Model-Based Methods for Clustering of Spatial Time Series Data—Lecture 2: Markov Random Fields

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

Session Content

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.

DEMP-Merged GMM Clustering with 5 Clusters

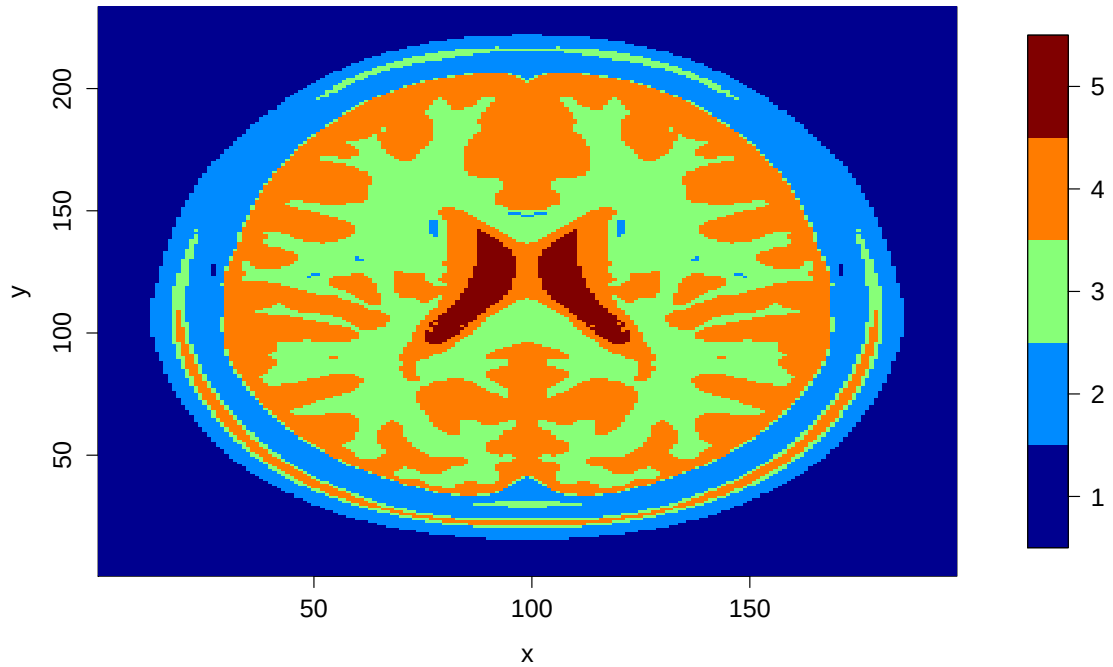

Figure 1: Location of clusters after automatic cluster merging of a $g = 11$ component GMM clustering to 5 clusters, via the DEMP method of [Hennig, 2000].

Infinite Square Lattice

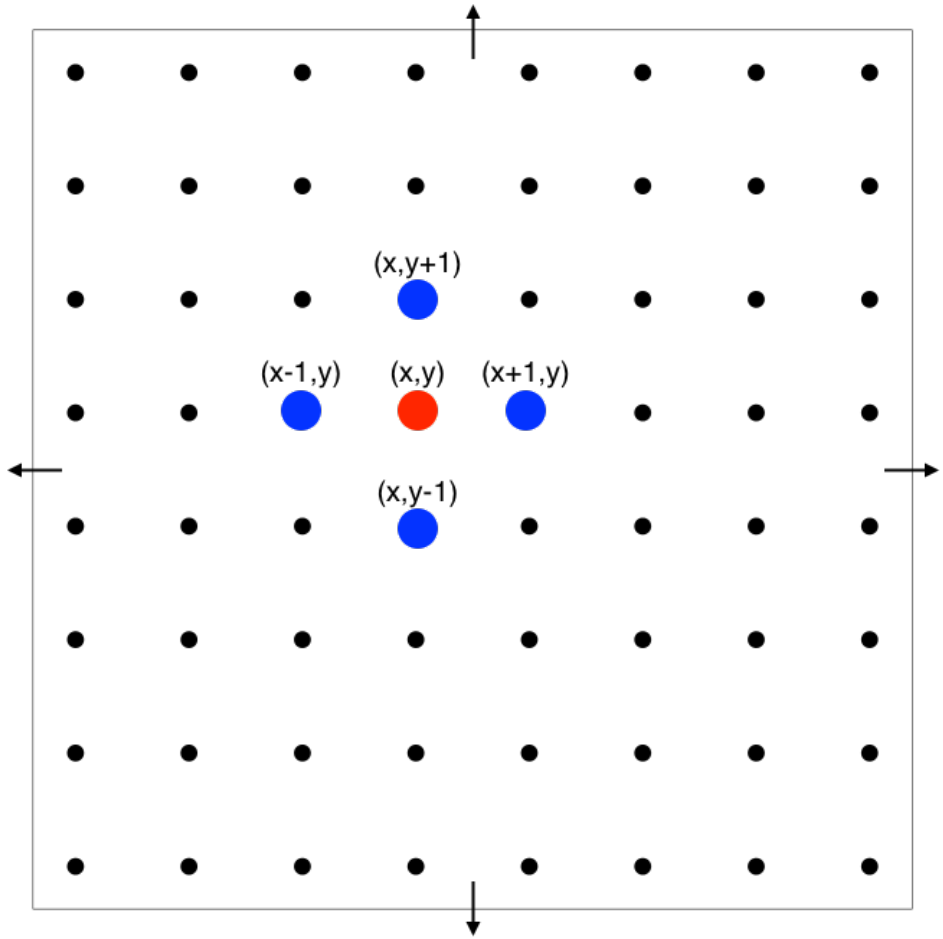

Figure 2: Schematic of an infinite square lattice.

Random Variable on the Infinite Square Lattice

- Let $\boldsymbol{C_z} \in \mathbb{K}$ be a random variable, where $\mathbb{K} \subset \mathbb{N}^q$ is a finite set of elements and $\boldsymbol{z} = (x, y) \in \mathbb{Z}^2$.
  - We say that $C_{\boldsymbol{z}}$ is a g-dimensional **categorical random variable** on the infinite square lattice if $\mathbb{K} = \{1, ..., g\}$.
- Let $N_{\boldsymbol{z}}$ denote the **neighborhood** around the coordinate $\boldsymbol{z}$, where the neighborhood is a set of valid locations on the square lattice that depend on $\boldsymbol{z}$.

Random Variable on the Infinite Square Lattice

- Let $\boldsymbol{C_z} \in \mathbb{K}$ be a random variable, where $\mathbb{K} \subset \mathbb{N}^q$ is a finite set of elements and $\boldsymbol{z} = (x, y) \in \mathbb{Z}^2$.
  - We say that $C_{\boldsymbol{z}}$ is a g-dimensional **categorical random variable** on the infinite square lattice if $\mathbb{K} = \{1, ..., g\}$.
- Let $N_{\boldsymbol{z}}$ denote the **neighborhood** around the coordinate $\boldsymbol{z}$, where the neighborhood is a set of valid locations on the square lattice that depend on $\boldsymbol{z}$.

Examples of Neighborhoods

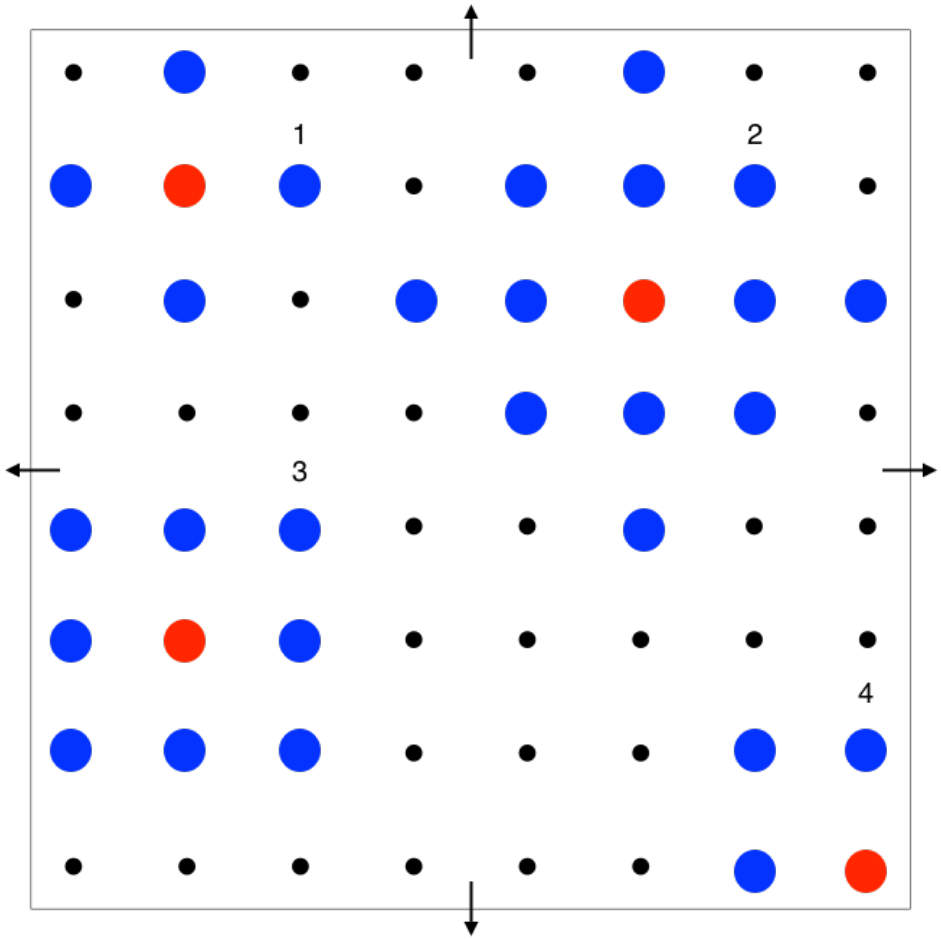

Figure 3: Example of neighborhoods on the square lattice. The red dot indicates the coordinate z and the blue dots indicate its neighbours N_z.

Examples of Neighborhoods—2

1. The **von Neumann neighborhood** of distance $d = 1$:

$$N_z = \{(i,j) : |x-i| + |y-j| \leq 1, (i,j) \neq (x,y)\}.$$

2. The von Neumann neighborhood of distance $d = 2$:

$$N_z = \{(i,j) : |x-i| + |y-j| \leq 2, (i,j) \neq (x,y)\}.$$

3. The **Moore neighborhood** of distance $d = 1$:

$$N_z = \{(i,j) : \max\{|x-i|, |y-j|\} \leq 1, (i,j) \neq (x,y)\}.$$

4. The neighborhood:

$$N_z = \{(x-1, y+1), (x-1, y), (x, y+1)\}.$$

Functions Over Neighborhoods

- Let $\mathbb{C}_{\boldsymbol{z}} = \left\{ C_{\boldsymbol{\zeta}} : \zeta \in N_{\boldsymbol{z}} \right\}$ be the set of random observations in the neighborhood of $\boldsymbol{z}$.
- Let $\boldsymbol{\eta}_{\boldsymbol{z}} = \boldsymbol{\eta}\left(\mathbb{C}_{\boldsymbol{z}}\right) \in \mathbb{R}^h$ be a h-dimensional function that takes a neighborhood a set containing the observations belonging to a neighborhood as input.

Functions Over Neighborhoods

- Let $\mathbb{C}_{\boldsymbol{z}} = \left\{ C_{\boldsymbol{\zeta}} : \zeta \in N_{\boldsymbol{z}} \right\}$ be the set of random observations in the neighborhood of $\boldsymbol{z}$.
- Let $\boldsymbol{\eta}_{\boldsymbol{z}} = \boldsymbol{\eta}\left(\mathbb{C}_{\boldsymbol{z}}\right) \in \mathbb{R}^h$ be a h-dimensional function that takes a neighborhood a set containing the observations belonging to a neighborhood as input.

Functions Over Neighborhoods—2

Example 1

Consider the g-dimensional categorical random variable $C_{\boldsymbol{z}}$. A potential function $\boldsymbol{\eta}_{\boldsymbol{z}}^{\top} = (\eta_{1,\boldsymbol{z}}, ..., \eta_{h,\boldsymbol{z}})$ may be such that $h = g$ and

$$\eta_{i,\boldsymbol{z}} = \frac{1}{|\mathbb{C}_{\boldsymbol{z}}|} \sum_{\boldsymbol{\zeta} \in \mathbb{C}_{\boldsymbol{z}}} \mathbb{I}\left\{C_{\boldsymbol{\zeta}} = i\right\}.$$

- The first and second functions counts the number of observations in each category, and the number of observations that are similar to $C_{\boldsymbol{z}}$, in the neighborhood of $\boldsymbol{z}$, respectively.

