analysis. To appear.
  - Nguyen, H.D., McLachlan G.J., Ullmann, J.F.P., and Janke, A.L. (2016). **Laplace mixture autoregressive models**. Statistics & Probability Letters.
  - Nguyen, H.D., McLachlan, G.J., and Wood, I.A. (2016). **Mixtures of spatial spline regressions for clustering and classification**. Computational Statistics & Data Analysis.
- Big Data-appropriate algorithms for statistical models.
  - Nguyen, H.D. and McLachlan G.J. (2015). **Maximum likelihood estimation of Gaussian mixture models without matrix operations**. Advances in Data Analysis and Classification.
  - Nguyen, H.D. and McLachlan G.J. (2016). **Laplace mixtures of linear experts**. Computational Statistics & Data Analysis.

Current Research

- Model-based clustering for unconventional regression data.
  - Nguyen, H.D. and McLachlan G.J. (2016). **Linear mixed models with marginally symmetric nonparametric random effects**. Computational Statistics & Data Analysis. To appear.
  - Nguyen, H.D., McLachlan G.J., Ullmann, J.F.P., and Janke, A.L. (2016). **Laplace mixture autoregressive models**. Statistics & Probability Letters.
  - Nguyen, H.D., McLachlan, G.J., and Wood, I.A. (2016). **Mixtures of spatial spline regressions for clustering and classification**. Computational Statistics & Data Analysis.
- Big Data-appropriate algorithms for statistical models.
  - Nguyen, H.D. and McLachlan G.J. (2015). **Maximum likelihood estimation of Gaussian mixture models without matrix operations**. Advances in Data Analysis and Classification.
  - Nguyen, H.D. and McLachlan G.J. (2016). **Laplace mixtures of linear experts**. Computational Statistics & Data Analysis.

Current Research—2

- Large-scale testing in proteomics experiments.
  - Nguyen, H.D., Hill, M.M., and Wood, I.A. (2012), **A robust permutation test for quantitative SILAC proteomics experiments**, Journal of Integrated OMICS
  - Chen, D., Shah, A., Nguyen, H., Loo, D., Inder, K., and Hill, M. (2014), **Online quantitative proteomics p-value calculator for permutation-based statistical testing of peptide ratios**. Journal of Proteomics Research.
- Statistical Inference of Neural Networks.
  - Nguyen, H.D., and Wood, I.A. (2016), **A block successive lower-bound maximization algorithm for the maximum pseudolikelihood estimation of fully visible Boltzmann machines**, Neural Computation.
  - Nguyen, H.D., and Wood, I.A. (2016), **Asymptotic normality of the maximum pseudolikelihood estimator for fully-visible Boltzmann machines**, IEEE Transactions on Neural Networks and Learning Systems.

Current Research—2

- Large-scale testing in proteomics experiments.
  - Nguyen, H.D., Hill, M.M., and Wood, I.A. (2012), **A robust permutation test for quantitative SILAC proteomics experiments**, Journal of Integrated OMICS
  - Chen, D., Shah, A., Nguyen, H., Loo, D., Inder, K., and Hill, M. (2014), **Online quantitative proteomics p-value calculator for permutation-based statistical testing of peptide ratios**. Journal of Proteomics Research.
- Statistical Inference of Neural Networks.
  - Nguyen, H.D., and Wood, I.A. (2016), **A block successive lower-bound maximization algorithm for the maximum pseudolikelihood estimation of fully visible Boltzmann machines**, Neural Computation.
  - Nguyen, H.D., and Wood, I.A. (2016), **Asymptotic normality of the maximum pseudolikelihood estimator for fully-visible Boltzmann machines**, IEEE Transactions on Neural Networks and Learning Systems.

Current Research—3

Figure 1: Zebrafish. [Wikipedia]

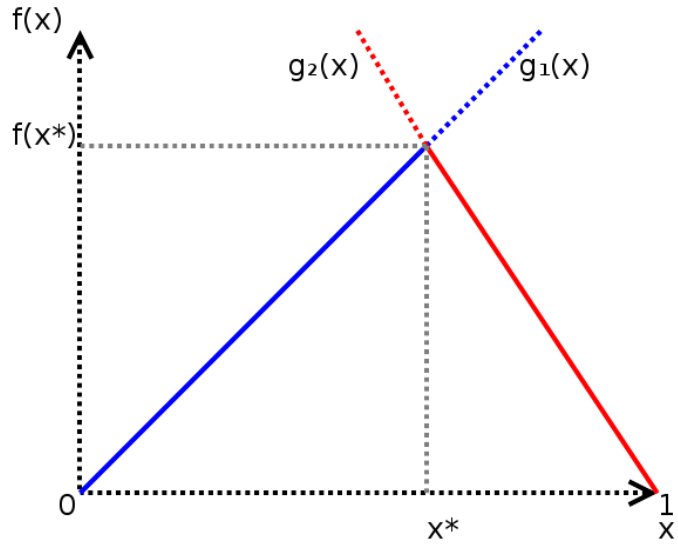

Figure 2: Schematic of a triangular density function.
[Nguyen and McLachlan, 2016]

About This Workshop

- Based on the recently-accepted publication:
  - Nguyen, H.D., McLachlan G.J., Ullmann, J.F.P., and Janke, A.L. (2016). **Spatial clustering of time-series via mixtures of autoregressive models and Markov random fields.** Statistica Neerlandica. To appear.
  - Available on ArXiv at http://arxiv.org/abs/1601.03517.
- Develops upon previous work in model-based clustering of spatial data.
  - Nguyen, H.D., McLachlan, G.J., Cherbuin, N., and Janke, A.L. (2014). **False discovery rate control in magnetic resonance imaging studies via Markov random fields**. IEEE Transactions on Medical Imaging.

About This Workshop

- Based on the recently-accepted publication:
  - Nguyen, H.D., McLachlan G.J., Ullmann, J.F.P., and Janke, A.L. (2016). **Spatial clustering of time-series via mixtures of autoregressive models and Markov random fields.** Statistica Neerlandica. To appear.
  - Available on ArXiv at http://arxiv.org/abs/1601.03517.
- Develops upon previous work in model-based clustering of spatial data.
  - Nguyen, H.D., McLachlan, G.J., Cherbuin, N., and Janke, A.L. (2014). **False discovery rate control in magnetic resonance imaging studies via Markov random fields**. IEEE Transactions on Medical Imaging.

Format of Workshop

- 2×1 hour Lectures $+$ 2×1.5 hour Lectures.
- 4×2 hour Computer Practicals.
  - Conducted using the **R** programming language via **RStudio**.
  - Demonstrate how methods can be applied in practice using open-source packages.

Format of Workshop

- 2×1 hour Lectures $+$ 2×1.5 hour Lectures.
- 4×2 hour Computer Practicals.
  - Conducted using the **R** programming language via **RStudio**.
  - Demonstrate how methods can be applied in practice using open-source packages.

Content of Workshop

Session 1 Model-based clustering of spatially-correlated data.

1. Finite mixtures of Gaussian distributions.
2. Bayes-optimal clustering.
3. Maximum marginal-likelihood estimation.
4. Asymptotic inference under spatial correlation.
5. Model selection via the Bayesian information criterion.

Session 2 Markov random fields for spatial modeling.

1. Infinite square lattices
2. Neighborhood structures
3. Markov random fields
4. Maximum pseudolikelihood estimation.
5. Minorization–maximization algorithms.
6. Convergence and asymptotic theory.
7. Model selection via the pseudolikelihood information criterion.

Content of Workshop

Session 1 Model-based clustering of spatially-correlated data.

1. Finite mixtures of Gaussian distributions.
2. Bayes-optimal clustering.
3. Maximum marginal-likelihood estimation.
4. Asymptotic inference under spatial correlation.
5. Model selection via the Bayesian information criterion.

Session 2 Markov random fields for spatial modeling.

1. Infinite square lattices
2. Neighborhood structures
3. Markov random fields
4. Maximum pseudolikelihood estimation.
5. Minorization–maximization algorithms.
6. Convergence and asymptotic theory.
7. Model selection via the pseudolikelihood information criterion.

Content of Workshop—2

Session 3 Time series modeling.

1. Autoregressive-moving average models.
2. Stationarity and invertibility.
3. Autocorrelation functions and correlograms.
4. Maximum likelihood estimation.
5. Asymptotic Inference.
6. Extremum estimators.
7. Model selection via the Bayesian information criterion.

Session 4 Model-based time series clustering.

1. Finite mixture of autoregressive models.
2. Maximum marginal-likelihood estimation.
3. MM Algorithm.
4. Asymptotic inference under spatial correlation.
5. Information theoretic model selection.
6. Markov random fields for spatial modeling.

Content of Workshop—2

Session 3 Time series modeling.

1. Autoregressive-moving average models.
2. Stationarity and invertibility.
3. Autocorrelation functions and correlograms.
4. Maximum likelihood estimation.
5. Asymptotic Inference.
6. Extremum estimators.
7. Model selection via the Bayesian information criterion.

Session 4 Model-based time series clustering.

1. Finite mixture of autoregressive models.
2. Maximum marginal-likelihood estimation.
3. MM Algorithm.
4. Asymptotic inference under spatial correlation.
5. Information theoretic model selection.
6. Markov random fields for spatial modeling.

Clustering

- [Jain and Dubes, 1988] writes that a **cluster** is comprised of a number of similar objects collected or grouped together.

- In [Everitt et al., 2011], examples of definitions for clusters include the following.

  1. A cluster is a set of entities which are alike, and entities from different clusters are not alike.
  2. A cluster is an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it.
  3. Clusters may be described as connected regions of a multi-dimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points.

- **Clustering** is the task of deducing objects into clusters, via one of the numerous definitions of clusters.

Clustering

- [Jain and Dubes, 1988] writes that a **cluster** is comprised of a number of similar objects collected or grouped together.

- In [Everitt et al., 2011], examples of definitions for clusters include the following.

  1. A cluster is a set of entities which are alike, and entities from different clusters are not alike.
  2. A cluster is an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it.
  3. Clusters may be described as connected regions of a multi-dimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points.

- **Clustering** is the task of deducing objects into clusters, via one of the numerous definitions of clusters.

Clustering

- [Jain and Dubes, 1988] writes that a **cluster** is comprised of a number of similar objects collected or grouped together.

- In [Everitt et al., 2011], examples of definitions for clusters include the following.

  1. A cluster is a set of entities which are alike, and entities from different clusters are not alike.
  2. A cluster is an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it.
  3. Clusters may be described as connected regions of a multi-dimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points.

- **Clustering** is the task of deducing objects into clusters, via one of the numerous definitions of clusters.

Iris Data (2D)

- Analyzed by R. A. Fisher in the seminal paper [Fisher, 1936].
- 150 data points in $\mathbb{R}^2$ (petal lengths and sepal lengths).
- Three subspecies of iris flowers (setosa, versicolor, virginica).
- 50 flowers from each subspecies.

Iris Data—2

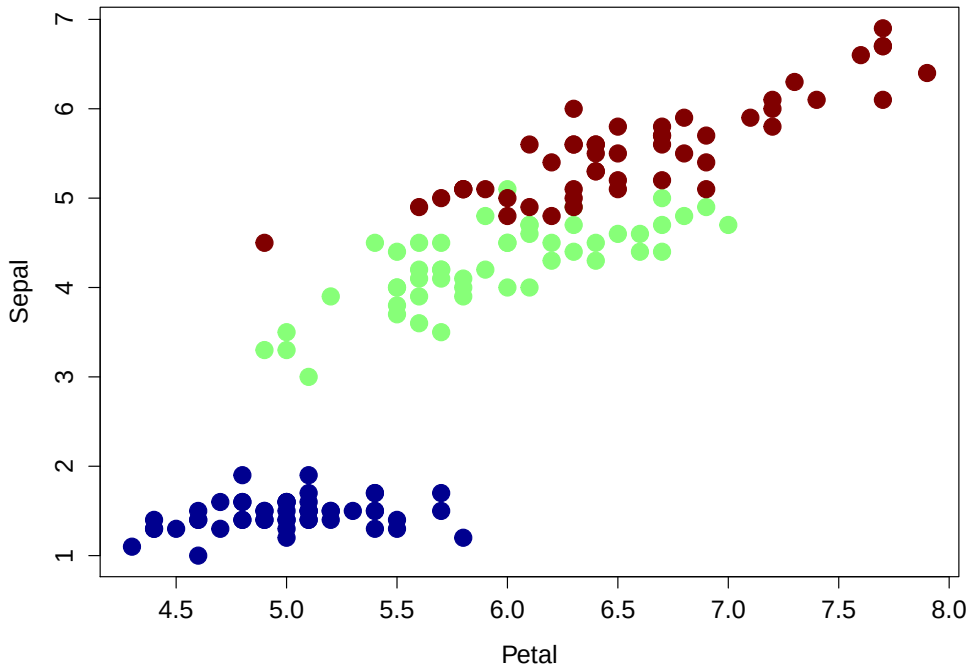

Figure 3: Blue: setosa. Green: versicolor. Maroon: virginica.

Iris Data—3

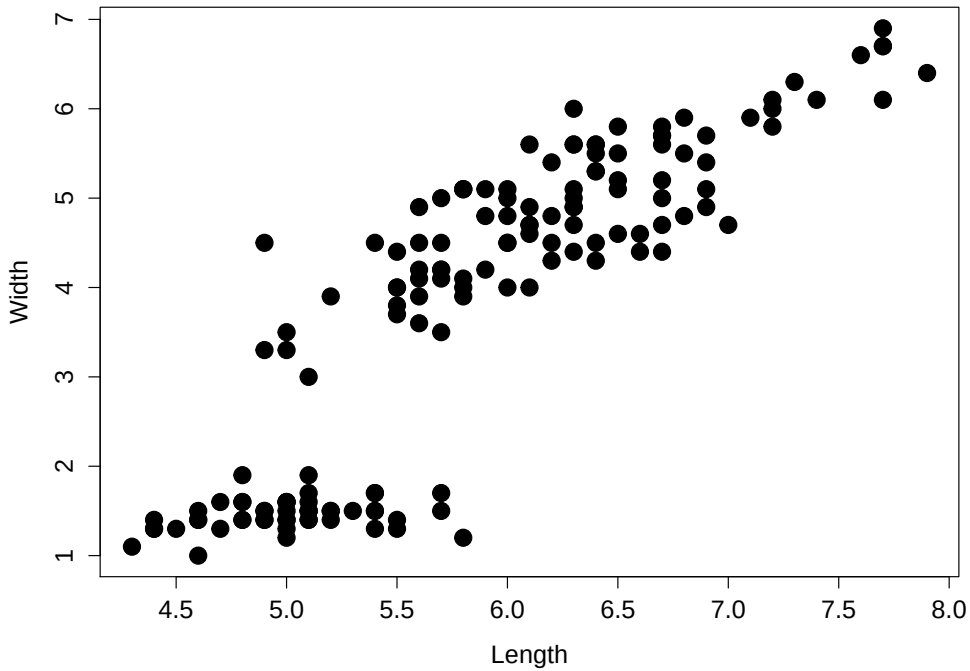

Figure 4: Unlabeled Iris Data.

Model-based Clustering

- Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \mathbb{X}$ be an ID (identically distributed) random sample of n observations, from some population with **population probability density function** $f(\boldsymbol{x})$ with support $\mathbb{X}$.
- Suppose that we know that there are $g \in \mathbb{N}$ subpopulations that form the population.
- Associate with each $\boldsymbol{X}_j$ $(j = 1, ..., n)$ a **class label** $C_j = 1, ..., g$, where C_j indicates the membership of observation $\boldsymbol{X}_j$ to one of the g subpopulations.

Model-based Clustering

- Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \mathbb{X}$ be an ID (identically distributed) random sample of n observations, from some population with **population probability density function** $f(\boldsymbol{x})$ with support $\mathbb{X}$.
- Suppose that we know that there are $g \in \mathbb{N}$ subpopulations that form the population.
- Associate with each $\boldsymbol{X}_j$ $(j = 1, ..., n)$ a **class label** $C_j = 1, ..., g$, where C_j indicates the membership of observation $\boldsymbol{X}_j$ to one of the g subpopulations.

Model-based Clustering

- Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \mathbb{X}$ be an ID (identically distributed) random sample of n observations, from some population with **population probability density function** $f(\boldsymbol{x})$ with support $\mathbb{X}$.
- Suppose that we know that there are $g \in \mathbb{N}$ subpopulations that form the population.
- Associate with each $\boldsymbol{X}_j$ $(j = 1, ..., n)$ a **class label** $C_j = 1, ..., g$, where C_j indicates the membership of observation $\boldsymbol{X}_j$ to one of the g subpopulations.

Finite Mixture Models

- Suppose that a priori to observing $\boldsymbol{X}_j$, $C_j = i$ with probability $\pi_i = \mathbb{P}(C_j = i)$, for $i = 1, ..., g$.

- Upon observing $\boldsymbol{X}_j$, if $C_j = i$, then $\boldsymbol{X}_j$ has (conditional) **component probability density function**

$$f(\boldsymbol{x}_j | C_j = i) = f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i$ is a parameter vector that f_i may depend upon.

- The characterization yields the expression for the population probability density function

$$f(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i),$$

which we call a finite mixture model, where we put all of the probabilities π_i and parameter vectors $\boldsymbol{\theta}_i$ into $\boldsymbol{\theta}$.

Finite Mixture Models

- Suppose that a priori to observing $\boldsymbol{X}_j$, $C_j = i$ with probability $\pi_i = \mathbb{P}(C_j = i)$, for $i = 1, ..., g$.
- Upon observing $\boldsymbol{X}_j$, if $C_j = i$, then $\boldsymbol{X}_j$ has (conditional) **component probability density function**

$$f(\boldsymbol{x}_j | C_j = i) = f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i$ is a parameter vector that f_i may depend upon.

- The characterization yields the expression for the population probability density function

$$f(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i),$$

which we call a finite mixture model, where we put all of the probabilities π_i and parameter vectors $\boldsymbol{\theta}_i$ into $\boldsymbol{\theta}$.

Finite Mixture Models

- Suppose that a priori to observing $\boldsymbol{X}_j$, $C_j = i$ with probability $\pi_i = \mathbb{P}(C_j = i)$, for $i = 1, \ldots, g$.

- Upon observing $\boldsymbol{X}_j$, if $C_j = i$, then $\boldsymbol{X}_j$ has (conditional) **component probability density function**

$$f(\boldsymbol{x}_j | C_j = i) = f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i)$$

  where $\boldsymbol{\theta}_i$ is a parameter vector that f_i may depend upon.

- The characterization yields the expression for the population probability density function

$$f(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{x}_j; \boldsymbol{\theta}_i),$$

  which we call a finite mixture model, where we put all of the probabilities π_i and parameter vectors $\boldsymbol{\theta}_i$ into $\boldsymbol{\theta}$.

Bayes-Optimal Clustering

- Suppose that we observe an observation $\boldsymbol{X}^*$ from the population with known form $f(\boldsymbol{x};\boldsymbol{\theta})$ and **parameter vector** $\boldsymbol{\theta}$, but unknown class label C^*.
- The **Bayes-optimal rule** for clustering is to estimate C^* by

$$\hat{c}(\boldsymbol{X}^*) = \arg\max_{i=1,\ldots,g} \frac{f_i(\boldsymbol{X}^*;\boldsymbol{\theta}_i)}{\sum_{k=1}^{g} \pi_k f_k(\boldsymbol{X}^*;\boldsymbol{\theta}_k)}.$$

- The clustering $\hat{c}(\boldsymbol{X}^*)$ maximizes the a posteriori density

$$f(C^* = \hat{c}(\boldsymbol{X}^*)|\boldsymbol{X}^*;\boldsymbol{\theta}) = \frac{f_i(\boldsymbol{X}^*;\boldsymbol{\theta}_i)}{f(\boldsymbol{X}^*;\boldsymbol{\theta})},$$

over all possible clusterings $\hat{c}(\boldsymbol{X}^*) = 1,\ldots,g$.

Bayes-Optimal Clustering

- Suppose that we observe an observation $\boldsymbol{X}^*$ from the population with known form $f(\boldsymbol{x};\boldsymbol{\theta})$ and **parameter vector** $\boldsymbol{\theta}$, but unknown class label C^*.
- The **Bayes-optimal rule** for clustering is to estimate C^* by

$$\hat{c}(\boldsymbol{X}^*) = \arg\max_{i=1,\ldots,g} \frac{f_i(\boldsymbol{X}^*;\boldsymbol{\theta}_i)}{\sum_{k=1}^{g} \pi_k f_k(\boldsymbol{X}^*;\boldsymbol{\theta}_k)}.$$

- The clustering $\hat{c}(\boldsymbol{X}^*)$ maximizes the a posteriori density

$$f(C^* = \hat{c}(\boldsymbol{X}^*)|\boldsymbol{X}^*;\boldsymbol{\theta}) = \frac{f_i(\boldsymbol{X}^*;\boldsymbol{\theta}_i)}{f(\boldsymbol{X}^*;\boldsymbol{\theta})},$$

over all possible clusterings $\hat{c}(\boldsymbol{X}^*) = 1,\ldots,g$.

Bayes-Optimal Clustering

- Suppose that we observe an observation $\boldsymbol{X}^*$ from the population with known form $f(\boldsymbol{x};\boldsymbol{\theta})$ and **parameter vector** $\boldsymbol{\theta}$, but unknown class label C^*.
- The **Bayes-optimal rule** for clustering is to estimate C^* by

$$\hat{c}(\boldsymbol{X}^*) = \arg\max_{i=1,\ldots,g} \frac{f_i(\boldsymbol{X}^*;\boldsymbol{\theta}_i)}{\sum_{k=1}^{g} \pi_k f_k(\boldsymbol{X}^*;\boldsymbol{\theta}_k)}.$$

- The clustering $\hat{c}(\boldsymbol{X}^*)$ maximizes the a posteriori density

$$f(C^* = \hat{c}(\boldsymbol{X}^*)|\boldsymbol{X}^*;\boldsymbol{\theta}) = \frac{f_i(\boldsymbol{X}^*;\boldsymbol{\theta}_i)}{f(\boldsymbol{X}^*;\boldsymbol{\theta})},$$

over all possible clusterings $\hat{c}(\boldsymbol{X}^*) = 1,\ldots,g$.

Optimality of $\hat{c}(\boldsymbol{X}^*)$

Theorem 1

The Bayes-optimal clustering rule is optimal in the sense that $\hat{c}(\boldsymbol{X}^*)$ *minimizes the loss function*

$$\mathscr{L}(c) = \mathbb{P}(c(\boldsymbol{X}^*) \neq C^*)$$

over all possible assignment functions c*. That is, if* c *is any other clustering rule, then* $\mathscr{L}(\hat{c}) \leq \mathscr{L}(c)$ *[Wasserman, 2004, Thm. 22.6].*

Mixture of Gaussian Distributions

- Suppose that we are operating in the real vector space $\mathbb{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$.

- A common choice for component density functions $f(\boldsymbol{x}_j | C_j = i)$ in such a space is the Gaussian density

$$\phi(\boldsymbol{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = |2\pi\Sigma|^{-1/2} \exp\left[-\frac{(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)}{2} \right],$$

  where $\boldsymbol{\mu}_i \in \mathbb{R}^d$ is the mean vector and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{d \times d}$ is the positive-definite covariance matrix.

- We say that

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

  is a **Gaussian mixture model**, where we put the parameter components π_i, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ in $\boldsymbol{\theta}$.

Mixture of Gaussian Distributions

- Suppose that we are operating in the real vector space $\mathbb{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$.

- A common choice for component density functions $f(\boldsymbol{x}_j | C_j = i)$ in such a space is the Gaussian density

$$\phi(\boldsymbol{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = |2\pi\Sigma|^{-1/2} \exp\left[-\frac{(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)}{2} \right],$$

  where $\boldsymbol{\mu}_i \in \mathbb{R}^d$ is the mean vector and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{d \times d}$ is the positive-definite covariance matrix.

- We say that

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

  is a **Gaussian mixture model**, where we put the parameter components π_i, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ in $\boldsymbol{\theta}$.

Mixture of Gaussian Distributions

- Suppose that we are operating in the real vector space $\mathbb{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$.

- A common choice for component density functions $f(\boldsymbol{x}_j | C_j = i)$ in such a space is the Gaussian density

$$\phi(\boldsymbol{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = |2\pi\Sigma|^{-1/2} \exp\left[-\frac{(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)}{2} \right],$$

  where $\boldsymbol{\mu}_i \in \mathbb{R}^d$ is the mean vector and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{d \times d}$ is the positive-definite covariance matrix.

- We say that

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

  is a **Gaussian mixture model**, where we put the parameter components π_i, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ in $\boldsymbol{\theta}$.