Functions Over Neighborhoods—2

Example 1

Consider the g-dimensional categorical random variable $C_{\boldsymbol{z}}$. A potential function $\boldsymbol{\eta}_{\boldsymbol{z}}^{\top} = (\eta_{1,\boldsymbol{z}}, ..., \eta_{h,\boldsymbol{z}})$ may be such that $h = g$ and

$$\eta_{i,\boldsymbol{z}} = \frac{1}{|\mathbb{C}_{\boldsymbol{z}}|} \sum_{\boldsymbol{\zeta} \in \mathbb{C}_{\boldsymbol{z}}} \mathbb{I}\left\{C_{\boldsymbol{\zeta}} = i\right\}.$$

- The first and second functions counts the number of observations in each category, and the number of observations that are similar to $C_{\boldsymbol{z}}$, in the neighborhood of $\boldsymbol{z}$, respectively.

Markov Random Field

- If the random variables $\boldsymbol{C}_{\boldsymbol{z}}$ for $\boldsymbol{z} \in \mathbb{Z}^2$ satisfy, then we call the conditional relationship

$$\mathbb{P}(\boldsymbol{C}_{\boldsymbol{z}} = \boldsymbol{\gamma} | \mathbb{C}_{\boldsymbol{z}}; \boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{\top} \boldsymbol{\eta}_{\boldsymbol{z}}\right)}{\sum_{\boldsymbol{\kappa} \in \mathbb{K}} \exp\left(\boldsymbol{\theta}_{\boldsymbol{\kappa}}^{\top} \boldsymbol{\eta}_{\boldsymbol{z}}\right)}$$

  a discrete-multinomial **Markov random field** (cf. [Geman and Graffigne, 1986]).

  - The model has parameter $\boldsymbol{\theta}$, which contains the parameter components $\boldsymbol{\theta}_{\boldsymbol{\kappa}}$for all $\boldsymbol{\kappa} \in \mathbb{K}$.

Markov Random Field—2

Example 2

Consider the g-dimensional categorical random variable C_z. a discrete-multinomial Markov random field can be defined via the conditional probabilities:

$$\mathbb{P}(C_z = i | \mathbb{C}_z; \boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta}_i^\top \boldsymbol{\eta}_z\right)}{\sum_{k=1}^{g} \exp\left(\boldsymbol{\theta}_k^\top \boldsymbol{\eta}_z\right)}.$$

- If $g = 2$, then we have the **discrete-logistic MRF**:

$$\mathbb{P}(C_z = 1 | \mathbb{C}_z; \boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta}_1^\top \boldsymbol{\eta}_z\right)}{1 + \exp\left(\boldsymbol{\theta}_1^\top \boldsymbol{\eta}_z\right)},$$

where we can set $\boldsymbol{\theta}_2 = \mathbf{0}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1$.

Markov Random Field—2

Example 2

Consider the g-dimensional categorical random variable C_z. a discrete-multinomial Markov random field can be defined via the conditional probabilities:

$$\mathbb{P}(C_z = i|\mathbb{C}_z;\boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta}_i^\top \boldsymbol{\eta}_z\right)}{\sum_{k=1}^{g}\exp\left(\boldsymbol{\theta}_k^\top \boldsymbol{\eta}_z\right)}.$$

- If $g = 2$, then we have the **discrete-logistic MRF**:

$$\mathbb{P}(C_z = 1|\mathbb{C}_z;\boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta}_1^\top \boldsymbol{\eta}_z\right)}{1+\exp\left(\boldsymbol{\theta}_1^\top \boldsymbol{\eta}_z\right)},$$

where we can set $\boldsymbol{\theta}_2 = \mathbf{0}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1$.

Maximum Pseudolikelihood Estimation

- Let $z_1, \ldots, z_n$ be a set of n coordinates in $\mathbb{Z}^2$ at which the random sample

$$c_{z_1}, \ldots, c_{z_n} = c_1, \ldots, c_n$$

corresponding to the variables $C_{z_1}, \ldots, C_{z_n}$ are observed.

- Suppose that $C_j \in \{1, \ldots, g\}$ (i.e. C_j is a g-dimensional categorical random variable) is realized from an MRF with neighborhood functions $\boldsymbol{\eta}_{z_j} = \boldsymbol{\eta}_j$ and parameter $\boldsymbol{\theta}_0$.

- We can estimate $\boldsymbol{\theta}_0$ by computing the **maximum pseudolikelihood** (MPL) **estimator**:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^g} \mathscr{P}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^g} \sum_{j=1}^{n} \sum_{i=1}^{g} \mathbb{I}\{c_j = i\} \log \mathbb{P}(C_j = i | \mathbb{C}_j; \boldsymbol{\theta}),$$

where $\mathscr{P}(\boldsymbol{\theta})$ is the **log-pseudolikelihood function** (e.g., see [Besag, 1974]).

Maximum Pseudolikelihood Estimation

- Let $z_1, \ldots, z_n$ be a set of n coordinates in $\mathbb{Z}^2$ at which the random sample

$$c_{z_1}, \ldots, c_{z_n} = c_1, \ldots, c_n$$

corresponding to the variables $C_{z_1}, \ldots, C_{z_n}$ are observed.

- Suppose that $C_j \in \{1, \ldots, g\}$ (i.e. C_j is a g-dimensional categorical random variable) is realized from an MRF with neighborhood functions $\boldsymbol{\eta}_{z_j} = \boldsymbol{\eta}_j$ and parameter $\boldsymbol{\theta}_0$.

- We can estimate $\boldsymbol{\theta}_0$ by computing the **maximum pseudolikelihood** (MPL) **estimator**:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^g} \mathscr{P}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^g} \sum_{j=1}^{n} \sum_{i=1}^{g} \mathbb{I}\{c_j = i\} \log \mathbb{P}(C_j = i | \mathbb{C}_j; \boldsymbol{\theta}),$$

where $\mathscr{P}(\boldsymbol{\theta})$ is the **log-pseudolikelihood function** (e.g., see [Besag, 1974]).

Maximum Pseudolikelihood Estimation

- Let $z_1, \ldots, z_n$ be a set of n coordinates in $\mathbb{Z}^2$ at which the random sample

$$c_{z_1}, \ldots, c_{z_n} = c_1, \ldots, c_n$$

corresponding to the variables $C_{z_1}, \ldots, C_{z_n}$ are observed.

- Suppose that $C_j \in \{1, \ldots, g\}$ (i.e. C_j is a g-dimensional categorical random variable) is realized from an MRF with neighborhood functions $\boldsymbol{\eta}_{z_j} = \boldsymbol{\eta}_j$ and parameter $\boldsymbol{\theta}_0$.

- We can estimate $\boldsymbol{\theta}_0$ by computing the **maximum pseudolikelihood** (MPL) **estimator**:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^g} \mathscr{P}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^g} \sum_{j=1}^{n} \sum_{i=1}^{g} \mathbb{I}\{c_j = i\} \log \mathbb{P}(C_j = i | \mathbb{C}_j; \boldsymbol{\theta}),$$

where $\mathscr{P}(\boldsymbol{\theta})$ is the **log-pseudolikelihood function** (e.g., see [Besag, 1974]).

Maximum Pseudolikelihood Estimation—2

- Expanded, the log-PL function can be written as

$$\begin{aligned}\mathscr{P}(\boldsymbol{\theta}) &= \sum_{j=1}^{n}\sum_{i=1}^{g}\mathbb{I}\{c_j = i\}\log\mathbb{P}(C_j = i|\mathbb{C}_j;\boldsymbol{\theta}) \\ &= \sum_{j=1}^{n}\sum_{i=1}^{g}\mathbb{I}\{c_j = i\}\,\boldsymbol{\theta}_i^{\top}\boldsymbol{\eta}_j - \sum_{j=1}^{n}\log\sum_{i=1}^{g}\exp\left(\boldsymbol{\theta}_i^{\top}\boldsymbol{\eta}_j\right),\end{aligned}$$

  where $\boldsymbol{\theta}_g = \mathbf{0}$, and we put $\boldsymbol{\theta}_i$ for $i = 1,\ldots,g-1$ into $\boldsymbol{\theta}$.

- The negative-log-sum-exp form is concave and linear compositions and additions are concave, thus $\mathscr{P}$ is concave in $\boldsymbol{\theta}$.

- We cannot solve the first order condition $\nabla\mathscr{P} = \mathbf{0}$ in closed form.

Maximum Pseudolikelihood Estimation—2

- Expanded, the log-PL function can be written as

$$\begin{aligned}\mathscr{P}(\boldsymbol{\theta}) &= \sum_{j=1}^{n}\sum_{i=1}^{g}\mathbb{I}\{c_j = i\}\log\mathbb{P}(C_j = i|\mathbb{C}_j;\boldsymbol{\theta}) \\ &= \sum_{j=1}^{n}\sum_{i=1}^{g}\mathbb{I}\{c_j = i\}\,\boldsymbol{\theta}_i^{\top}\boldsymbol{\eta}_j - \sum_{j=1}^{n}\log\sum_{i=1}^{g}\exp\left(\boldsymbol{\theta}_i^{\top}\boldsymbol{\eta}_j\right),\end{aligned}$$

  where $\boldsymbol{\theta}_g = \mathbf{0}$, and we put $\boldsymbol{\theta}_i$ for $i = 1, \ldots, g-1$ into $\boldsymbol{\theta}$.

- The negative-log-sum-exp form is concave and linear compositions and additions are concave, thus $\mathscr{P}$ is concave in $\boldsymbol{\theta}$.

- We cannot solve the first order condition $\nabla\mathscr{P} = \mathbf{0}$ in closed form.

Maximum Pseudolikelihood Estimation—2

- Expanded, the log-PL function can be written as

$$\begin{aligned}\mathscr{P}(\boldsymbol{\theta}) &= \sum_{j=1}^{n}\sum_{i=1}^{g}\mathbb{I}\{c_j = i\}\log\mathbb{P}(C_j = i|\mathbb{C}_j;\boldsymbol{\theta}) \\ &= \sum_{j=1}^{n}\sum_{i=1}^{g}\mathbb{I}\{c_j = i\}\,\boldsymbol{\theta}_i^{\top}\boldsymbol{\eta}_j - \sum_{j=1}^{n}\log\sum_{i=1}^{g}\exp\left(\boldsymbol{\theta}_i^{\top}\boldsymbol{\eta}_j\right),\end{aligned}$$

  where $\boldsymbol{\theta}_g = \mathbf{0}$, and we put $\boldsymbol{\theta}_i$ for $i = 1, \ldots, g-1$ into $\boldsymbol{\theta}$.

- The negative-log-sum-exp form is concave and linear compositions and additions are concave, thus $\mathscr{P}$ is concave in $\boldsymbol{\theta}$.

- We cannot solve the first order condition $\nabla\mathscr{P} = \mathbf{0}$ in closed form.

Minorization–Maximization Algorithm

- Suppose that we wish to maximize some difficult objective function $\mathscr{L}(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta$.
- Let $\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\psi})$ be an easy to maximize function over the parameter $\boldsymbol{\theta}$, conditioned on $\boldsymbol{\psi} \in \Theta$.
- If $\mathscr{Q}$ satisfies the conditions:

$$\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}) = \mathscr{L}(\boldsymbol{\theta}), \text{ for all } \boldsymbol{\theta} \in \Theta$$

and

$$\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\psi}) \leq \mathscr{L}(\boldsymbol{\theta}), \text{ for all } \boldsymbol{\theta} \neq \boldsymbol{\psi},$$

then we say that $\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\psi})$ is a **minorizer** at $\boldsymbol{\psi}$.

Minorization–Maximization Algorithm

- Suppose that we wish to maximize some difficult objective function $\mathscr{L}(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta$.
- Let $\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\psi})$ be an easy to maximize function over the parameter $\boldsymbol{\theta}$, conditioned on $\boldsymbol{\psi} \in \Theta$.
- If $\mathscr{Q}$ satisfies the conditions:

$$\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}) = \mathscr{L}(\boldsymbol{\theta}), \text{ for all } \boldsymbol{\theta} \in \Theta$$

and

$$\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\psi}) \leq \mathscr{L}(\boldsymbol{\theta}), \text{ for all } \boldsymbol{\theta} \neq \boldsymbol{\psi},$$

then we say that $\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\psi})$ is a **minorizer** at $\boldsymbol{\psi}$.

Minorization–Maximization Algorithm

- Suppose that we wish to maximize some difficult objective function $\mathscr{L}(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta$.
- Let $\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\psi})$ be an easy to maximize function over the parameter $\boldsymbol{\theta}$, conditioned on $\boldsymbol{\psi} \in \Theta$.
- If $\mathscr{Q}$ satisfies the conditions:

$$\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}) = \mathscr{L}(\boldsymbol{\theta}), \text{ for all } \boldsymbol{\theta} \in \Theta$$

and

$$\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\psi}) \leq \mathscr{L}(\boldsymbol{\theta}), \text{ for all } \boldsymbol{\theta} \neq \boldsymbol{\psi},$$

then we say that $\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\psi})$ is a **minorizer** at $\boldsymbol{\psi}$.