Gaussian Distributions

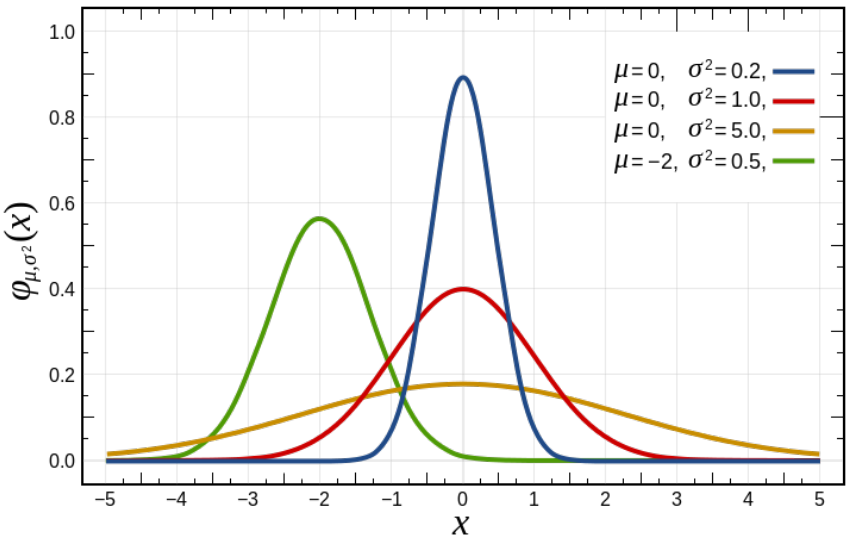

Figure 5: Gaussian density functions in $\mathbb{R}$. [Wikipedia]

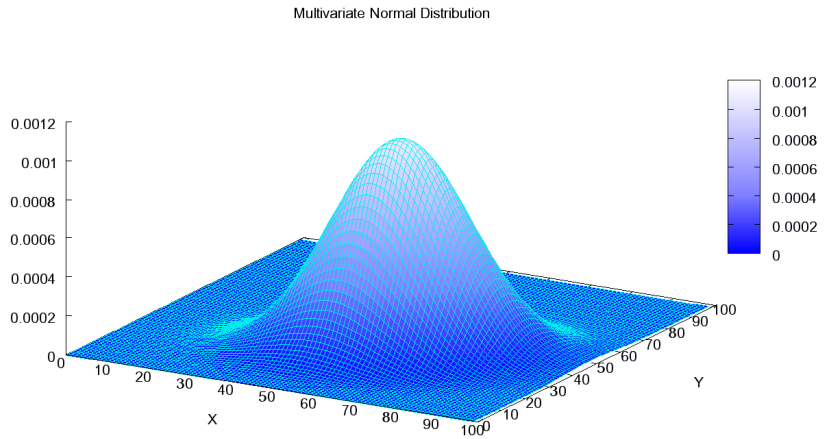

Figure 6: A Gaussian density function in $\mathbb{R}^2$. [Wikipedia]

GMM Clustering of Iris Data

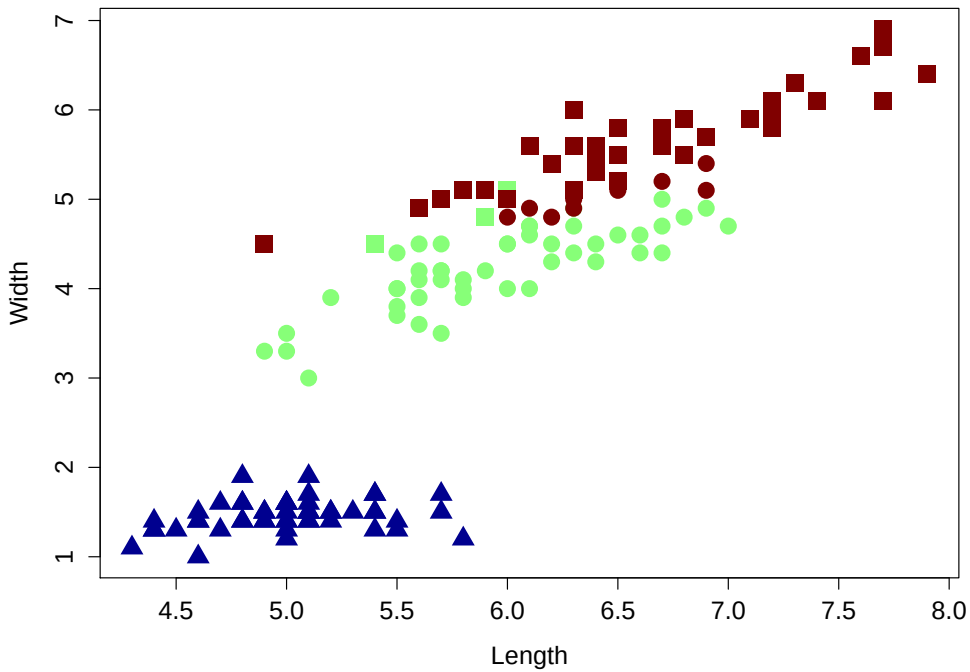

Figure 7: Blue: setosa. Green: versicolor. Maroon: virginica. Different shapes indicate three different clusters.

Goodness-of-Clustering

- Given the true labels, a measure of goodness-of-clustering to the true labels is the ARI (adjusted-Rand index) of [Hubert and Arabie, 1985]:

$$\text{ARI} \approx \frac{A+B}{\#\{\text{All pairs of observations}\}},$$

where

$$A = \#\{\text{Pairs correctly in same cluster}\}$$

and

$$B = \#\{\text{Pairs correctly in diff. cluster}\}.$$

- ARI of GMM clustering is 0.79 (max of 1).

Goodness-of-Clustering

- Given the true labels, a measure of goodness-of-clustering to the true labels is the ARI (adjusted-Rand index) of [Hubert and Arabie, 1985]:

$$\text{ARI} \approx \frac{A+B}{\#\{\text{All pairs of observations}\}},$$

where

$$A = \#\{\text{Pairs correctly in same cluster}\}$$

and

$$B = \#\{\text{Pairs correctly in diff. cluster}\}.$$

- ARI of GMM clustering is 0.79 (max of 1).

K-Means Clustering of Iris Data

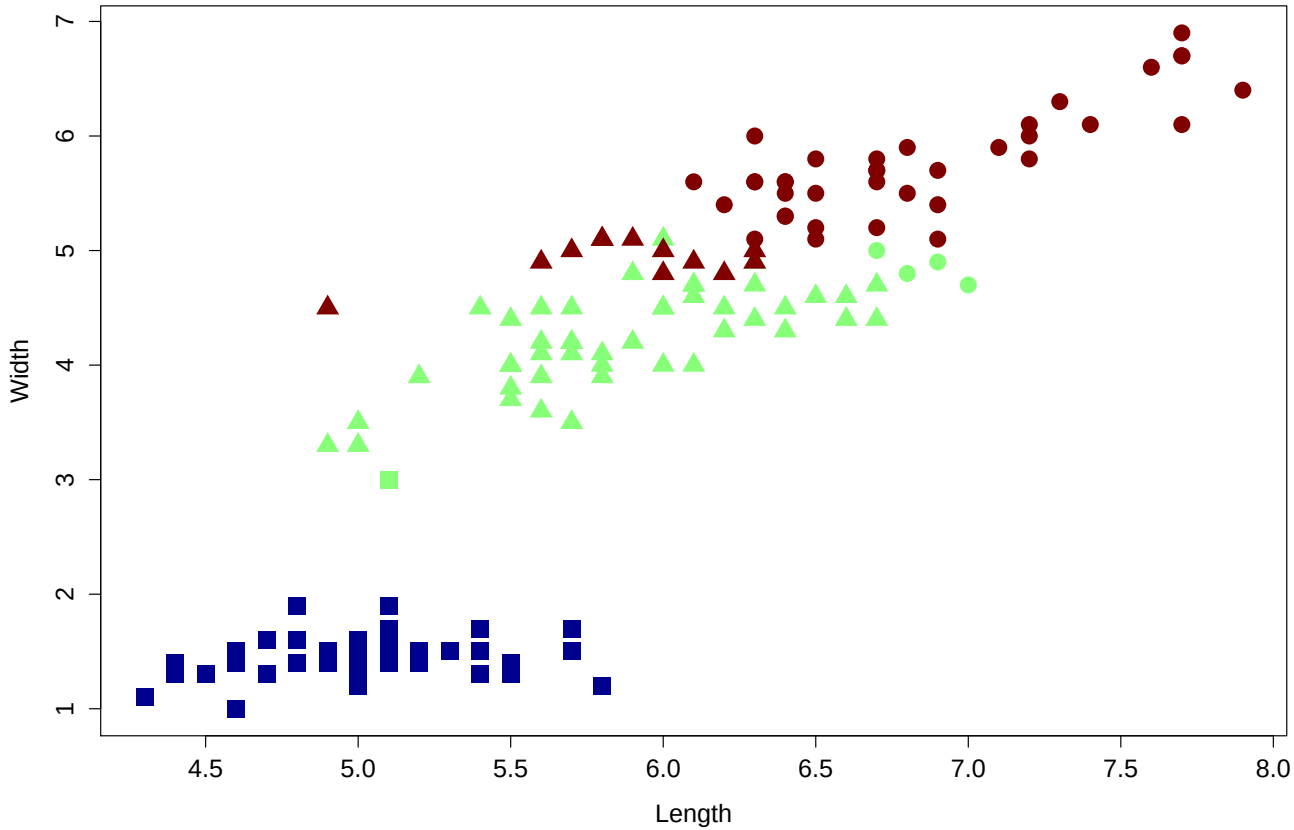

Figure 8: Blue: setosa. Green: versicolor. Maroon: virginica. Different shapes indicate three different clusters. Clustering ARI=0.70.

Estimated GMM

- Population density function is estimated to be

$$f\left(\boldsymbol{x};\hat{\boldsymbol{\theta}}\right)=\pi_1\phi\left(\boldsymbol{x};\hat{\boldsymbol{\mu}}_1,\hat{\boldsymbol{\Sigma}}_1\right)+\pi_2\phi\left(\boldsymbol{x};\hat{\boldsymbol{\mu}}_2,\hat{\boldsymbol{\Sigma}}_2\right)+\pi_3\phi\left(\boldsymbol{x};\hat{\boldsymbol{\mu}}_3,\hat{\boldsymbol{\Sigma}}_3\right)$$

where $\hat{\pi}_1=0.302$, $\hat{\pi}_2=0.364$, and $\hat{\pi}_3=0.334$;

$$\hat{\boldsymbol{\mu}}_1=\begin{bmatrix}6.54\\6.57\end{bmatrix},\ \hat{\boldsymbol{\mu}}_2=\begin{bmatrix}6.02\\4.36\end{bmatrix},\text{ and }\hat{\boldsymbol{\mu}}_3=\begin{bmatrix}5.01\\1.46\end{bmatrix};\text{ and}$$

$$\hat{\boldsymbol{\Sigma}}_1=\begin{bmatrix}0.480&0.373\\0.373&0.349\end{bmatrix},\ \hat{\boldsymbol{\Sigma}}_2=\begin{bmatrix}0.288&0.237\\0.237&0.290\end{bmatrix},$$

and

$$\hat{\boldsymbol{\Sigma}}_3=\begin{bmatrix}0.124&0.016\\0.016&0.030\end{bmatrix}.$$

Random Data from Estimated Model

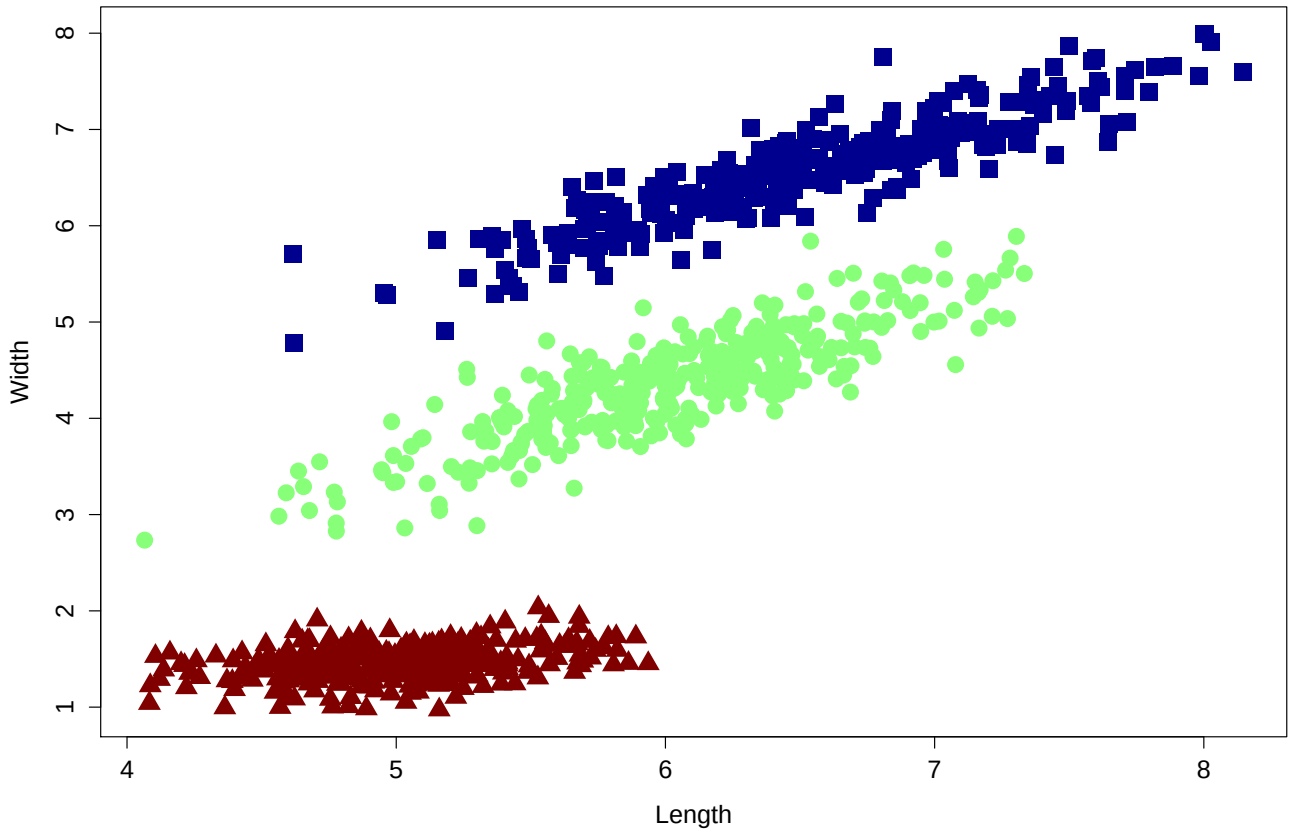

Figure 9: 1000 randomly generated observations from the estimated GMM. Different shapes indicate three different clusters.