Minorization–Maximization Algorithm—2

- Let $\boldsymbol{\theta}^{(0)}$ be some initial values, and let $\boldsymbol{\theta}^{(r)}$ be rth iteration of the MM algorithm.
- The **MM algorithm** is defined via the update scheme:

$$\boldsymbol{\theta}^{(r+1)} = \arg\max_{\boldsymbol{\theta} \in \not\leq} \mathscr{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right).$$

Theorem 3

The MM algorithm generates sequences of iterates $\boldsymbol{\theta}^{(k)}$ that satisfy the inequalities:

$$\mathscr{L}\left(\boldsymbol{\theta}^{(r)}\right) \leq \mathscr{Q}\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)}\right) \leq \mathscr{Q}\left(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}\right) \leq \mathscr{L}\left(\boldsymbol{\theta}^{(r+1)}\right).$$

- The MM algorithm is strictly increasing in the objective sequence $\mathscr{L}\left(\boldsymbol{\theta}^{(r)}\right)$.

Minorization–Maximization Algorithm—2

- Let $\boldsymbol{\theta}^{(0)}$ be some initial values, and let $\boldsymbol{\theta}^{(r)}$ be rth iteration of the MM algorithm.
- The **MM algorithm** is defined via the update scheme:

$$\boldsymbol{\theta}^{(r+1)} = \arg\max_{\boldsymbol{\theta} \in \not\leq} \mathscr{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right).$$

Theorem 3

The MM algorithm generates sequences of iterates $\boldsymbol{\theta}^{(k)}$ that satisfy the inequalities:

$$\mathscr{L}\left(\boldsymbol{\theta}^{(r)}\right) \leq \mathscr{Q}\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)}\right) \leq \mathscr{Q}\left(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}\right) \leq \mathscr{L}\left(\boldsymbol{\theta}^{(r+1)}\right).$$

- The MM algorithm is strictly increasing in the objective sequence $\mathscr{L}\left(\boldsymbol{\theta}^{(r)}\right)$.

Minorization–Maximization Algorithm—2

- Let $\boldsymbol{\theta}^{(0)}$ be some initial values, and let $\boldsymbol{\theta}^{(r)}$ be rth iteration of the MM algorithm.
- The **MM algorithm** is defined via the update scheme:

$$\boldsymbol{\theta}^{(r+1)} = \arg\max_{\boldsymbol{\theta} \in \not\leq} \mathscr{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right).$$

Theorem 3

The MM algorithm generates sequences of iterates $\boldsymbol{\theta}^{(k)}$ that satisfy the inequalities:

$$\mathscr{L}\left(\boldsymbol{\theta}^{(r)}\right) \leq \mathscr{Q}\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)}\right) \leq \mathscr{Q}\left(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}\right) \leq \mathscr{L}\left(\boldsymbol{\theta}^{(r+1)}\right).$$

- The MM algorithm is strictly increasing in the objective sequence $\mathscr{L}\left(\boldsymbol{\theta}^{(r)}\right)$.

Minorization–Maximization Algorithm—3

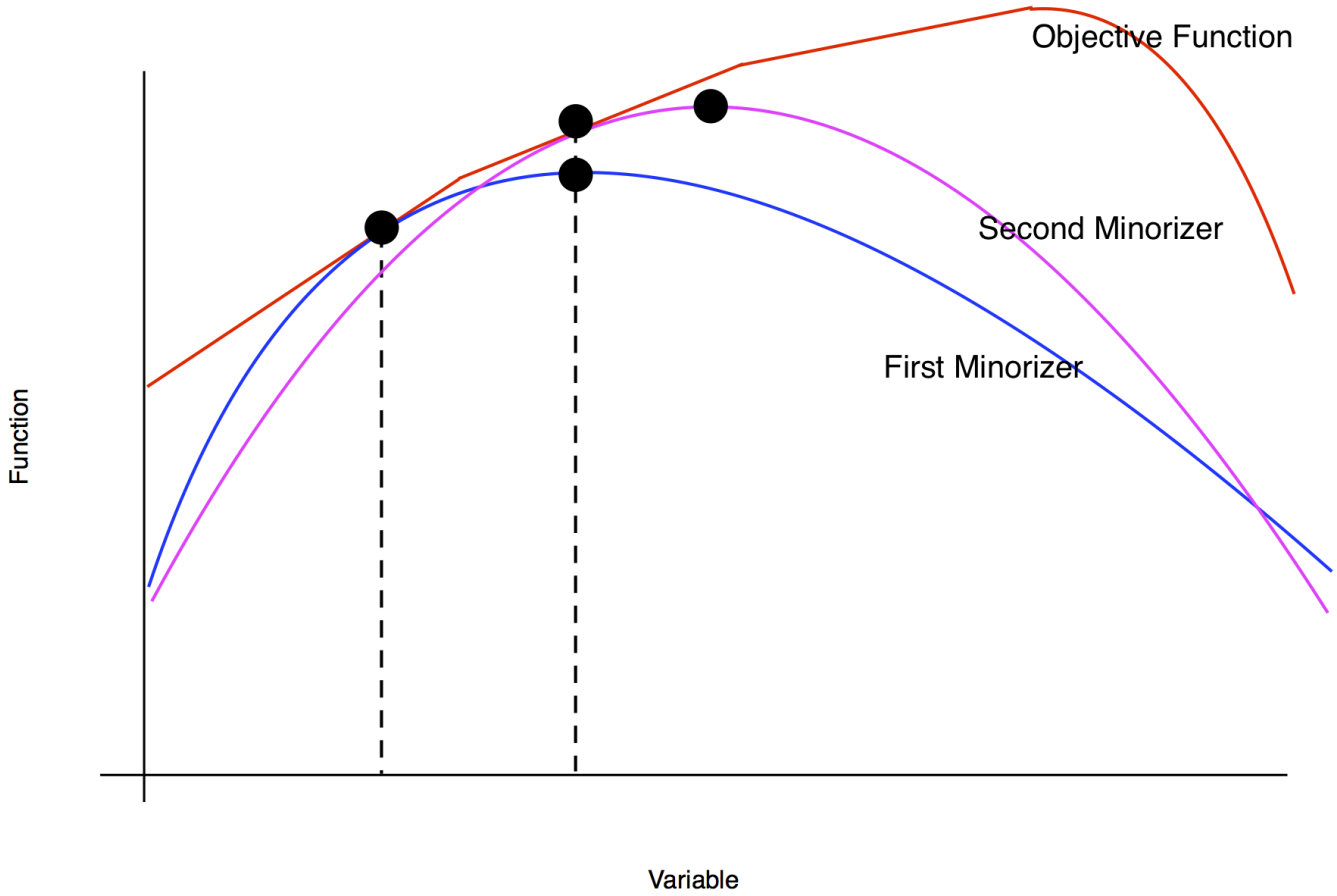

Figure 4: Schematic diagram of an MM algorithm.

A Useful Minorizer

Theorem 4

If $\Theta \subset \mathbb{R}^q$ and $h(\boldsymbol{\theta})$ is a function with Hessian $\mathscr{H}_h(\boldsymbol{\theta})$ such that $\boldsymbol{H} - \mathscr{H}_h(\boldsymbol{\theta})$ is negative semidefinite, for all $\boldsymbol{\theta}$, then $\mathscr{L}(\boldsymbol{\theta}) = h(\boldsymbol{\theta})$ can be minorized by

$$\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\psi}) = h(\boldsymbol{\psi}) + (\boldsymbol{\theta} - \boldsymbol{\psi})^\top \nabla h(\boldsymbol{\psi}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\psi})^\top \boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\psi}).$$

- This is known as the **Quadratic lower-bound minorizer** (cf. [Bohning and Lindsay, 1988, Bohning, 1992]).

A Useful Minorizer

Theorem 4

If $\Theta \subset \mathbb{R}^q$ and $h(\boldsymbol{\theta})$ is a function with Hessian $\mathscr{H}_h(\boldsymbol{\theta})$ such that $\boldsymbol{H} - \mathscr{H}_h(\boldsymbol{\theta})$ is negative semidefinite, for all $\boldsymbol{\theta}$, then $\mathscr{L}(\boldsymbol{\theta}) = h(\boldsymbol{\theta})$ can be minorized by

$$\mathscr{Q}(\boldsymbol{\theta}; \boldsymbol{\psi}) = h(\boldsymbol{\psi}) + (\boldsymbol{\theta} - \boldsymbol{\psi})^\top \nabla h(\boldsymbol{\psi}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\psi})^\top \boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\psi}).$$

- This is known as the **Quadratic lower-bound minorizer** (cf. [Bohning and Lindsay, 1988, Bohning, 1992]).

MM Algorithm for MPL Estimation

- Consider that in any set of parameters $\boldsymbol{\theta}_i$ for $i = 1, .., g-1$, we have the partial gradient and Hessian,

$$\nabla_i \mathscr{P}(\boldsymbol{\theta}) = \sum_{s=1}^{n} \left[\mathbb{I}\{c_j = i\} - \frac{\exp\left(\boldsymbol{\theta}_i^\top \boldsymbol{\eta}_j\right)}{\sum_{k=1}^{g} \exp\left(\boldsymbol{\theta}_k^\top \boldsymbol{\eta}_j\right)} \right] \boldsymbol{\eta}_j$$

and

$$\mathscr{H}_i \mathscr{P}(\boldsymbol{\theta}) = -\sum_{s=1}^{n} \boldsymbol{\eta}_j \boldsymbol{\eta}_j^\top p_{ij}(1 - p_{ij}),$$

respectively, where

$$p_{ij} = \frac{\exp\left(\boldsymbol{\theta}_i^\top \boldsymbol{\eta}_j\right)}{\sum_{k=1}^{g} \exp\left(\boldsymbol{\theta}_k^\top \boldsymbol{\eta}_j\right)}.$$

MM Algorithm for MPL Estimation—2

- Consider that for any i and j,

$$p_{ij}(1-p_{ij}) \leq \frac{1}{4},$$

since $p_{ij} = 1/2$ maximizes the left-hand side.

- Thus, we have $\boldsymbol{H} - \mathscr{H}_i \mathscr{P}(\boldsymbol{\theta})$ is negative definite, if we take

$$\boldsymbol{H} = \frac{1}{4} \sum_{s=1}^{n} \boldsymbol{\eta}_j \boldsymbol{\eta}_j^\top.$$

MM Algorithm for MPL Estimation—2

- Consider that for any i and j,

$$p_{ij}(1-p_{ij}) \leq \frac{1}{4},$$

since $p_{ij} = 1/2$ maximizes the left-hand side.

- Thus, we have $\boldsymbol{H} - \mathscr{H}_l \mathscr{P}(\boldsymbol{\theta})$ is negative definite, if we take

$$\boldsymbol{H} = \frac{1}{4} \sum_{s=1}^{n} \boldsymbol{\eta}_j \boldsymbol{\eta}_j^{\top}.$$

MM Algorithm for MPL Estimation—3

- We can minorize $\mathscr{P}(\boldsymbol{\theta})$ with respect to the element$\boldsymbol{\theta}_i$ at $\boldsymbol{\psi}$, for each $i = 1, \ldots, g-1$ by

$$\begin{aligned}\mathscr{Q}_i(\boldsymbol{\theta}_i; \boldsymbol{\psi}) &= \mathscr{P}(\boldsymbol{\psi}) + (\boldsymbol{\theta}_i - \boldsymbol{\psi}_i)^\top \nabla_i \mathscr{P}(\boldsymbol{\psi}) \\ &\quad -\frac{1}{8}(\boldsymbol{\theta}_i - \boldsymbol{\psi}_i)^\top \boldsymbol{H}(\boldsymbol{\theta}_i - \boldsymbol{\psi}_i),\end{aligned}$$

- The function $\mathscr{Q}_i(\boldsymbol{\theta}_i; \boldsymbol{\psi})$ is quadratic, and thus is concave in $\boldsymbol{\theta}_i$.

- Solving the first order condition $\nabla \mathscr{Q}_i(\boldsymbol{\theta}_i; \boldsymbol{\psi}) = \mathbf{0}$ yields the solution

$$\boldsymbol{\theta}_i^* = 4\boldsymbol{H}^{-1}\nabla_i \mathscr{P}(\boldsymbol{\psi}) + \boldsymbol{\psi}_i.$$

MM Algorithm for MPL Estimation—3

- We can minorize $\mathscr{P}(\boldsymbol{\theta})$ with respect to the element$\boldsymbol{\theta}_i$ at $\boldsymbol{\psi}$, for each $i = 1, \ldots, g-1$ by

$$\begin{aligned} \mathscr{Q}_i(\boldsymbol{\theta}_i; \boldsymbol{\psi}) &= \mathscr{P}(\boldsymbol{\psi}) + (\boldsymbol{\theta}_i - \boldsymbol{\psi}_i)^\top \nabla_i \mathscr{P}(\boldsymbol{\psi}) \\ &\quad -\frac{1}{8}(\boldsymbol{\theta}_i - \boldsymbol{\psi}_i)^\top \boldsymbol{H}(\boldsymbol{\theta}_i - \boldsymbol{\psi}_i), \end{aligned}$$

- The function $\mathscr{Q}_i(\boldsymbol{\theta}_i; \boldsymbol{\psi})$ is quadratic, and thus is concave in $\boldsymbol{\theta}_i$.