Maximum Marginal-likelihood Estimation

- Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ be an ID random sample from a population with GMM density $f(\boldsymbol{x}; \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is unknown.
- The marginal-likelihood and log-marginal-likelihood functions are

$$L_n(\boldsymbol{\theta}) = \prod_{j=1}^{n} f(\boldsymbol{x}; \boldsymbol{\theta}_0) = \prod_{j=1}^{n} \sum_{i=1}^{g} \pi_{i,0} \phi\left(\boldsymbol{x}; \boldsymbol{\mu}_{i,0}, \boldsymbol{\Sigma}_{i,0}\right),$$

and

$$\ell_n(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log f(\boldsymbol{x}; \boldsymbol{\theta}_0) = \sum_{j=1}^{n} \log \sum_{i=1}^{g} \pi_{i,0} \phi\left(\boldsymbol{x}; \boldsymbol{\mu}_{i,0}, \boldsymbol{\Sigma}_{i,0}\right).$$

- Define the MML (maximum marginal-likelihood) estimator as $\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is an appropriate root to the system $\nabla \ell_n(\boldsymbol{\theta}) = \mathbf{0}$.

Maximum Marginal-likelihood Estimation

- Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ be an ID random sample from a population with GMM density $f(\boldsymbol{x}; \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is unknown.
- The marginal-likelihood and log-marginal-likelihood functions are

$$L_n(\boldsymbol{\theta}) = \prod_{j=1}^{n} f(\boldsymbol{x}; \boldsymbol{\theta}_0) = \prod_{j=1}^{n} \sum_{i=1}^{g} \pi_{i,0} \phi\left(\boldsymbol{x}; \boldsymbol{\mu}_{i,0}, \boldsymbol{\Sigma}_{i,0}\right),$$

and

$$\ell_n(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log f(\boldsymbol{x}; \boldsymbol{\theta}_0) = \sum_{j=1}^{n} \log \sum_{i=1}^{g} \pi_{i,0} \phi\left(\boldsymbol{x}; \boldsymbol{\mu}_{i,0}, \boldsymbol{\Sigma}_{i,0}\right).$$

- Define the MML (maximum marginal-likelihood) estimator as $\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is an appropriate root to the system $\nabla \ell_n(\boldsymbol{\theta}) = \mathbf{0}$.

Maximum Marginal-likelihood Estimation

- Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ be an ID random sample from a population with GMM density $f(\boldsymbol{x}; \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is unknown.

- The marginal-likelihood and log-marginal-likelihood functions are

$$L_n(\boldsymbol{\theta}) = \prod_{j=1}^{n} f(\boldsymbol{x}; \boldsymbol{\theta}_0) = \prod_{j=1}^{n} \sum_{i=1}^{g} \pi_{i,0} \phi\left(\boldsymbol{x}; \boldsymbol{\mu}_{i,0}, \boldsymbol{\Sigma}_{i,0}\right),$$

  and

$$\ell_n(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log f(\boldsymbol{x}; \boldsymbol{\theta}_0) = \sum_{j=1}^{n} \log \sum_{i=1}^{g} \pi_{i,0} \phi\left(\boldsymbol{x}; \boldsymbol{\mu}_{i,0}, \boldsymbol{\Sigma}_{i,0}\right).$$

- Define the MML (maximum marginal-likelihood) estimator as $\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is an appropriate root to the system $\nabla \ell_n(\boldsymbol{\theta}) = \mathbf{0}$.

Minorization–Maximization Algorithm

- Since $\ell_n(\boldsymbol{\theta})$ contains the log-sum-exp form, it is not possible to solve $\nabla \ell_n(\boldsymbol{\theta}) = \mathbf{0}$ via elementary calculus arguments.
- A minorization–maximization algorithm (cf. [Hunter and Lange, 2004]) can be constructed for maximizing $\ell_n(\boldsymbol{\theta})$.
- Let $\boldsymbol{\theta}^{(0)}$ be some initial estimate, and let $\boldsymbol{\theta}^{(r)}$ be the rth iterate of the MM algorithm.
  - MM algorithms are iterative algorithms that generate stable and convergence sequences of estimates $\boldsymbol{\theta}^{(r)}$.

Minorization–Maximization Algorithm

- Since $\ell_n(\boldsymbol{\theta})$ contains the log-sum-exp form, it is not possible to solve $\nabla \ell_n(\boldsymbol{\theta}) = \mathbf{0}$ via elementary calculus arguments.
- A minorization–maximization algorithm (cf. [Hunter and Lange, 2004]) can be constructed for maximizing $\ell_n(\boldsymbol{\theta})$.
- Let $\boldsymbol{\theta}^{(0)}$ be some initial estimate, and let $\boldsymbol{\theta}^{(r)}$ be the rth iterate of the MM algorithm.
  - MM algorithms are iterative algorithms that generate stable and convergence sequences of estimates $\boldsymbol{\theta}^{(r)}$.

Minorization–Maximization Algorithm

- Since $\ell_n(\boldsymbol{\theta})$ contains the log-sum-exp form, it is not possible to solve $\nabla \ell_n(\boldsymbol{\theta}) = \mathbf{0}$ via elementary calculus arguments.
- A minorization–maximization algorithm (cf. [Hunter and Lange, 2004]) can be constructed for maximizing $\ell_n(\boldsymbol{\theta})$.
- Let $\boldsymbol{\theta}^{(0)}$ be some initial estimate, and let $\boldsymbol{\theta}^{(r)}$ be the rth iterate of the MM algorithm.
  - MM algorithms are iterative algorithms that generate stable and convergence sequences of estimates $\boldsymbol{\theta}^{(r)}$.

MM Algorithm for GMMs

1. At the $(r+1)$thiteration, compute

$$\pi_i^{(r+1)} = n^{-1} \sum_{j=1}^{n} \tau_i \left(\boldsymbol{x}_j; \boldsymbol{\theta}^{(r)} \right),$$

$$\boldsymbol{\mu}_i^{(r+1)} = \left[\sum_{j=1}^{n} \tau_i \left(\boldsymbol{x}_j; \boldsymbol{\theta}^{(r)} \right) \right]^{-1} \sum_{j=1}^{n} \tau_i \left(\boldsymbol{x}_j; \boldsymbol{\theta}^{(r)} \right) \boldsymbol{x}_j,$$

and

$$\boldsymbol{\Sigma}_i^{(r+1)} = \frac{\sum_{j=1}^{n} \tau_i \left(\boldsymbol{x}_j; \boldsymbol{\theta}^{(r)} \right) \left[\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(r+1)} \right] \left[\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(r+1)} \right]^{\top}}{\sum_{j=1}^{n} \tau_i \left(\boldsymbol{x}_j; \boldsymbol{\theta}^{(r)} \right)},$$

where $\tau_i \left(\boldsymbol{x}_j; \boldsymbol{\theta} \right) = \pi_i \phi \left(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \right) / f \left(\boldsymbol{x}; \boldsymbol{\theta} \right)$.

2. Put $\pi_i^{(r+1)}$, $\boldsymbol{\mu}_i^{(r+1)}$, and $\boldsymbol{\Sigma}_i^{(r+1)}$ into $\boldsymbol{\theta}^{(r+1)}$.
3. Stop when $\ell_n \left(\boldsymbol{\theta}^{(r+1)} \right) - \ell_n \left(\boldsymbol{\theta}^{(r)} \right) < \text{TOL}$ for some $\text{TOL} > 0$.

Convergence of MM Algorithm

- Upon stopping the algorithm, we declare the final iterate the MML estimate $\hat{\boldsymbol{\theta}}$.
- Let $\lim_{r\to\infty} \boldsymbol{\theta}^{(r)} = \boldsymbol{\theta}^{(\infty)}$ (or $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^{(\infty)}$, as TOL $\to 0$).
- Using the theorems of [Razaviyayn et al., 2013], we have the following result.

Theorem 2

Let $\boldsymbol{\theta}^{(r)}$ be a sequence of MM algorithm iterates that converge to some limit point $\boldsymbol{\theta}^{(\infty)}$. The sequence $\boldsymbol{\theta}^{(r)}$ has the following properties:

1. *The sequence $\ell_n\left(\boldsymbol{\theta}^{(r)}\right)$ is monotonically increasing as r increases.*
2. *The limit point $\boldsymbol{\theta}^{(\infty)}$ is a stationary point of $\ell_n(\boldsymbol{\theta})$.*

Convergence of MM Algorithm

- Upon stopping the algorithm, we declare the final iterate the MML estimate $\hat{\boldsymbol{\theta}}$.
- Let $\lim_{r\to\infty} \boldsymbol{\theta}^{(r)} = \boldsymbol{\theta}^{(\infty)}$ (or $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^{(\infty)}$, as TOL $\to 0$).
- Using the theorems of [Razaviyayn et al., 2013], we have the following result.

Theorem 2

Let $\boldsymbol{\theta}^{(r)}$ be a sequence of MM algorithm iterates that converge to some limit point $\boldsymbol{\theta}^{(\infty)}$. The sequence $\boldsymbol{\theta}^{(r)}$ has the following properties:

1. *The sequence $\ell_n\left(\boldsymbol{\theta}^{(r)}\right)$ is monotonically increasing as r increases.*
2. *The limit point $\boldsymbol{\theta}^{(\infty)}$ is a stationary point of $\ell_n(\boldsymbol{\theta})$.*

Convergence of MM Algorithm

- Upon stopping the algorithm, we declare the final iterate the MML estimate $\hat{\boldsymbol{\theta}}$.
- Let $\lim_{r\to\infty} \boldsymbol{\theta}^{(r)} = \boldsymbol{\theta}^{(\infty)}$ (or $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^{(\infty)}$, as TOL $\to 0$).
- Using the theorems of [Razaviyayn et al., 2013], we have the following result.

Theorem 2

Let $\boldsymbol{\theta}^{(r)}$ be a sequence of MM algorithm iterates that converge to some limit point $\boldsymbol{\theta}^{(\infty)}$. The sequence $\boldsymbol{\theta}^{(r)}$ has the following properties:

1. *The sequence $\ell_n\left(\boldsymbol{\theta}^{(r)}\right)$ is monotonically increasing as r increases.*
2. *The limit point $\boldsymbol{\theta}^{(\infty)}$ is a stationary point of $\ell_n(\boldsymbol{\theta})$.*

Asymptotics of MML Estimator

- Extremum estimator theory of [Amemiya, 1985] suggests the following result.

Theorem 3

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be an ID, stationary, and ergodic (or α-mixing) random sample, such that for each j, $\boldsymbol{X}_j$ arises from a population with density function $f(\boldsymbol{x}_j; \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is a strict-local maximizer of $\mathbb{E} \log f(\boldsymbol{X}_j; \boldsymbol{\theta})$. If $\Theta_n = \{\boldsymbol{\theta} : \nabla \ell_n = 0\}$ (where we take $\Theta_n = \{\bar{\boldsymbol{\theta}}\}$, for some $\bar{\boldsymbol{\theta}} \in \Theta$, if $\nabla \ell_n = 0$ has no solution), then for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left[\inf_{\boldsymbol{\theta} \in \Theta_n} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T (\boldsymbol{\theta} - \boldsymbol{\theta}_0) > \varepsilon\right] = 0.$$

- If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are not too dependent on one another, then there exists a consistent root to $\nabla \ell_n = 0$, in the sense that the set of roots contains a sequence that converge to a strict-local maximizer of $\mathbb{E} \log f(\boldsymbol{X}_j; \boldsymbol{\theta})$, as n grows.

Asymptotics of MML Estimator

- Extremum estimator theory of [Amemiya, 1985] suggests the following result.

Theorem 3

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be an ID, stationary, and ergodic (or α-mixing) random sample, such that for each j, $\boldsymbol{X}_j$ arises from a population with density function $f(\boldsymbol{x}_j; \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is a strict-local maximizer of $\mathbb{E} \log f(\boldsymbol{X}_j; \boldsymbol{\theta})$. If $\Theta_n = \{\boldsymbol{\theta} : \nabla \ell_n = 0\}$ (where we take $\Theta_n = \{\bar{\boldsymbol{\theta}}\}$, for some $\bar{\boldsymbol{\theta}} \in \Theta$, if $\nabla \ell_n = 0$ has no solution), then for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left[\inf_{\boldsymbol{\theta} \in \Theta_n} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T (\boldsymbol{\theta} - \boldsymbol{\theta}_0) > \varepsilon\right] = 0.$$

- If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are not too dependent on one another, then there exists a consistent root to $\nabla \ell_n = 0$, in the sense that the set of roots contains a sequence that converge to a strict-local maximizer of $\mathbb{E} \log f(\boldsymbol{X}_j; \boldsymbol{\theta})$, as n grows.

Asymptotics of MML Estimator

- Extremum estimator theory of [Amemiya, 1985] suggests the following result.