- Solving the first order condition $\nabla \mathscr{Q}_i(\boldsymbol{\theta}_i; \boldsymbol{\psi}) = \mathbf{0}$ yields the solution

$$\boldsymbol{\theta}_i^* = 4\boldsymbol{H}^{-1}\nabla_i \mathscr{P}(\boldsymbol{\psi}) + \boldsymbol{\psi}_i.$$

MM Algorithm for MPL Estimation—3

- We can minorize $\mathscr{P}(\boldsymbol{\theta})$ with respect to the element$\boldsymbol{\theta}_i$ at $\boldsymbol{\psi}$, for each $i = 1, \ldots, g-1$ by

$$\begin{aligned} \mathscr{Q}_i(\boldsymbol{\theta}_i; \boldsymbol{\psi}) &= \mathscr{P}(\boldsymbol{\psi}) + (\boldsymbol{\theta}_i - \boldsymbol{\psi}_i)^\top \nabla_i \mathscr{P}(\boldsymbol{\psi}) \\ &\quad -\frac{1}{8}(\boldsymbol{\theta}_i - \boldsymbol{\psi}_i)^\top \boldsymbol{H}(\boldsymbol{\theta}_i - \boldsymbol{\psi}_i), \end{aligned}$$

- The function $\mathscr{Q}_i(\boldsymbol{\theta}_i; \boldsymbol{\psi})$ is quadratic, and thus is concave in $\boldsymbol{\theta}_i$.

- Solving the first order condition $\nabla \mathscr{Q}_i(\boldsymbol{\theta}_i; \boldsymbol{\psi}) = \mathbf{0}$ yields the solution

$$\boldsymbol{\theta}_i^* = 4\boldsymbol{H}^{-1}\nabla_i \mathscr{P}(\boldsymbol{\psi}) + \boldsymbol{\psi}_i.$$

MM Algorithm for MPL Estimation—4

- The minorizer and solution suggests the following MM algorithm update scheme: at the $(r+1)$th iteration of the algorithm, in the order $i = 1, \ldots, g-1$, compute

$$\boldsymbol{\theta}_i^{(r+1)} = 4\boldsymbol{H}^{-1}\nabla_i \mathscr{P}\left(\boldsymbol{\theta}_{[i]}^{(r+1)}\right) + \boldsymbol{\theta}_i^{(r)},$$

where

$$\boldsymbol{\theta}_{[i]}^{(r+1)\top} = \left(\boldsymbol{\theta}_{[i],1}^{(r+1)\top}, \ldots, \boldsymbol{\theta}_{[i],g}^{(r+1)\top}\right)$$

and

$$\boldsymbol{\theta}_{[i],k}^{(r+1)} = \begin{cases} \boldsymbol{\theta}_i^{(r+1)} & \text{if } k < i, \\ \boldsymbol{\theta}_i^{(r)} & \text{if } k \geq i. \end{cases}$$

- Stop when $\mathscr{P}\left(\boldsymbol{\theta}^{(r+1)}\right) - \mathscr{P}\left(\boldsymbol{\theta}^{(r)}\right) < \text{TOL}$ for some $\text{TOL} > 0$.

MM Algorithm for MPL Estimation—4

- The minorizer and solution suggests the following MM algorithm update scheme: at the $(r+1)$th iteration of the algorithm, in the order $i = 1, \ldots, g-1$, compute

$$\boldsymbol{\theta}_i^{(r+1)} = 4\boldsymbol{H}^{-1}\nabla_i \mathscr{P}\left(\boldsymbol{\theta}_{[i]}^{(r+1)}\right) + \boldsymbol{\theta}_i^{(r)},$$

where

$$\boldsymbol{\theta}_{[i]}^{(r+1)\top} = \left(\boldsymbol{\theta}_{[i],1}^{(r+1)\top}, \ldots, \boldsymbol{\theta}_{[i],g}^{(r+1)\top}\right)$$

and

$$\boldsymbol{\theta}_{[i],k}^{(r+1)} = \begin{cases} \boldsymbol{\theta}_i^{(r+1)} & \text{if } k < i, \\ \boldsymbol{\theta}_i^{(r)} & \text{if } k \geq i. \end{cases}$$

- Stop when $\mathscr{P}\left(\boldsymbol{\theta}^{(r+1)}\right) - \mathscr{P}\left(\boldsymbol{\theta}^{(r)}\right) < \text{TOL}$ for some $\text{TOL} > 0$.

Convergence of MM Algorithm

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

Theorem 5

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

Convergence of MM Algorithm

- Upon stopping the algorithm, we declare the final iterate the MPL 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 5

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

Convergence of MM Algorithm

- Upon stopping the algorithm, we declare the final iterate the MPL 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 5

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

Convergence of MM Algorithm—2

- Recall that the pseudolikelihood function $\mathscr{P}$ is concave in $\boldsymbol{\theta}$.

Theorem 6

Every limit point $\boldsymbol{\theta}^{(\infty)}$ of the MM algorithm is a global maximum of $\mathscr{P}(\boldsymbol{\theta})$.

Convergence of MM Algorithm—2

- Recall that the pseudolikelihood function $\mathscr{P}$ is concave in $\boldsymbol{\theta}$.

Theorem 6

Every limit point $\boldsymbol{\theta}^{(\infty)}$ of the MM algorithm is a global maximum of $\mathscr{P}(\boldsymbol{\theta})$.

Asymptotics of MPL Estimator

- Let the set of of maximizers of $\mathscr{P}$ from n observations be

$$\mathscr{M}_n = \left\{ \boldsymbol{\theta} \in \mathbb{R}^q : \mathscr{P}(\boldsymbol{\theta}) = \sup_{\boldsymbol{\psi} \in \mathbb{R}^q} \mathscr{P}(\boldsymbol{\psi}) \right\}.$$

- The following result was presented in [Geman and Graffigne, 1986].

Theorem 7

If the data $C_{z_1}, ..., C_{z_n} = C_1, ..., C_n$ arises from an identifiable MRF with parameter vector $\boldsymbol{\theta}_0$, then the following are true:

(a) $\mathbb{P}(\mathscr{P}(\boldsymbol{\theta})$ *is strictly concave*$) \to 1$, *as* $n \to \infty$.

(b) $\mathbb{P}(\mathscr{M}_n$ *contains a single element*$) \to 1$, *as* $n \to \infty$.

(c) $\mathbb{P}\left(\lim_{n\to\infty} \sup_{\boldsymbol{\theta} \in \mathscr{M}_n} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_0) = 0 \right) = 1.$

Asymptotics of MPL Estimator

- Let the set of of maximizers of $\mathscr{P}$ from n observations be

$$\mathscr{M}_n = \left\{ \boldsymbol{\theta} \in \mathbb{R}^q : \mathscr{P}(\boldsymbol{\theta}) = \sup_{\boldsymbol{\psi} \in \mathbb{R}^q} \mathscr{P}(\boldsymbol{\psi}) \right\}.$$

- The following result was presented in [Geman and Graffigne, 1986].

Theorem 7

If the data $C_{z_1}, ..., C_{z_n} = C_1, ..., C_n$ arises from an identifiable MRF with parameter vector $\boldsymbol{\theta}_0$, then the following are true:

(a) $\mathbb{P}(\mathscr{P}(\boldsymbol{\theta})$ *is strictly concave*$) \to 1$, *as* $n \to \infty$.

(b) $\mathbb{P}(\mathscr{M}_n$ *contains a single element*$) \to 1$, *as* $n \to \infty$.

(c) $\mathbb{P}\left(\lim_{n\to\infty} \sup_{\boldsymbol{\theta} \in \mathscr{M}_n} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_0) = 0 \right) = 1.$

Spatially-Informed Clustering

- Let $\boldsymbol{z}^*$ be a coordinate of interest, and let $C_{\boldsymbol{z}}^*$ be the unknown g-dimensional categorical random variable that is observed at $\boldsymbol{z}^*$.

- Based on the neighborhood function evaluates $\boldsymbol{\eta}_{\boldsymbol{z}}^*$ (that is computed from the neighborhood $\mathbb{C}_{\boldsymbol{z}}^*$ of $\boldsymbol{z}^*$), the **maximum a posteriori** (MAP) estimate of the category $C_{\boldsymbol{z}}^*$ is

$$\hat{c} = \hat{c}\left(\mathbb{C}_{\boldsymbol{z}}^*\right) = \arg \max_{i=1,\ldots g} \mathbb{P}\left(C_{\boldsymbol{z}}^* = i | \mathbb{C}_{\boldsymbol{z}}^*; \boldsymbol{\theta}\right).$$

- The MAP clustering rule can be approximated by substituting $\hat{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$.

Spatially-Informed Clustering

- Let $\boldsymbol{z}^*$ be a coordinate of interest, and let $C^*_{\boldsymbol{z}}$ be the unknown g-dimensional categorical random variable that is observed at $\boldsymbol{z}^*$.

- Based on the neighborhood function evaluates $\boldsymbol{\eta}^*_{\boldsymbol{z}}$ (that is computed from the neighborhood $\mathbb{C}^*_{\boldsymbol{z}}$ of $\boldsymbol{z}^*$), the **maximum a posteriori** (MAP) estimate of the category $C^*_{\boldsymbol{z}}$ is

$$\hat{c} = \hat{c}\left(\mathbb{C}^*_{\boldsymbol{z}}\right) = \arg\max_{i=1,\ldots g} \mathbb{P}\left(C^*_{\boldsymbol{z}} = i | \mathbb{C}^*_{\boldsymbol{z}}; \boldsymbol{\theta}\right).$$

- The MAP clustering rule can be approximated by substituting $\hat{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$.

Spatially-Informed Clustering

- Let $\boldsymbol{z}^*$ be a coordinate of interest, and let $C^*_{\boldsymbol{z}}$ be the unknown g-dimensional categorical random variable that is observed at $\boldsymbol{z}^*$.

- Based on the neighborhood function evaluates $\boldsymbol{\eta}^*_{\boldsymbol{z}}$ (that is computed from the neighborhood $\mathbb{C}^*_{\boldsymbol{z}}$ of $\boldsymbol{z}^*$), the **maximum a posteriori** (MAP) estimate of the category $C^*_{\boldsymbol{z}}$ is

$$\hat{c} = \hat{c}\left(\mathbb{C}^*_{\boldsymbol{z}}\right) = \arg \max_{i=1,\ldots g} \mathbb{P}\left(C^*_{\boldsymbol{z}} = i | \mathbb{C}^*_{\boldsymbol{z}}; \boldsymbol{\theta}\right).$$

- The MAP clustering rule can be approximated by substituting $\hat{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$.

Model Selection

- Suppose that the data $C_{z_1}, ..., C_{z_n} = C_1, ..., C_n$ is generated from a MRF with unknown neighborhood N_z and function η_z (put into $\mathbb{M} = \{\nu_1, .., \nu_M\}$), and unknown parameter vector $\theta_{0,m}$, for $m = 1, ..., M$.

- We can select the best neighborhood and functions combination by $\hat{\nu}(X_1, ..., X_n) = \hat{N}$, where

$$\hat{N} = \arg\min_{m \in M} \; -2 \times \mathscr{P}\left(\hat{\theta}_m\right) + \mathscr{D}_m \log n,$$

  where $\hat{\theta}_m$ is the MPL estimate, under the assumption that g is the number of subpopulations, and $\mathscr{D}_m$ is the size of the vector $\hat{\theta}_m$.

- $\hat{N}$ is the PLIC (**pseudolikelihood information criterion**) rule (cf. [Ji and Seymour, 1996, Stanford and Raftery, 2002]).

Model Selection

- Suppose that the data $C_{z_1}, ..., C_{z_n} = C_1, ..., C_n$ is generated from a MRF with unknown neighborhood N_z and function $\boldsymbol{\eta}_z$ (put into $\mathbb{M} = \{\nu_1, .., \nu_M\}$), and unknown parameter vector $\boldsymbol{\theta}_{0,m}$, for $m = 1, ..., M$.

- We can select the best neighborhood and functions combination by $\hat{\nu}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) = \hat{N}$, where

$$\hat{N} = \arg\min_{m \in M} \; -2 \times \mathscr{P}\left(\hat{\boldsymbol{\theta}}_m\right) + \mathscr{D}_m \log n,$$

  where $\hat{\boldsymbol{\theta}}_m$ is the MPL estimate, under the assumption that g is the number of subpopulations, and $\mathscr{D}_m$ is the size of the vector $\hat{\boldsymbol{\theta}}_m$.