Theorem 3

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be an ID, stationary, and ergodic (or α-mixing) random sample, such that for each j, $\boldsymbol{X}_j$ arises from a population with density function $f(\boldsymbol{x}_j; \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is a strict-local maximizer of $\mathbb{E} \log f(\boldsymbol{X}_j; \boldsymbol{\theta})$. If $\Theta_n = \{\boldsymbol{\theta} : \nabla \ell_n = 0\}$ (where we take $\Theta_n = \{\bar{\boldsymbol{\theta}}\}$, for some $\bar{\boldsymbol{\theta}} \in \Theta$, if $\nabla \ell_n = 0$ has no solution), then for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left[\inf_{\boldsymbol{\theta} \in \Theta_n} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T (\boldsymbol{\theta} - \boldsymbol{\theta}_0) > \varepsilon\right] = 0.$$

- If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are not too dependent on one another, then there exists a consistent root to $\nabla \ell_n = 0$, in the sense that the set of roots contains a sequence that converge to a strict-local maximizer of $\mathbb{E} \log f(\boldsymbol{X}_j; \boldsymbol{\theta})$, as n grows.

Ergodicity

- We say that $\boldsymbol{X}_j$ is **stationary** if "the joint distribution of similarly spaced elements in the sequence is identical, regardless of the absolute location of those elements within the sequence".

- Let $\boldsymbol{X}_j$ be a stationary sequence and let T be a measure-preserving transformation (technical; cf. [White, 2001]). We say that $\boldsymbol{X}_j$ is **ergodic** if

$$\lim_{n\to\infty} n^{-1} \sum_{j=1}^{n} \mathbb{P}\left(F \cap T^{\top} G\right) = \mathbb{P}(F)\mathbb{P}(G)$$

  for all events $F, G \in \mathscr{F}$, where $\mathscr{F}$ is the σ-algebra generated by the sequence $\boldsymbol{X}_j$.

  - Ergodicity is a form of "average asymptotic independence".

Ergodicity

- We say that $\boldsymbol{X}_j$ is **stationary** if "the joint distribution of similarly spaced elements in the sequence is identical, regardless of the absolute location of those elements within the sequence".

- Let $\boldsymbol{X}_j$ be a stationary sequence and let T be a measure-preserving transformation (technical; cf. [White, 2001]). We say that $\boldsymbol{X}_j$ is **ergodic** if

$$\lim_{n\to\infty} n^{-1} \sum_{j=1}^{n} \mathbb{P}\left(F \cap T^{\top} G\right) = \mathbb{P}(F)\mathbb{P}(G)$$

  for all events $F, G \in \mathscr{F}$, where $\mathscr{F}$ is the σ-algebra generated by the sequence $\boldsymbol{X}_j$.

  - Ergodicity is a form of "average asymptotic independence".

Ergodic Theorem

Theorem 4

If $\boldsymbol{X}_j$ is stationary and ergodic with finite expectation $\mathbb{E}(\boldsymbol{X}_j)$, then

$$n^{-1} \sum_{j=1}^{n} \boldsymbol{X}_j \to \mathbb{E}(\boldsymbol{X}_j) .$$

Further, if h is a measurable function and

$$\boldsymbol{Y}_j = h(\ldots, \boldsymbol{X}_{j-1}, \boldsymbol{X}_j, \boldsymbol{X}_{j+1}, \ldots)$$

then $\boldsymbol{Y}_j$ is stationary if $\boldsymbol{X}_j$ is stationary, and $\boldsymbol{Y}_j$ is stationary and ergodic if $\boldsymbol{X}_j$ is stationary and ergodic. [White, 2001, Thms. 3.34 and 3.35]

α-Mixing

- Let

$$\alpha(m) = \sup_j \left\{ |\mathbb{P}(F \cap G) - \mathbb{P}(F)\mathbb{P}(G)| : F \in \mathscr{F}_{-\infty}^{j}, G \in \mathscr{F}_{j+m}^{\infty} \right\},$$

where $\mathscr{F}_a^b$ is the σ-algebra generated by $\{\boldsymbol{X}_a, ..., \boldsymbol{X}_b\}$.

- If $\boldsymbol{X}_j$ is such that $\alpha(m) \to \infty$ as $m \to \infty$, then we say that $\boldsymbol{X}_j$ is an α-mixing (or **strongly-mixing**) sequence.

α-Mixing

- Let

$$\alpha(m) = \sup_j \left\{ |\mathbb{P}(F \cap G) - \mathbb{P}(F)\mathbb{P}(G)| : F \in \mathscr{F}_{-\infty}^{j}, G \in \mathscr{F}_{j+m}^{\infty} \right\},$$

where $\mathscr{F}_a^b$ is the σ-algebra generated by $\{\boldsymbol{X}_a, ..., \boldsymbol{X}_b\}$.

- If $\boldsymbol{X}_j$ is such that $\alpha(m) \to \infty$ as $m \to \infty$, then we say that $\boldsymbol{X}_j$ is an α-mixing (or **strongly-mixing**) sequence.

Useful α-Mixing Theorems

Theorem 5

If $\boldsymbol{X}_j$ is stationary and strongly-mixing sequence, then $\boldsymbol{X}_j$ is stationary and ergodic.

- If $\boldsymbol{X}_j$ and $\boldsymbol{X}_k$ are independent when $|j-k| > M$, for all $j, k \in \mathbb{Z}$ and some $M < \infty$, then we say that the sequence $\boldsymbol{X}_j$ is M-dependent.

Theorem 6

If $\boldsymbol{X}_j$ is M-dependent, then $\boldsymbol{X}_j$ is strongly-mixing (cf. [Bradley, 1989]).

- Theorems 4–6 imply that a stationary and M-dependent sequence $\boldsymbol{X}_j$ will fulfill the conditions of Theorem 3.

Useful α-Mixing Theorems

Theorem 5

If $\boldsymbol{X}_j$ is stationary and strongly-mixing sequence, then $\boldsymbol{X}_j$ is stationary and ergodic.

- If $\boldsymbol{X}_j$ and $\boldsymbol{X}_k$ are independent when $|j-k| > M$, for all $j, k \in \mathbb{Z}$ and some $M < \infty$, then we say that the sequence $\boldsymbol{X}_j$ is M-dependent.

Theorem 6

If $\boldsymbol{X}_j$ is M-dependent, then $\boldsymbol{X}_j$ is strongly-mixing (cf. [Bradley, 1989]).

- Theorems 4–6 imply that a stationary and M-dependent sequence $\boldsymbol{X}_j$ will fulfill the conditions of Theorem 3.

Useful α-Mixing Theorems

Theorem 5

If $\boldsymbol{X}_j$ is stationary and strongly-mixing sequence, then $\boldsymbol{X}_j$ is stationary and ergodic.

- If $\boldsymbol{X}_j$ and $\boldsymbol{X}_k$ are independent when $|j-k| > M$, for all $j, k \in \mathbb{Z}$ and some $M < \infty$, then we say that the sequence $\boldsymbol{X}_j$ is M-dependent.

Theorem 6

If $\boldsymbol{X}_j$ is M-dependent, then $\boldsymbol{X}_j$ is strongly-mixing (cf. [Bradley, 1989]).

- Theorems 4–6 imply that a stationary and M-dependent sequence $\boldsymbol{X}_j$ will fulfill the conditions of Theorem 3.

Useful α-Mixing Theorems

Theorem 5

If $\boldsymbol{X}_j$ is stationary and strongly-mixing sequence, then $\boldsymbol{X}_j$ is stationary and ergodic.

- If $\boldsymbol{X}_j$ and $\boldsymbol{X}_k$ are independent when $|j-k| > M$, for all $j, k \in \mathbb{Z}$ and some $M < \infty$, then we say that the sequence $\boldsymbol{X}_j$ is M-dependent.

Theorem 6

If $\boldsymbol{X}_j$ is M-dependent, then $\boldsymbol{X}_j$ is strongly-mixing (cf. [Bradley, 1989]).

- Theorems 4–6 imply that a stationary and M-dependent sequence $\boldsymbol{X}_j$ will fulfill the conditions of Theorem 3.

Model Selection

- Suppose that the data $\boldsymbol{X}_j$ is generated from a GMM with unknown number of subpopulations g_0, and unknown parameter vector $\boldsymbol{\theta}_{0,g_0}$.
- Suppose that it is known that $g_0 \in \{\gamma_1, \gamma_2, ...\} = \mathbb{G}$.
- We can estimate g_0 by $\hat{g}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) = \hat{G}$, where

$$\hat{G} = \arg\min_{g \in \mathbb{G}} \; -2 \times \ell_n\left(\hat{\boldsymbol{\theta}}_g\right) + \mathscr{D}_g \log n,$$

  where $\hat{\boldsymbol{\theta}}_g$ is the MML estimate, under the assumption that g is the number of subpopulations, and

$$\mathscr{D}_g = g \times \left[\frac{d \times (d-1)}{2} + 2d + 1\right] - 1.$$

- $\hat{G}$ is the BIC (**Bayesian information criterion**) rule.

Model Selection

- Suppose that the data $\boldsymbol{X}_j$ is generated from a GMM with unknown number of subpopulations g_0, and unknown parameter vector $\boldsymbol{\theta}_{0,g_0}$.
- Suppose that it is known that $g_0 \in \{\gamma_1, \gamma_2, ...\} = \mathbb{G}$.
- We can estimate g_0 by $\hat{g}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) = \hat{G}$, where

$$\hat{G} = \arg\min_{g \in \mathbb{G}} \; -2 \times \ell_n\left(\hat{\boldsymbol{\theta}}_g\right) + \mathscr{D}_g \log n,$$

where $\hat{\boldsymbol{\theta}}_g$ is the MML estimate, under the assumption that g is the number of subpopulations, and

$$\mathscr{D}_g = g \times \left[\frac{d \times (d-1)}{2} + 2d + 1\right] - 1.$$

- $\hat{G}$ is the BIC (**Bayesian information criterion**) rule.

Model Selection

- Suppose that the data $\boldsymbol{X}_j$ is generated from a GMM with unknown number of subpopulations g_0, and unknown parameter vector $\boldsymbol{\theta}_{0,g_0}$.
- Suppose that it is known that $g_0 \in \{\gamma_1, \gamma_2, ...\} = \mathbb{G}$.
- We can estimate g_0 by $\hat{g}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) = \hat{G}$, where

$$\hat{G} = \arg\min_{g \in \mathbb{G}} \; -2 \times \ell_n\left(\hat{\boldsymbol{\theta}}_g\right) + \mathscr{D}_g \log n,$$

where $\hat{\boldsymbol{\theta}}_g$ is the MML estimate, under the assumption that g is the number of subpopulations, and

$$\mathscr{D}_g = g \times \left[\frac{d \times (d-1)}{2} + 2d + 1\right] - 1.$$

- $\hat{G}$ is the BIC (**Bayesian information criterion**) rule.

Model Selection

- Suppose that the data $\boldsymbol{X}_j$ is generated from a GMM with unknown number of subpopulations g_0, and unknown parameter vector $\boldsymbol{\theta}_{0,g_0}$.
- Suppose that it is known that $g_0 \in \{\gamma_1, \gamma_2, ...\} = \mathbb{G}$.
- We can estimate g_0 by $\hat{g}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) = \hat{G}$, where

$$\hat{G} = \arg\min_{g \in \mathbb{G}} \; -2 \times \ell_n\left(\hat{\boldsymbol{\theta}}_g\right) + \mathscr{D}_g \log n,$$

where $\hat{\boldsymbol{\theta}}_g$ is the MML estimate, under the assumption that g is the number of subpopulations, and

$$\mathscr{D}_g = g \times \left[\frac{d \times (d-1)}{2} + 2d + 1\right] - 1.$$

- $\hat{G}$ is the BIC (**Bayesian information criterion**) rule.

Summary of Model-Based Clustering via GMM

1. Determine a set of plausible number of clusters $\mathbb{G}$.
2. Use the MM algorithm to estimate the MML estimator $\hat{\boldsymbol{\theta}}_g$, for each $g \in \mathbb{G}$.
3. Estimate the number of clusters using the BIC rule, $\hat{G}$.
4. Using the GMM $f\left(\boldsymbol{X}; \boldsymbol{\theta}_{\hat{G}}\right)$ to estimate the cluster allocation of new observations (or observations used for estimation), $\boldsymbol{X}^*$ using the Bayes-optimal clustering rule, $\hat{c}\left(\boldsymbol{X}^*\right)$.

Neuroimaging Segmentation Example

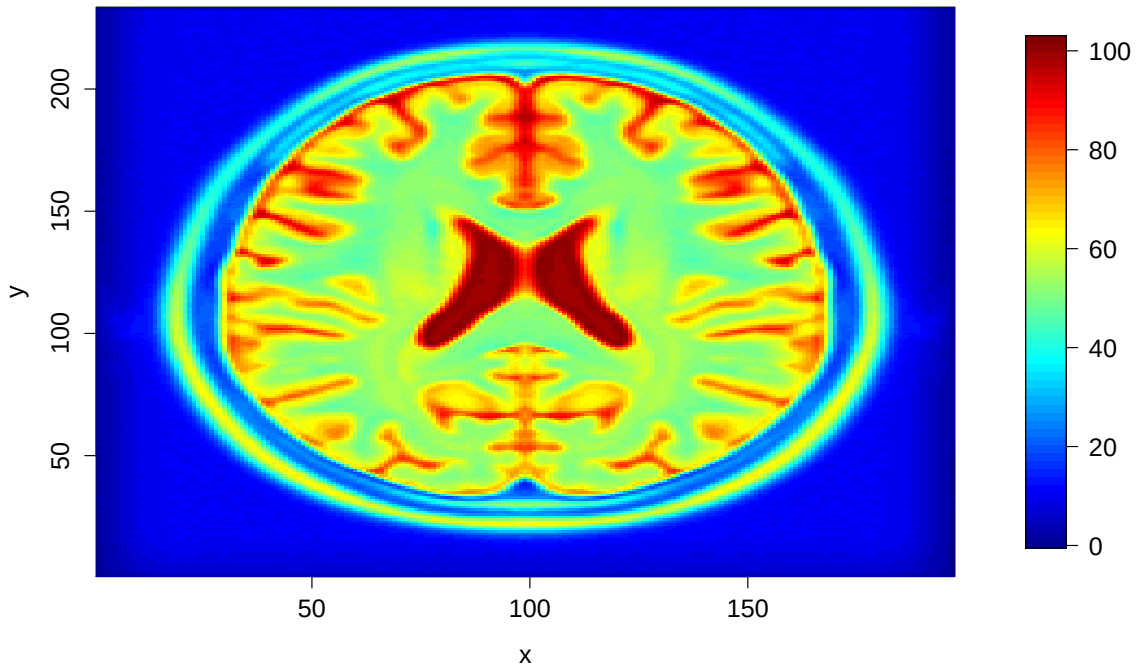

Figure 10: 95th z-slice of the ICBM 2009a nonlinear symmetric T2-weighted MRI atlas. [Fonov et al., 2011]

Image Detail

- Freely available from www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009.
- Full image is a nonlinear symmetric template atlas at the anatomic resolution $1 \times 1 \times 1$mm.
- Data volume of size $197 \times 233 \times 189$ $(x \times y \times z)$.
- Total volume of 8675289 voxels; z-slice area of $n = 45901$.
- At the 95th z-slice, the image intensities x_j are in the range $[0.27, 102.26]$, for $j = 1, \ldots, n$.

Image Detail

- Freely available from www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009.
- Full image is a nonlinear symmetric template atlas at the anatomic resolution $1 \times 1 \times 1$mm.
- Data volume of size $197 \times 233 \times 189$ $(x \times y \times z)$.
- Total volume of 8675289 voxels; z-slice area of $n = 45901$.
- At the 95th z-slice, the image intensities x_j are in the range $[0.27, 102.26]$, for $j = 1, \ldots, n$.

Image Detail

- Freely available from www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009.
- Full image is a nonlinear symmetric template atlas at the anatomic resolution $1 \times 1 \times 1$mm.
- Data volume of size $197 \times 233 \times 189$ $(x \times y \times z)$.
- Total volume of 8675289 voxels; z-slice area of $n = 45901$.
- At the 95th z-slice, the image intensities x_j are in the range $[0.27, 102.26]$, for $j = 1, \ldots, n$.

Image Detail

- Freely available from www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009.
- Full image is a nonlinear symmetric template atlas at the anatomic resolution $1 \times 1 \times 1$mm.
- Data volume of size $197 \times 233 \times 189$ $(x \times y \times z)$.
- Total volume of 8675289 voxels; z-slice area of $n = 45901$.
- At the 95th z-slice, the image intensities x_j are in the range $[0.27, 102.26]$, for $j = 1, \ldots, n$.

Image Detail

- Freely available from www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009.
- Full image is a nonlinear symmetric template atlas at the anatomic resolution $1 \times 1 \times 1$mm.
- Data volume of size $197 \times 233 \times 189$ $(x \times y \times z)$.
- Total volume of 8675289 voxels; z-slice area of $n = 45901$.
- At the 95th z-slice, the image intensities x_j are in the range $[0.27, 102.26]$, for $j = 1, ..., n$.

Marginal Density of 95th Slice

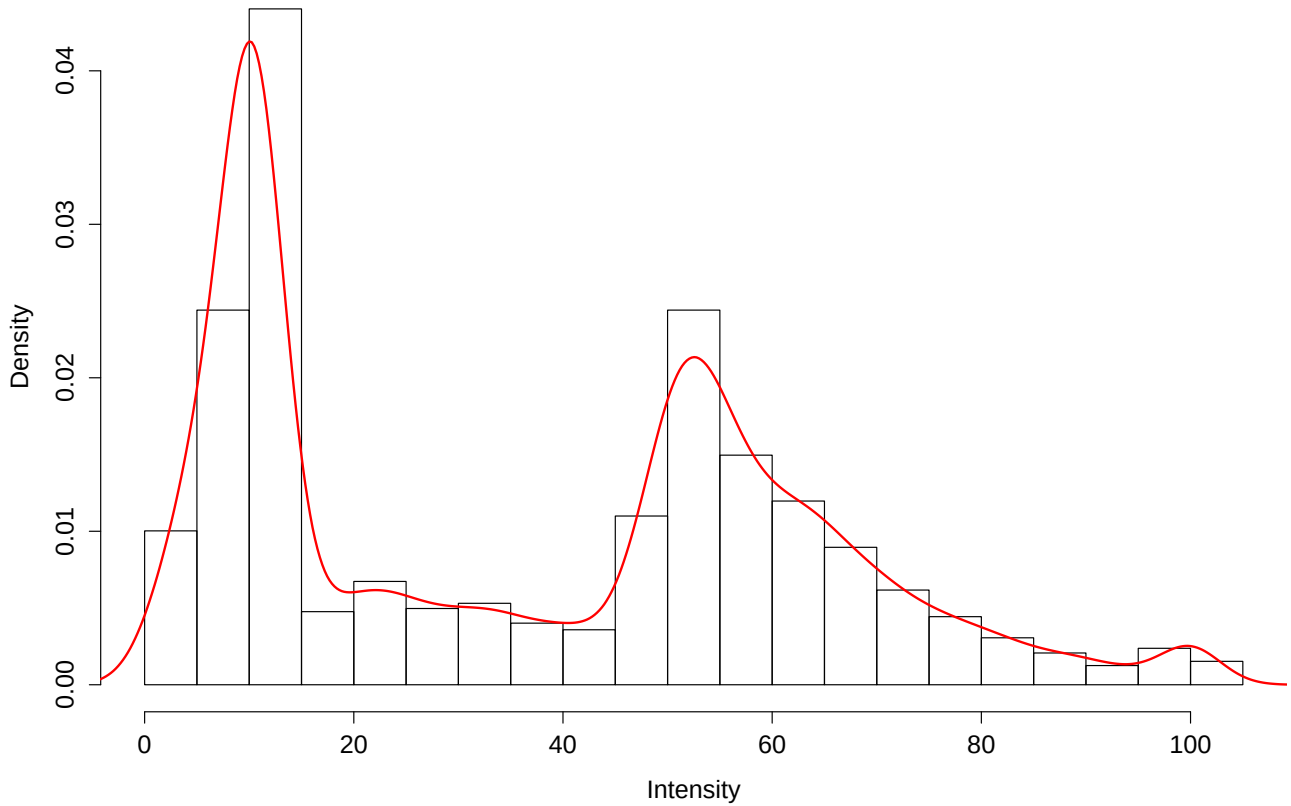

Figure 11: Marginal probability density (kernel density estimate) and histogram of the intensities on the 95th z-slice.

Observations from Density

- Density appears to show 2 main modes and flat area in the middle and end.
- Consistent with usual segmentation of human brain MRIs:
  - Grey matter, white matter, cerebrospinal fluid, and background.

Observations from Density

- Density appears to show 2 main modes and flat area in the middle and end.
- Consistent with usual segmentation of human brain MRIs:
  - Grey matter, white matter, cerebrospinal fluid, and background.

Model Selection via BIC

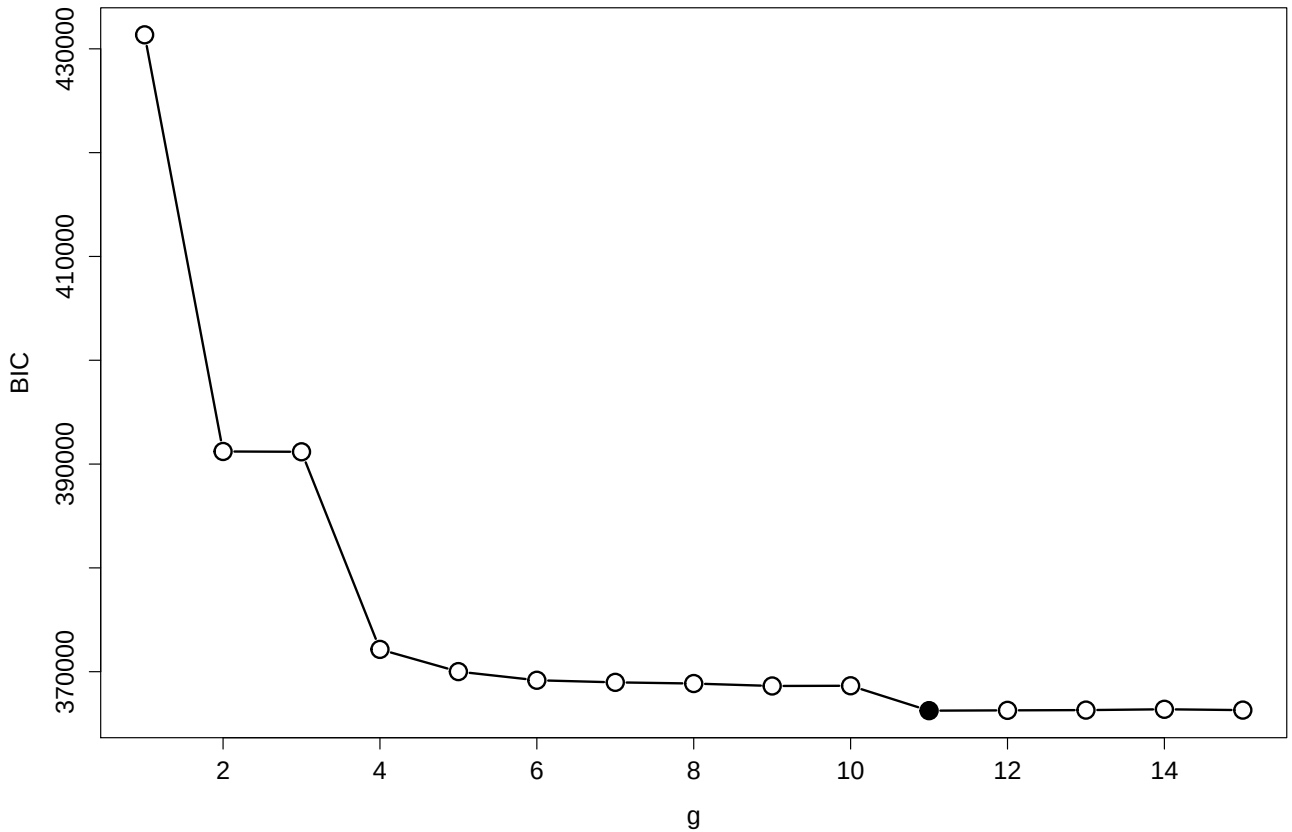

Figure 12: BIC values $-2 \times \ell_n\left(\hat{\boldsymbol{\theta}}_g\right) + \mathscr{D}_g \log n$ for conducting model selection. Colored dot indicates selected model.

Estimated Clustering

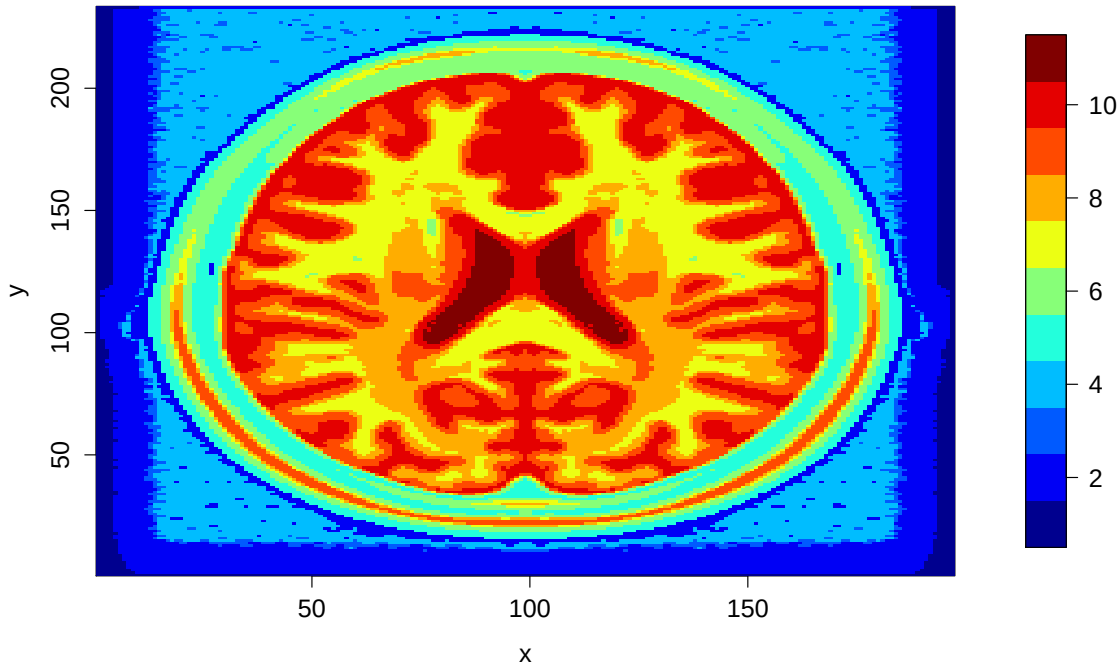

Figure 13: MML estimated clustering via the rule $\hat{c}(\boldsymbol{X}^*)$ obtained using $g = \hat{G} = 11$.