- $\hat{N}$ is the PLIC (**pseudolikelihood information criterion**) rule (cf. [Ji and Seymour, 1996, Stanford and Raftery, 2002]).

Model Selection

- Suppose that the data $C_{z_1}, ..., C_{z_n} = C_1, ..., C_n$ is generated from a MRF with unknown neighborhood N_z and function $\boldsymbol{\eta}_z$ (put into $\mathbb{M} = \{\nu_1, .., \nu_M\}$), and unknown parameter vector $\boldsymbol{\theta}_{0,m}$, for $m = 1, ..., M$.

- We can select the best neighborhood and functions combination by $\hat{\nu}(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) = \hat{N}$, where

$$\hat{N} = \arg\min_{m \in M} \; -2 \times \mathscr{P}\left(\hat{\boldsymbol{\theta}}_m\right) + \mathscr{D}_m \log n,$$

  where $\hat{\boldsymbol{\theta}}_m$ is the MPL estimate, under the assumption that g is the number of subpopulations, and $\mathscr{D}_m$ is the size of the vector $\hat{\boldsymbol{\theta}}_m$.

- $\hat{N}$ is the PLIC (**pseudolikelihood information criterion**) rule (cf. [Ji and Seymour, 1996, Stanford and Raftery, 2002]).

Summary of MRF Clustering Algorithm

1. Determine a set of plausible neighborhoods $N_{\boldsymbol{z}}$ and functions $\boldsymbol{\eta}_{\boldsymbol{z}}$, and put these into $\mathbb{M}$.
2. Use the MM algorithm to estimate the MPL estimator $\hat{\boldsymbol{\theta}}_m$, for each $m \in \mathbb{M}$.
3. Estimate the best neighborhood and function using the PLIC rule, $\hat{N}$.
4. Using the MRF defined by the selected neighborhood and function, estimate the cluster allocation of each observation (or observations used for estimation), $C_{\boldsymbol{z}}^*$ observed at $\boldsymbol{z}^*$ using the MAP clustering rule, $\hat{c}\,(\mathbb{C}_{\boldsymbol{z}}^*)$.

Simulation Example

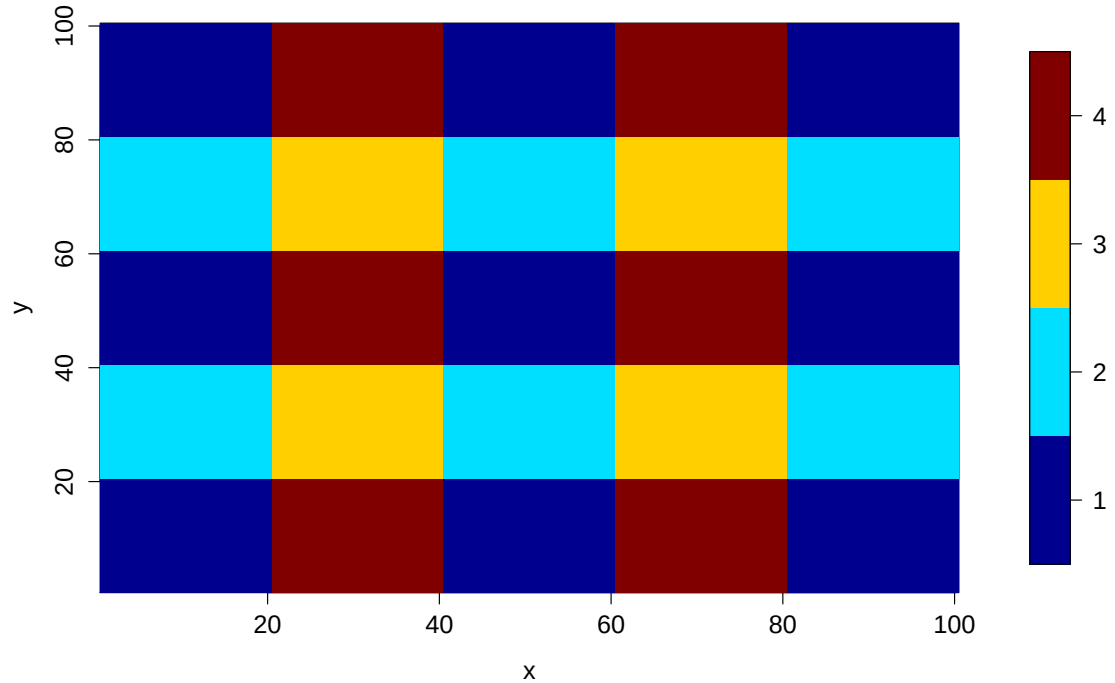

Figure 5: True categories $C_{\boldsymbol{z}} \in \{1, 2, 3, 4\}$ of simulation paradigm on a 100×100 $(x \times y)$ square lattice.

Simulation Example—2

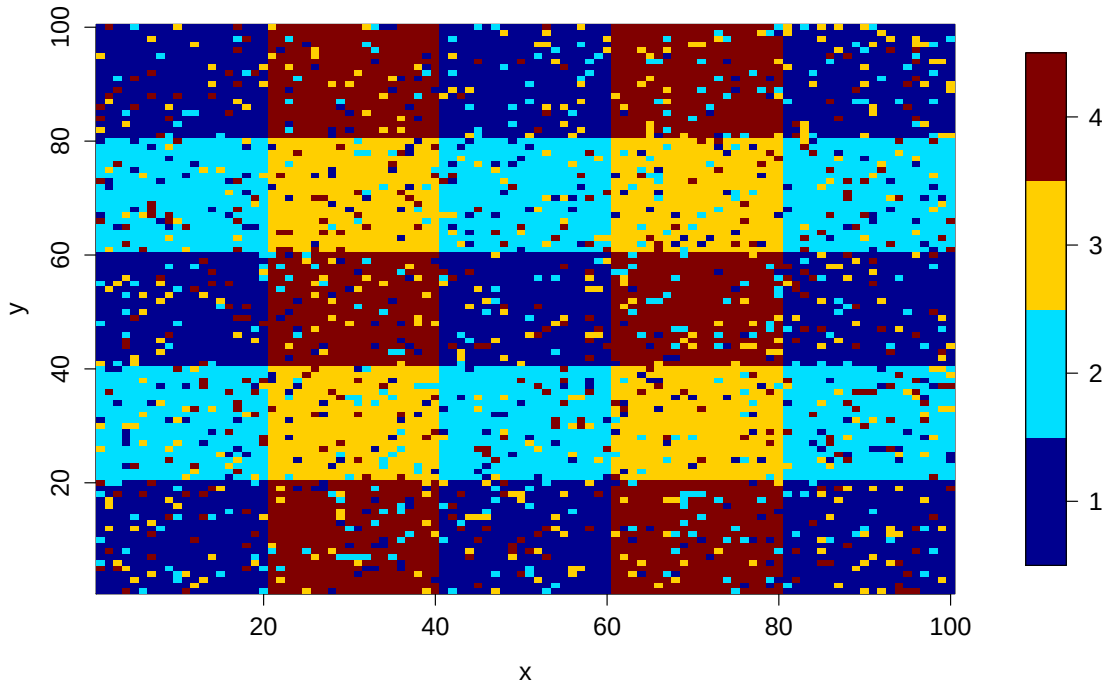

Figure 6: Simulated data that is corrupted, with 20% of the points being randomly mutated.

Simulation Example—3

- We utilize Moore neighborhoods of the form:

$$N_{\boldsymbol{z}} = \{(i,j) : \max\{|x-i|, |y-j|\} \leq d, (i,j) \neq (x,y)\},$$

for $d = 1, 2, 3, 4, ...$

  - Recall that $\mathbb{C}_{\boldsymbol{z}} = \{C_{\boldsymbol{\zeta}} : \boldsymbol{\zeta} \in N_{\boldsymbol{z}}\}$.

- We use functions over g-dimensional categorical random variable $C_{\boldsymbol{z}} \in \{1, 2, 3, 4\}$:$\boldsymbol{\eta}_{\boldsymbol{z}}^{\top} = (\eta_{1,\boldsymbol{z}}, ..., \eta_{g,\boldsymbol{z}})$, where

$$\eta_{i,\boldsymbol{z}} = \frac{1}{|\mathbb{C}_{\boldsymbol{z}}|} \sum_{\boldsymbol{\zeta} \in \mathbb{C}_{\boldsymbol{z}}} \mathbb{I}\{C_{\boldsymbol{\zeta}} = i\}$$

for $i = 1, ..., g$.

Simulation Example—3

- We utilize Moore neighborhoods of the form:

$$N_{\boldsymbol{z}} = \{(i,j) : \max\{|x-i|, |y-j|\} \leq d, (i,j) \neq (x,y)\},$$

for $d = 1, 2, 3, 4, ...$

  - Recall that $\mathbb{C}_{\boldsymbol{z}} = \{C_{\boldsymbol{\zeta}} : \boldsymbol{\zeta} \in N_{\boldsymbol{z}}\}$.

- We use functions over g-dimensional categorical random variable $C_{\boldsymbol{z}} \in \{1, 2, 3, 4\}$:$\boldsymbol{\eta}_{\boldsymbol{z}}^{\top} = (\eta_{1,\boldsymbol{z}}, ..., \eta_{g,\boldsymbol{z}})$, where

$$\eta_{i,\boldsymbol{z}} = \frac{1}{|\mathbb{C}_{\boldsymbol{z}}|} \sum_{\boldsymbol{\zeta} \in \mathbb{C}_{\boldsymbol{z}}} \mathbb{I}\{C_{\boldsymbol{\zeta}} = i\}$$

for $i = 1, ..., g$.

Simulation Example—4

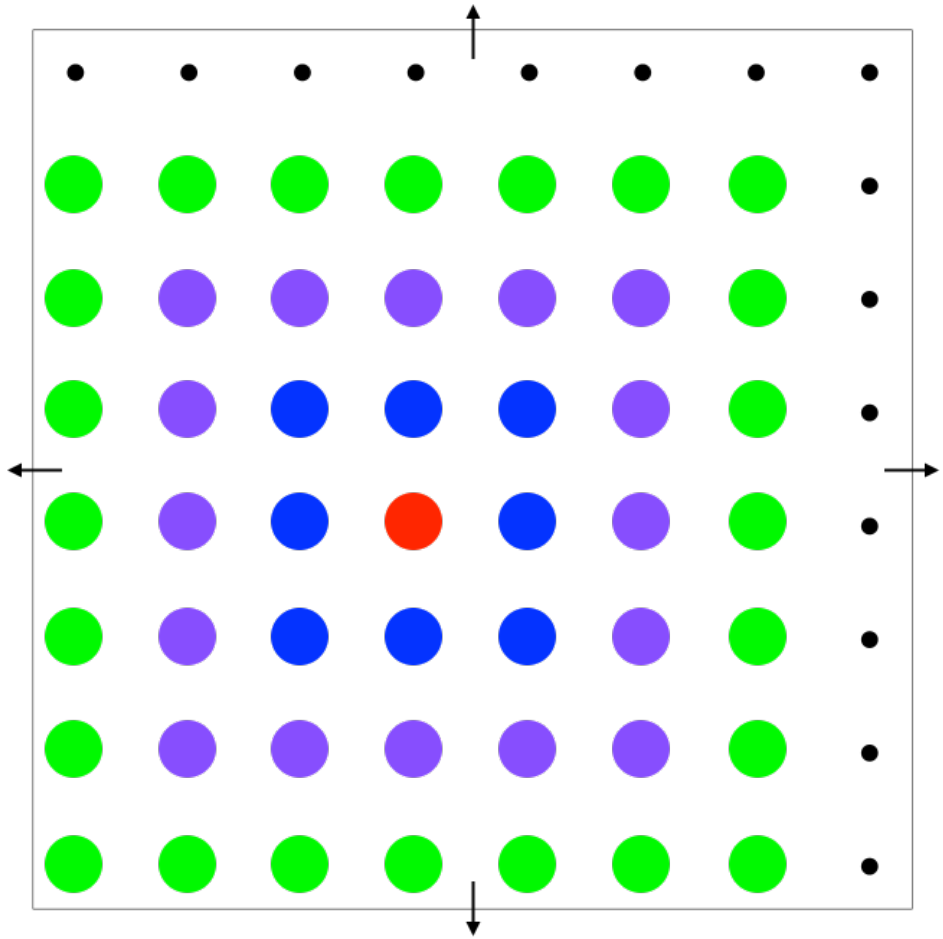

Figure 7: Schematic of Moore neighborhoods of sizes $d = 1, 2, 3$ (blue, purple, green).