Estimated GMM

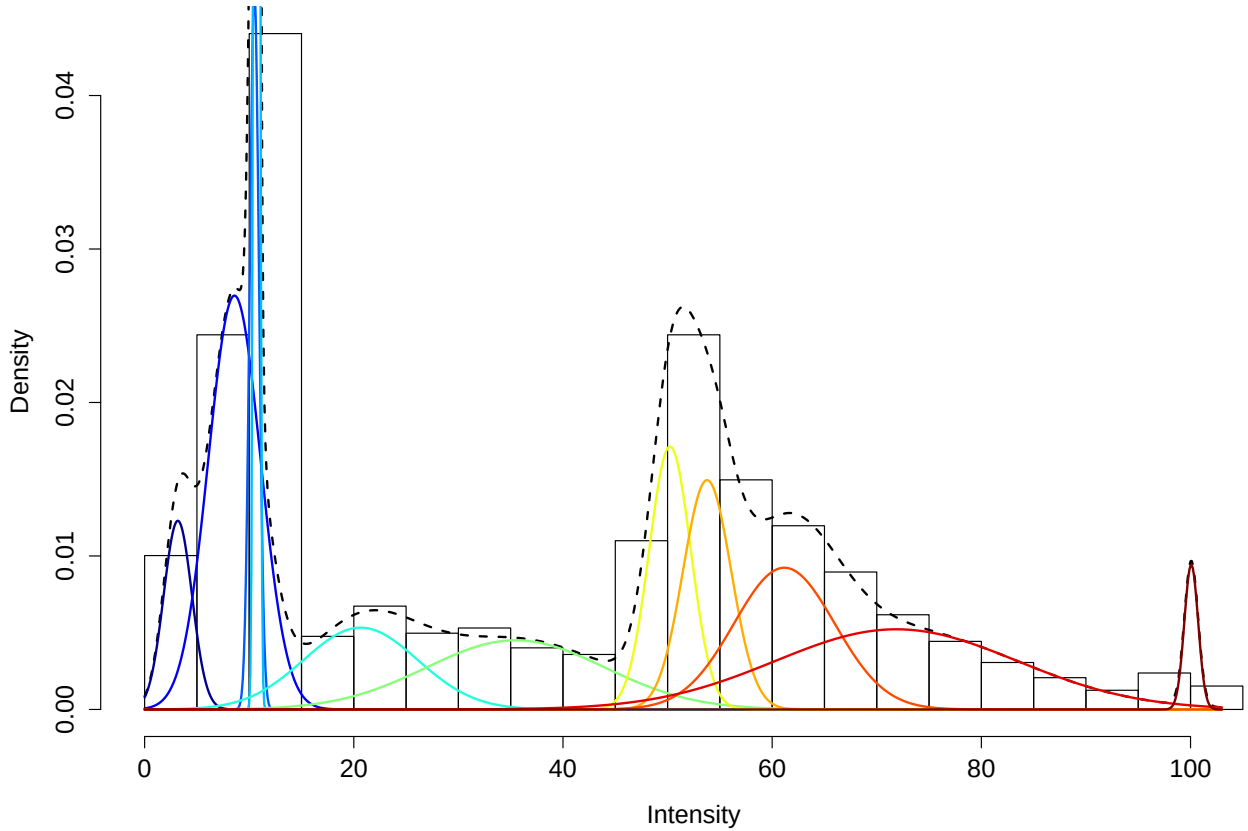

Figure 14: Estimated $g = 11$ components GMM. Black line indicates overall density $f\left(\boldsymbol{x}; \hat{\boldsymbol{\theta}}\right)$. Each color indicates a seperate component $\hat{\pi}_i \phi\left(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i\right)$.

Individual Clusters

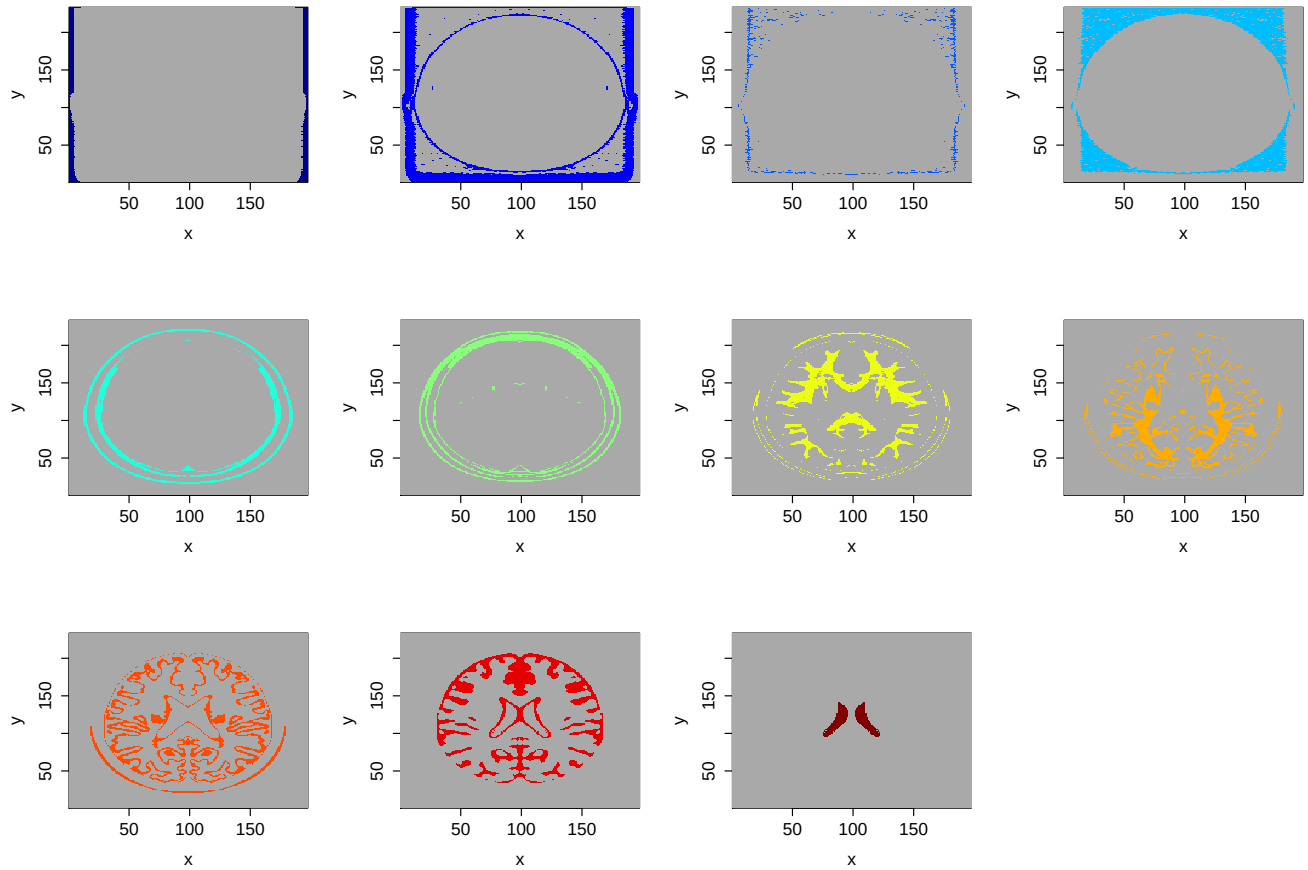

Figure 15: The panels running left to right, and top to bottom, display the location of the observations clustered to subpopulations $i = 1, \ldots, 11$.

Automatic Cluster Merging in GMMs

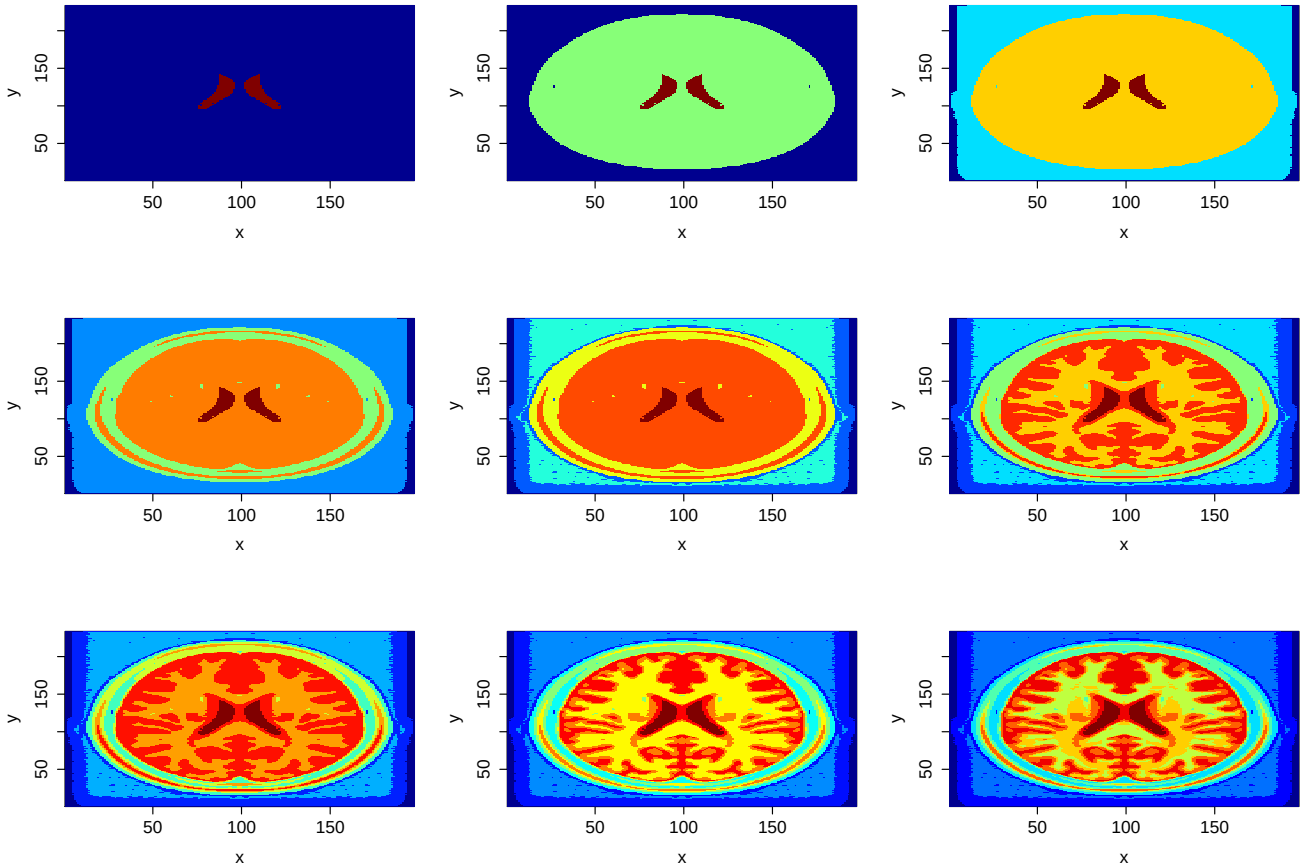

Figure 16: The panels running left to right, and top to bottom, display the clusterings after merging to a number of $2, \ldots, 10$ clusters, using the **Bhattacharyya distance method** (cf. [Hennig, 2010]).