Simulation Example—5

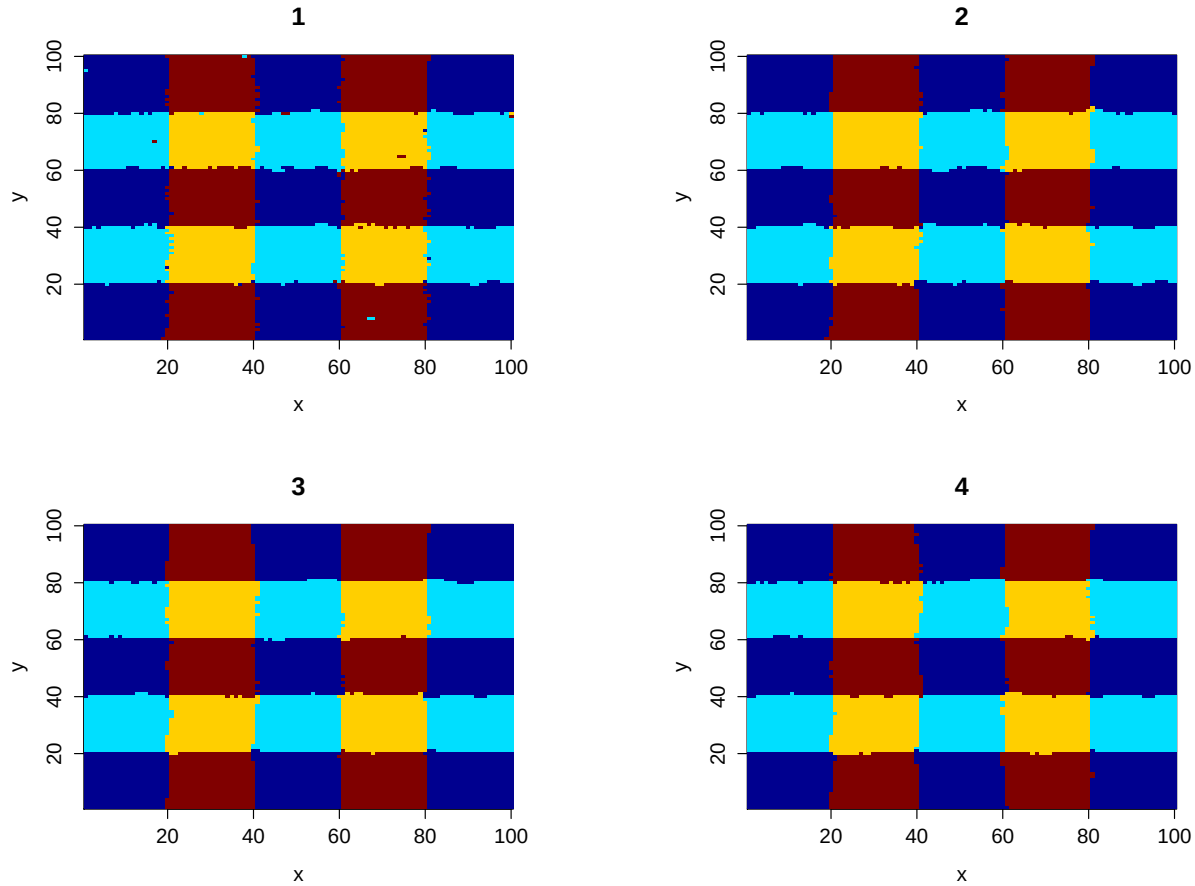

Figure 8: Clustering via MAP rule using MRF defined by Moore neighborhoods of sizes $d = 1, 2, 3, 4$, and average numbers of observations in each class.

Simulation Example—6

Table 1: PLIC values for each of the fitted MRFs, as well as adjusted Rand indices (cf. [Hubert and Arabie, 1985]) and accuracies for each of the clusterings via the MAP rule. Boldface indicates selection via PLIC rule.

d	PLIC value	ARI	Accuracy
1	12487	0.94	0.98
2	**12480**	**0.96**	**0.98**
3	12752	0.96	0.99
4	13099	0.96	0.98

More Simulation Examples—1

- We use functions over g-dimensional categorical random variable $C_z \in \{1,2,3,4\}$:$\boldsymbol{\eta}_z^\top = \left(\eta_{1,z}, ..., \eta_{g,z}, \boldsymbol{\omega}_z^\top\right)$, where

$$\eta_{i,z} = \frac{1}{|\mathbb{C}_z|} \sum_{\boldsymbol{\zeta} \in \mathbb{C}_z} \mathbb{I}\{C_{\boldsymbol{\zeta}} = i\}$$

for $i = 1, ..., g$, and $\boldsymbol{\omega}_z$ contains the interactions

$$\omega_{ij,z} = \eta_{i,z} \times \eta_{j,z}$$

for $i \neq j$.

More Simulation Examples—1

- We use functions over g-dimensional categorical random variable $C_{\boldsymbol{z}} \in \{1,2,3,4\}$:$\boldsymbol{\eta}_{\boldsymbol{z}}^{\top} = \left(\eta_{1,\boldsymbol{z}}, ..., \eta_{g,\boldsymbol{z}}, \boldsymbol{\omega}_{\boldsymbol{z}}^{\top}\right)$, where

$$\eta_{i,\boldsymbol{z}} = \frac{1}{|\mathbb{C}_{\boldsymbol{z}}|} \sum_{\boldsymbol{\zeta} \in \mathbb{C}_{\boldsymbol{z}}} \mathbb{I}\left\{C_{\boldsymbol{\zeta}} = i\right\}$$

for $i = 1, ..., g$, and $\boldsymbol{\omega}_{\boldsymbol{z}}$ contains the interactions

$$\omega_{ij,\boldsymbol{z}} = \eta_{i,\boldsymbol{z}} \times \eta_{j,\boldsymbol{z}}$$

for $i \neq j$.

More Simulation Examples—2

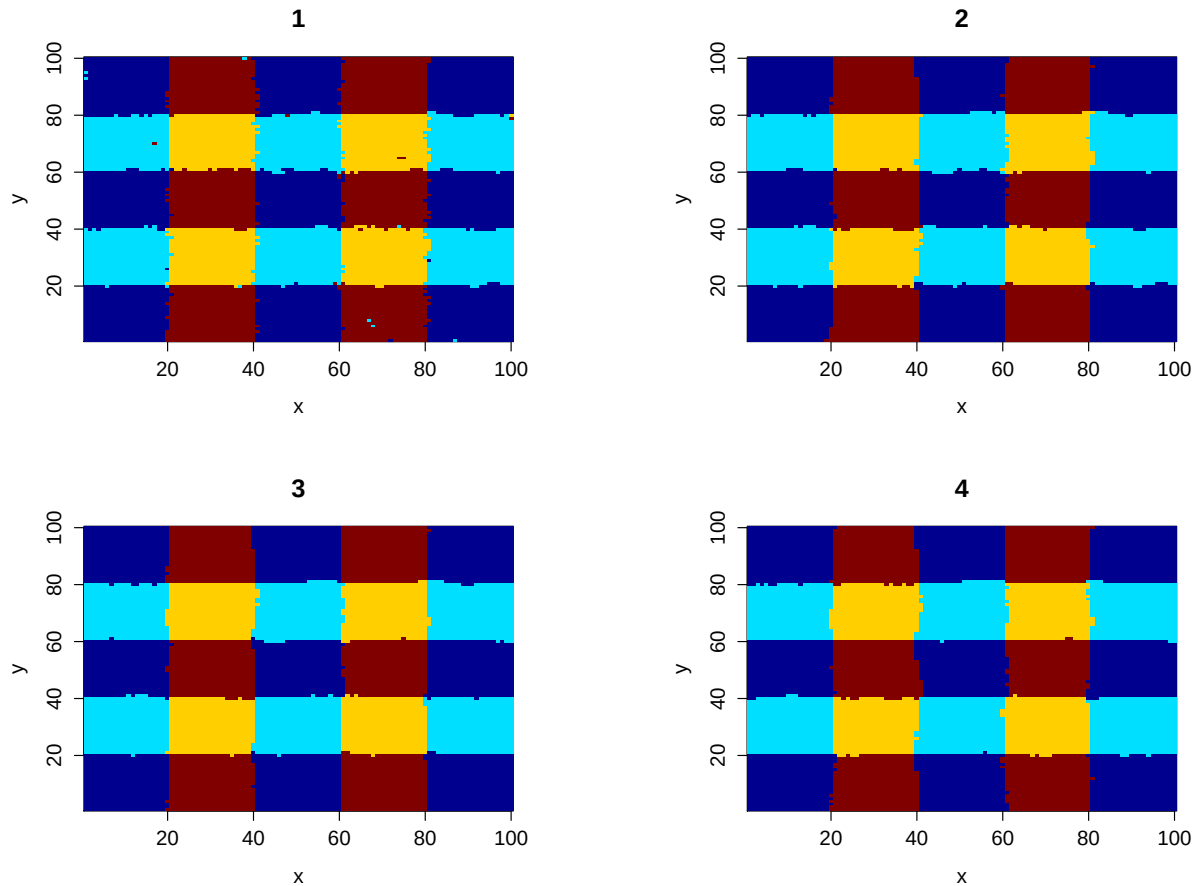

Figure 9: Clustering via MAP rule using MRF defined by Moore neighborhoods of sizes $d = 1, 2, 3, 4$, and average numbers of observations in each class, with interactions.

More Simulation Examples—3

Table 2: PLIC values for each of the fitted MRFs, as well as adjusted Rand indices (cf. [Hubert and Arabie, 1985]) and accuracies for each of the clusterings via the MAP rule. Boldface indicates selection via PLIC rule.

d	PLIC value	ARI	Accuracy
1	12513	0.94	0.98
2	**12482**	**0.96**	**0.98**
3	12737	0.97	0.99
4	13054	0.96	0.98

Von Neumann Simulation—1

- We now utilize von Neumann neighborhoods of the form:

$$N_{\boldsymbol{z}} = \{(i,j) : |x-i| + |y-j| \leq d, (i,j) \neq (x,y)\},$$

for $d = 1, 2, 3, 4, \ldots$.

  - Recall that $\mathbb{C}_{\boldsymbol{z}} = \left\{ C_{\boldsymbol{\zeta}} : \zeta \in N_{\boldsymbol{z}} \right\}$.

Von Neumann Simulation—2

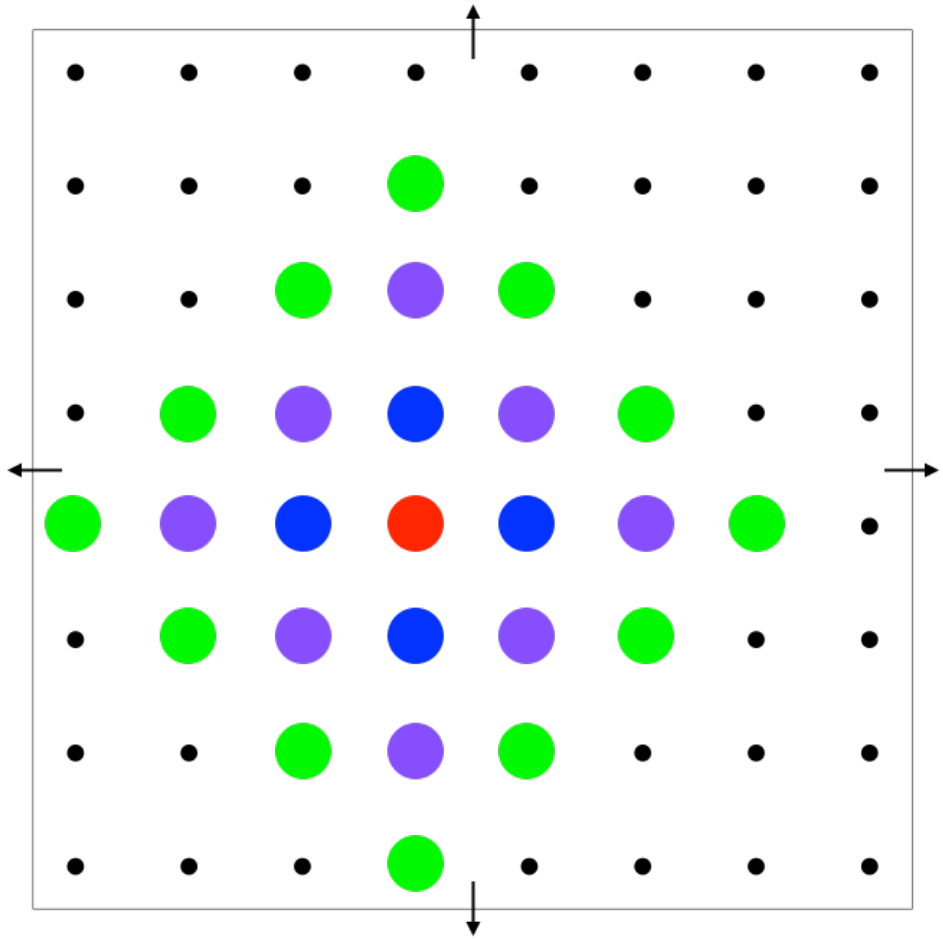

Figure 10: Schematic of von Neumann neighborhoods of sizes $d = 1, 2, 3$ (blue, purple, green).