Automatic Cluster Merging in GMMs—2

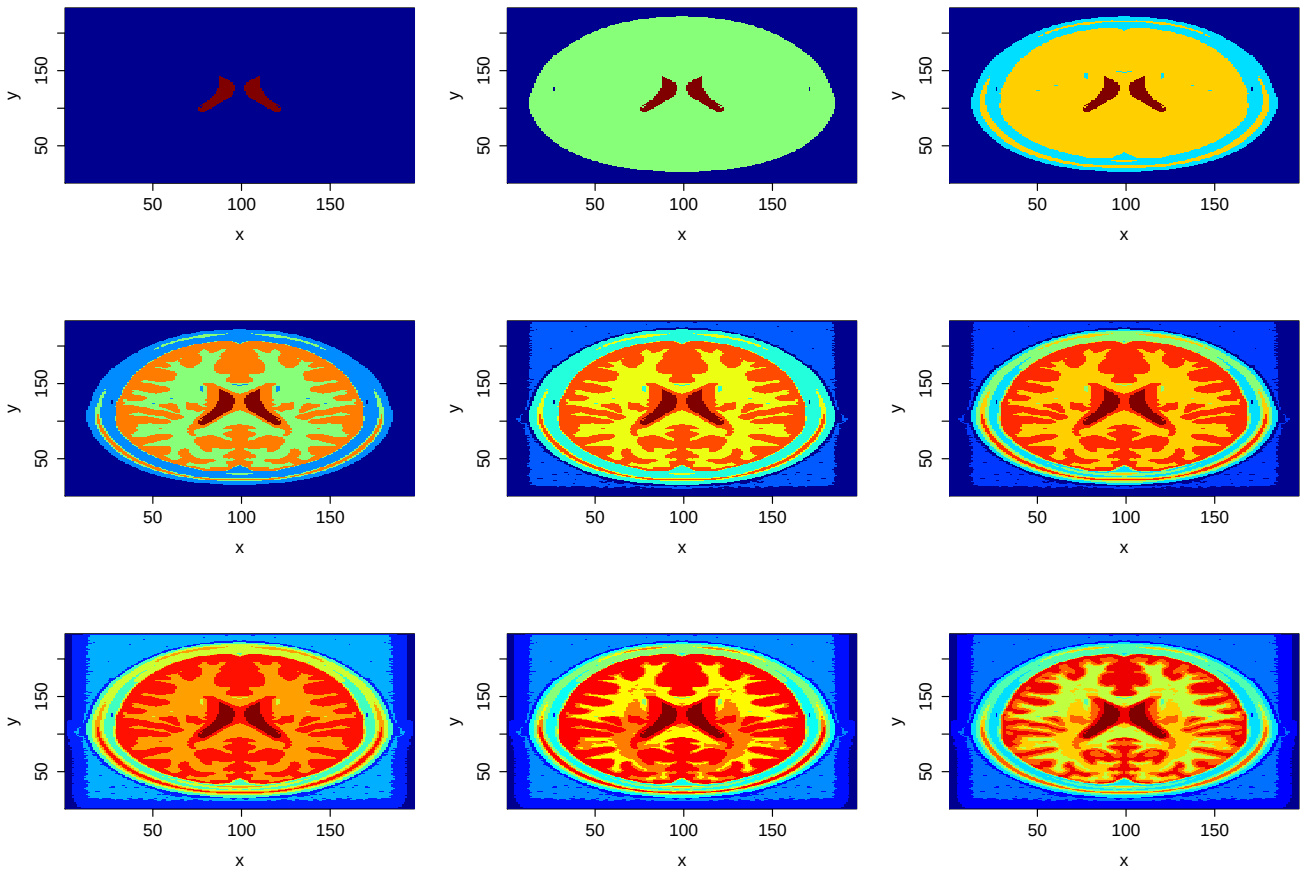

Figure 17: The panels running left to right, and top to bottom, display the clusterings after merging to a number of $2, \ldots, 10$ clusters, using the **DEMP** direct estimation of misclassification probability method (cf. [Hennig, 2010]).

DEMP with 5 Clusters

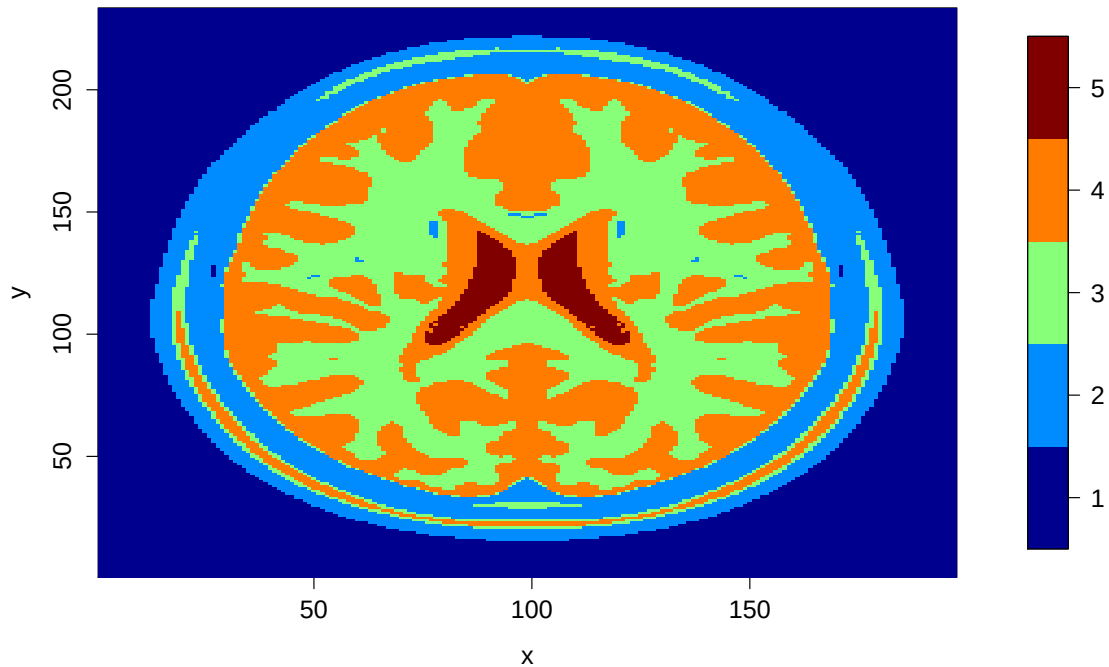

Figure 18: Location of clusters after automatic cluster merging from 11 to 5 clusters, via the DEMP method.

Extensions

- If the space of interest is $\mathbb{X} = \mathbb{R}^d$, then there are a variety of densities other than the Gaussian density functions to construct clustering with.
  - Mixtures of t-distributions. [Peel and McLachlan, 2000]
  - Skew normal and skew t-distributions. [Lee and McLachlan, 2014]
- If the space of interest is other than $\mathbb{R}^d$ (e.g. $\mathbb{X} = \mathbb{N}$, $\{1, 2, ..., g\}$, $\mathbb{R}^+$, or $[a, b]$), then it is possible to cluster using Poisson, multinomial, gamma, or beta mixture models (cf. [McLachlan and Peel, 2000, Ch. 6]).
- If the densities $f(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta})$ are dependent on some covariates $\boldsymbol{z} \in \mathbb{R}^q$, then mixtures of regressions or mixtures of experts can be used to conduct clustering (cf. [Nguyen, 2015]).

Extensions

- If the space of interest is $\mathbb{X} = \mathbb{R}^d$, then there are a variety of densities other than the Gaussian density functions to construct clustering with.
  - Mixtures of t-distributions. [Peel and McLachlan, 2000]
  - Skew normal and skew t-distributions. [Lee and McLachlan, 2014]
- If the space of interest is other than $\mathbb{R}^d$ (e.g. $\mathbb{X} = \mathbb{N}$, $\{1, 2, ..., g\}$, $\mathbb{R}^+$, or $[a, b]$), then it is possible to cluster using Poisson, multinomial, gamma, or beta mixture models (cf. [McLachlan and Peel, 2000, Ch. 6]).
- If the densities $f(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta})$ are dependent on some covariates $\boldsymbol{z} \in \mathbb{R}^q$, then mixtures of regressions or mixtures of experts can be used to conduct clustering (cf. [Nguyen, 2015]).

Extensions

- If the space of interest is $\mathbb{X} = \mathbb{R}^d$, then there are a variety of densities other than the Gaussian density functions to construct clustering with.
  - Mixtures of t-distributions. [Peel and McLachlan, 2000]
  - Skew normal and skew t-distributions. [Lee and McLachlan, 2014]
- If the space of interest is other than $\mathbb{R}^d$ (e.g. $\mathbb{X} = \mathbb{N}$, $\{1, 2, ..., g\}$, $\mathbb{R}^+$, or $[a, b]$), then it is possible to cluster using Poisson, multinomial, gamma, or beta mixture models (cf. [McLachlan and Peel, 2000, Ch. 6]).
- If the densities $f(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta})$ are dependent on some covariates $\boldsymbol{z} \in \mathbb{R}^q$, then mixtures of regressions or mixtures of experts can be used to conduct clustering (cf. [Nguyen, 2015]).

Next Lecture

- Introduce Markov random fields for categorical data.
- Consider the maximum pseudolikelihood estimator for MRF models.
- Discuss Minorization–Maximization algorithms in detail.
- Construct an MM algorithm for the MPL estimation of MRF models.
- Present theorems on the convergence and optimality of the MM algorithm.
- Discuss the asymptotic properties of the MPL estimator.
- Present a method for model selection between MRF models.

Next Lecture

- Introduce Markov random fields for categorical data.
- Consider the maximum pseudolikelihood estimator for MRF models.
- Discuss Minorization–Maximization algorithms in detail.
- Construct an MM algorithm for the MPL estimation of MRF models.
- Present theorems on the convergence and optimality of the MM algorithm.
- Discuss the asymptotic properties of the MPL estimator.
- Present a method for model selection between MRF models.

Next Lecture

- Introduce Markov random fields for categorical data.
- Consider the maximum pseudolikelihood estimator for MRF models.
- Discuss Minorization–Maximization algorithms in detail.
- Construct an MM algorithm for the MPL estimation of MRF models.
- Present theorems on the convergence and optimality of the MM algorithm.
- Discuss the asymptotic properties of the MPL estimator.
- Present a method for model selection between MRF models.

Next Lecture

- Introduce Markov random fields for categorical data.
- Consider the maximum pseudolikelihood estimator for MRF models.
- Discuss Minorization–Maximization algorithms in detail.
- Construct an MM algorithm for the MPL estimation of MRF models.
- Present theorems on the convergence and optimality of the MM algorithm.
- Discuss the asymptotic properties of the MPL estimator.
- Present a method for model selection between MRF models.

Next Lecture

- Introduce Markov random fields for categorical data.
- Consider the maximum pseudolikelihood estimator for MRF models.
- Discuss Minorization–Maximization algorithms in detail.
- Construct an MM algorithm for the MPL estimation of MRF models.
- Present theorems on the convergence and optimality of the MM algorithm.
- Discuss the asymptotic properties of the MPL estimator.
- Present a method for model selection between MRF models.

Next Lecture

- Introduce Markov random fields for categorical data.
- Consider the maximum pseudolikelihood estimator for MRF models.
- Discuss Minorization–Maximization algorithms in detail.
- Construct an MM algorithm for the MPL estimation of MRF models.
- Present theorems on the convergence and optimality of the MM algorithm.
- Discuss the asymptotic properties of the MPL estimator.
- Present a method for model selection between MRF models.

Next Lecture

- Introduce Markov random fields for categorical data.
- Consider the maximum pseudolikelihood estimator for MRF models.
- Discuss Minorization–Maximization algorithms in detail.
- Construct an MM algorithm for the MPL estimation of MRF models.
- Present theorems on the convergence and optimality of the MM algorithm.
- Discuss the asymptotic properties of the MPL estimator.
- Present a method for model selection between MRF models.

Questions?

tinyurl.com/hiendnguyen

References I

Amemiya, T. (1985).

Advanced Econometrics.
Harvard University Press, Cambridge.

Bradley, R. C. (1989).

A caution on mixing conditions for random fields.
Statistics and Probability Letters, 8:489–491.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011).

Cluster Analysis.
Wiley, New York.

Fisher, R. A. (1936).

The use of multiple measurements in taxonomic problems.
Annals of Eugenics, 7:179–188.

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., Collins, D. L., and the Brain Development Cooperative Group (2011).

Unbiased average age-approriate altases for pediatric studies.
NeuroImage, 54:313–327.

Hennig, C. (2010).

Methos for merging Gaussian mixture components.
Advances in Data Analysis and Classification, 4:3–34.

Hubert, L. and Arabie, P. (1985).

Comparing partitions.
Journal of Classification, 2:193–218.

Hunter, D. R. and Lange, K. (2004).

A tutorial on MM algorithms.
The American Statistician, 58:30–37.

References II

Jain, J. K. and Dubes, R. C. (1988).
Algorithm for Clustering Data.
Prentice Hall, Englewood Cliffs.

Lee, S. X. and McLachlan, G. J. (2014).
Finite mixtures of multivaraite skew t-distributions: some recent and new results.
Statistics and Computing, 24:181–202.

McLachlan, G. J. and Peel, D. (2000).
Finite Mixture Models.
Wiley, New York.

Nguyen, H. D. (2015).
Finite Mixture Models for Regression Problems.
PhD thesis, University of Queensland.

Nguyen, H. D. and McLachlan, G. J. (2016).
Maximum likelihood estimation of triangular and polygonal distributions.
Computational Statistics and Data Analysis, 102:23–36.

Peel, D. and McLachlan, G. J. (2000).
Robust mixture modelling using the t distribution.
Statistics and Computing, 10:335–344.

Razaviyayn, M., Hong, M., and Luo, Z.-Q. (2013).
A unified convergence analysis of block successive minimization methods for nonsmooth optimization.
SIAM Journal of Optimization, 23:1126–1153.

Wasserman, L. (2004).
All Of Statistics: A Concise Course In Statistical Inference.
Springer, New York.

References III

White, H. (2001).
Asymptotic Theory For Econometricians.
Academic Press, San Diego.