Von Neumann Simulation—3

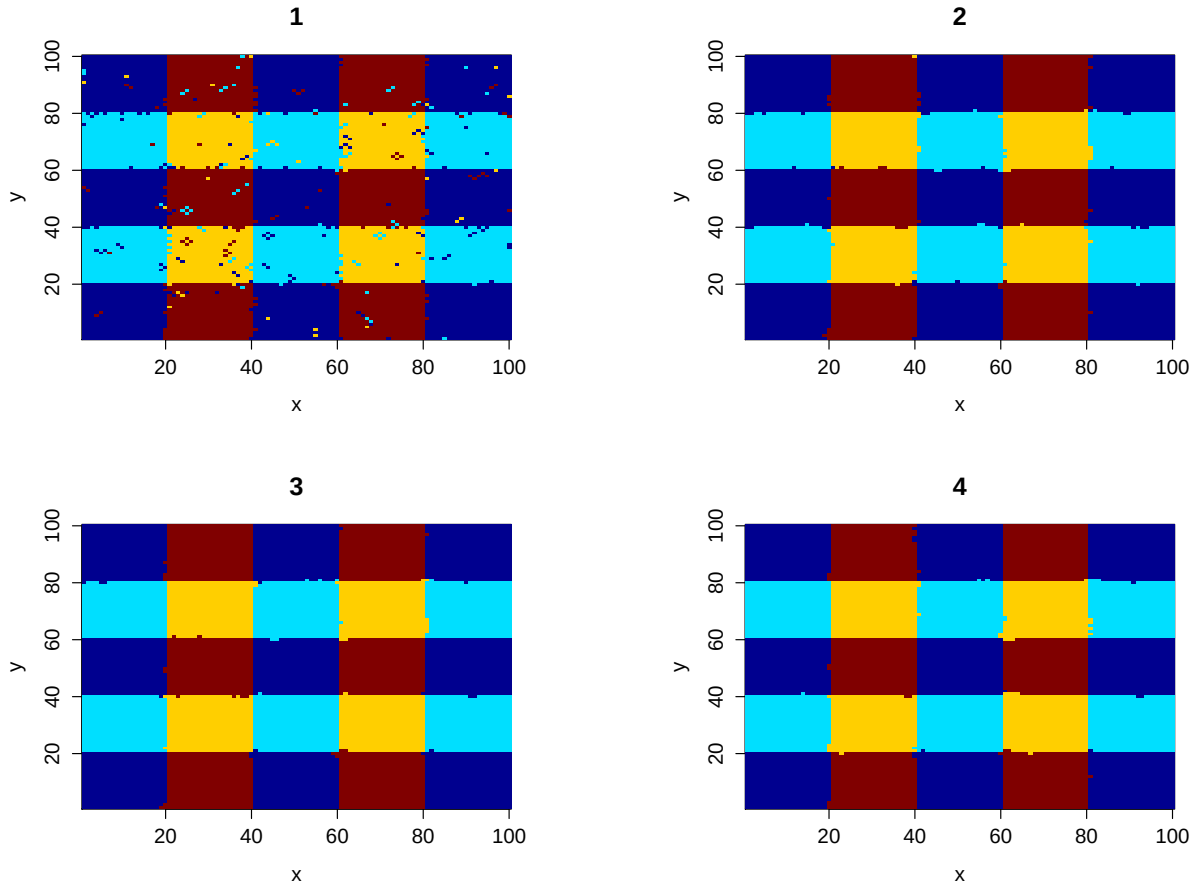

Figure 11: Clustering via MAP rule using MRF defined by von Neumann neighborhoods of sizes $d = 1, 2, 3, 4$, and average numbers of observations in each class.

Von Neumann Simulation—4

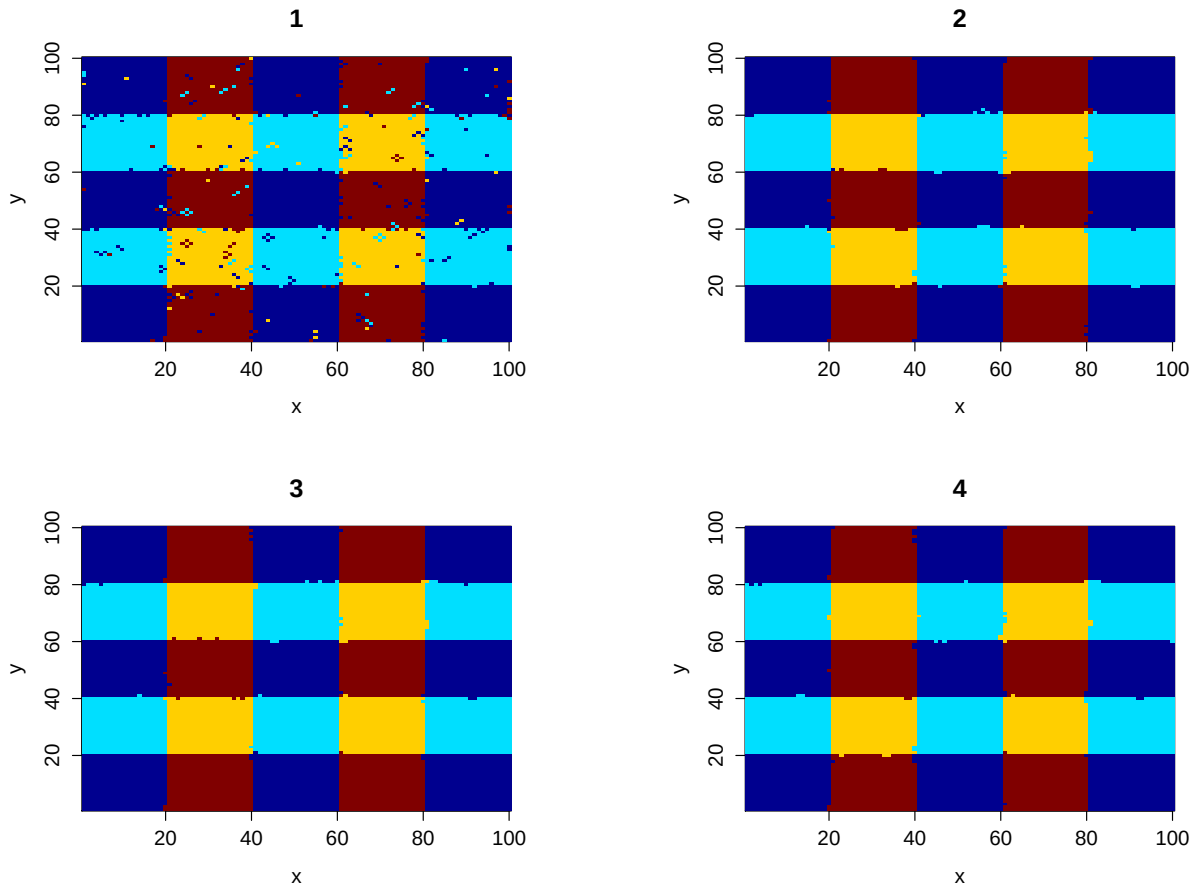

Figure 12: Clustering via MAP rule using MRF defined by von Neumann neighborhoods of sizes $d = 0, 1, 2, 3$, and average numbers of observations in each class, with interactions.

Von Neumann Simulation—5

Table 3: PLIC values for each of the fitted MRFs, as well as adjusted Rand indices (cf. [Hubert and Arabie, 1985]) and accuracies for each of the clusterings via the MAP rule. Boldface indicates selection via PLIC rule.

No Interactions				Interactions			
d	PLIC value	ARI	Accuracy	d	PLIC value	ARI	Accuracy
1	12914	0.92	0.97	1	12917	0.92	0.97
2	**12148**	**0.98**	**0.99**	2	12169	0.98	0.99
3	12205	0.98	0.99	3	12213	0.99	0.99
4	12355	0.99	0.99	4	12364	0.98	0.99

DEMP-Merged GMM Clustering with 5 Clusters—Again

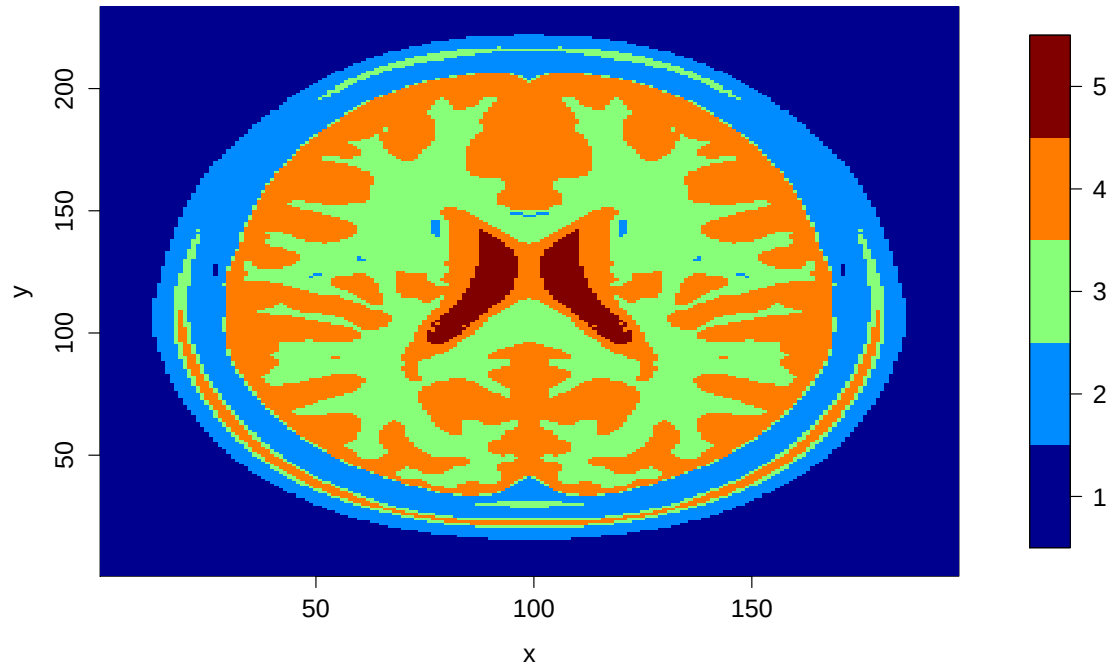

Figure 13: Location of clusters after automatic cluster merging of a $g = 11$ component GMM clustering to 5 clusters, via the DEMP method of [Hennig, 2000].

Brain Clustering

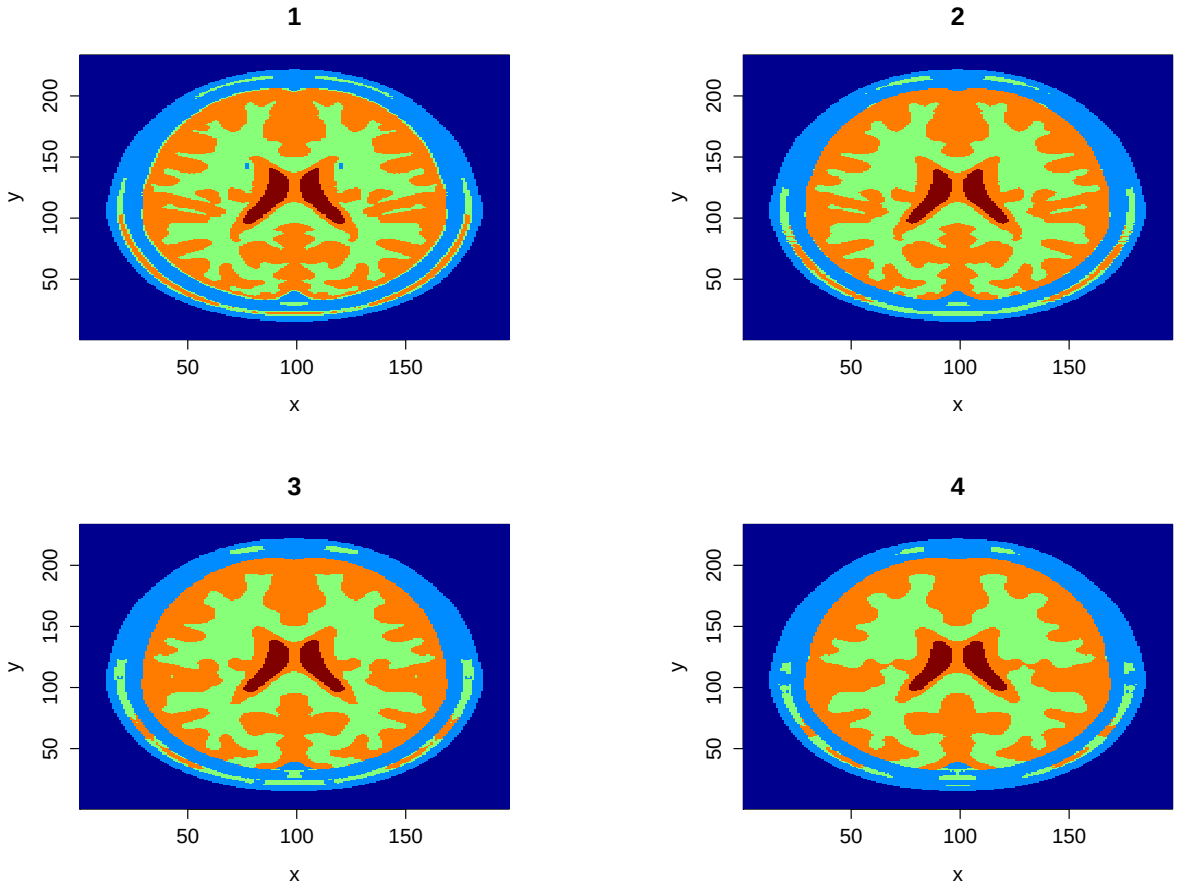

Figure 14: Clustering via MAP rule using MRF defined by Moore neighborhoods of sizes $d = 1, 2, 3, 4$, and average numbers of observations in each class.

Brain Clustering—2

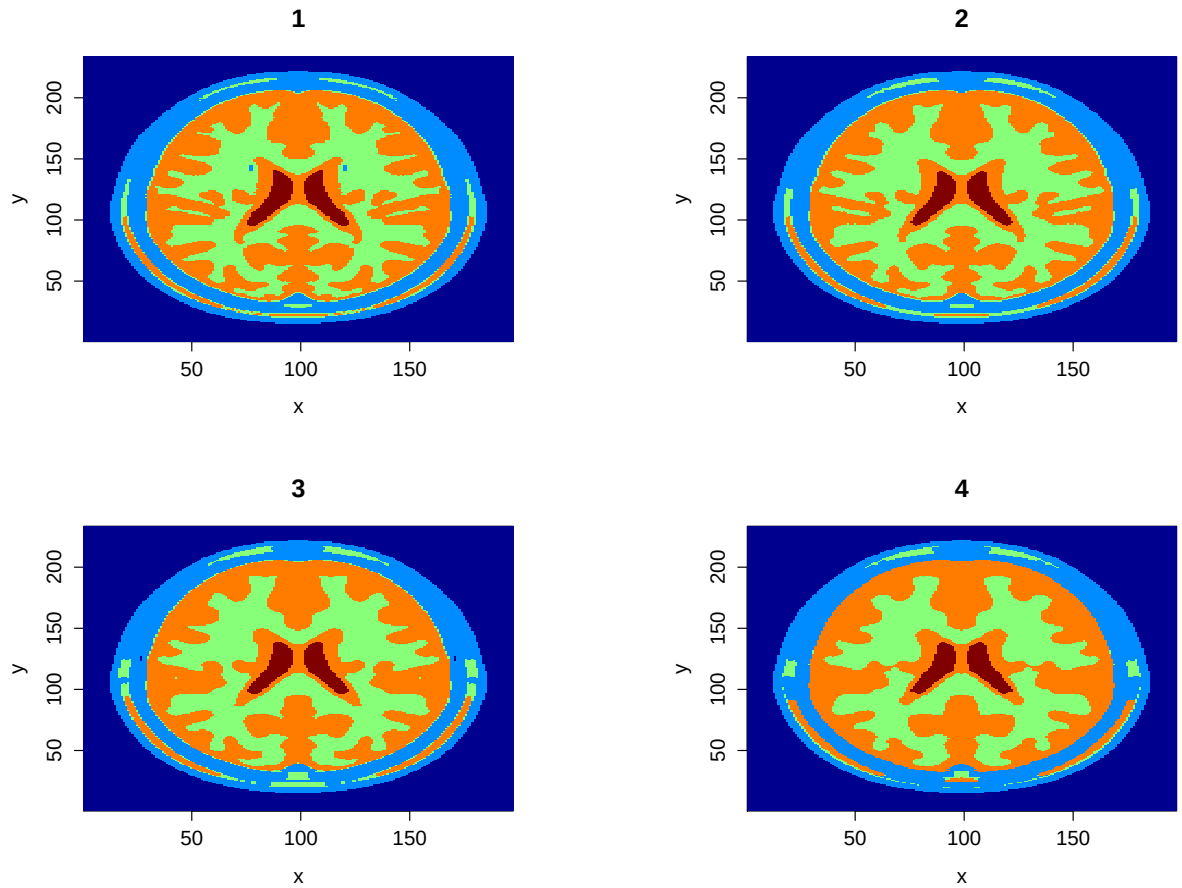

Figure 15: Clustering via MAP rule using MRF defined by Moore neighborhoods of sizes $d = 1, 2, 3, 4$, and average numbers of observations in each class, with interactions.

Brain Clustering—3

Table 4: PLIC values for each of the fitted MRFs, as well as adjusted Rand indices (cf. [Hubert and Arabie, 1985]) and agreement with initial clustering, for each of the clusterings via the MAP rule. Boldface indicates selection via PLIC rule.

No Interactions				Interactions			
d	PLIC value	ARI	Agree	d	PLIC value	ARI	Agree
1	6139	0.94	0.97	**1**	**6121**	**0.94**	**0.97**
2	9785	0.92	0.96	2	8431	0.93	0.96
3	13879	0.89	0.94	3	11834	0.90	0.95
4	17826	0.86	0.92	4	15672	0.87	0.93

Brain Clustering—4

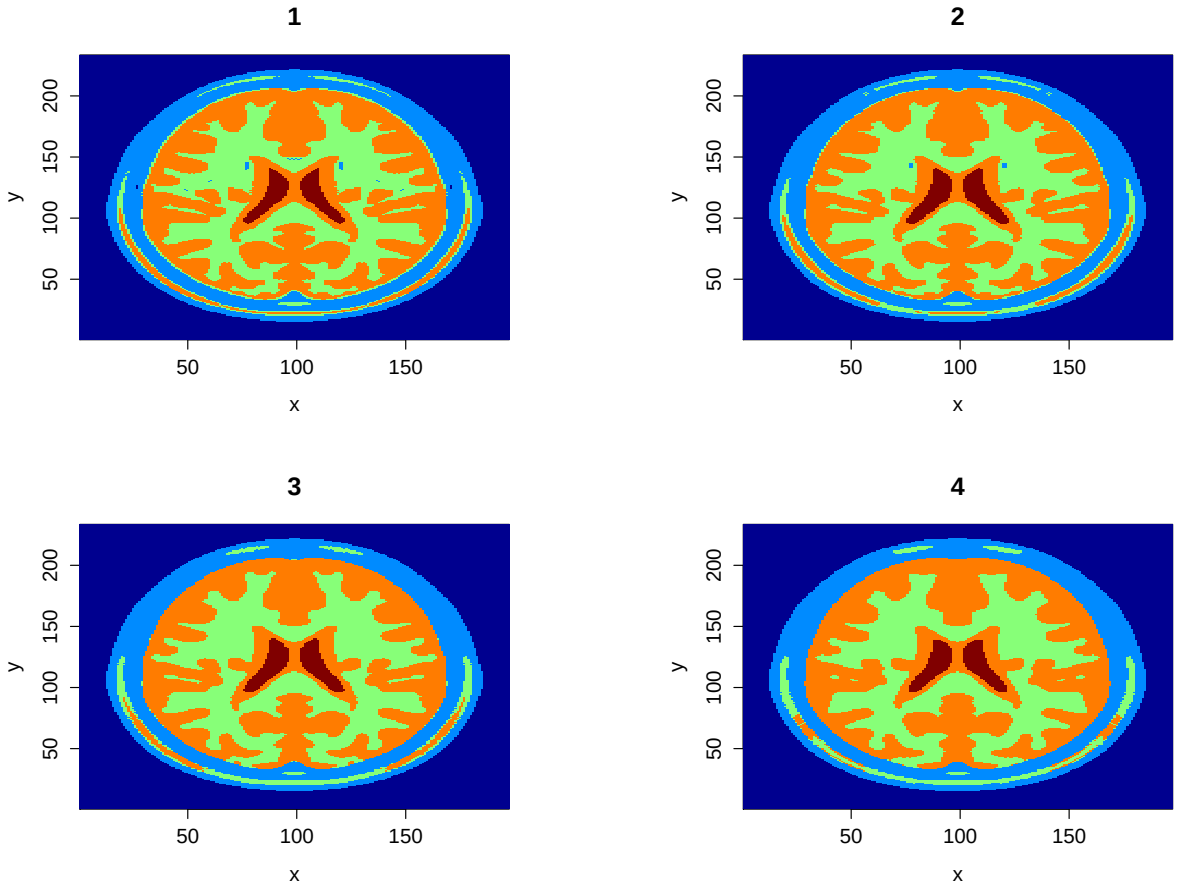

Figure 16: Clustering via MAP rule using MRF defined by von Neumann neighborhoods of sizes $d = 1, 2, 3, 4$, and average numbers of observations in each class.

Brain Clustering—5

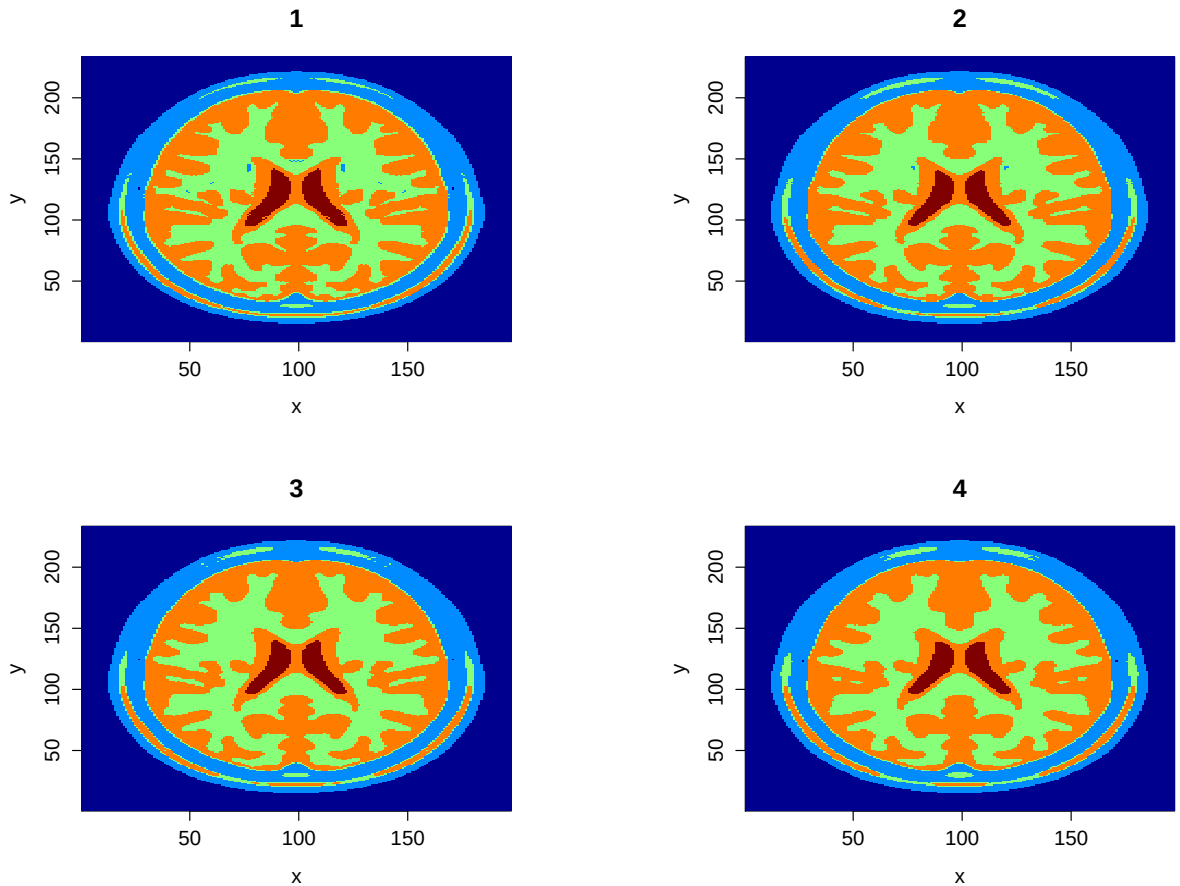

Figure 17: Clustering via MAP rule using MRF defined by von Neumann neighborhoods of sizes $d = 12, 3, 4$, and average numbers of observations in each class, with interactions.

Brain Clustering—6

Table 5: PLIC values for each of the fitted MRFs, as well as adjusted Rand indices (cf. [Hubert and Arabie, 1985]) and agreement with initial clustering, for each of the clusterings via the MAP rule. Boldface indicates selection via PLIC rule.

No Interactions				Interactions			
d	PLIC value	ARI	Agree	d	PLIC value	ARI	Agree
1	**6292**	**0.93**	**0.96**	1	6456	0.93	0.96
2	7023	0.94	0.97	2	6771	0.94	0.97
3	9646	0.92	0.96	3	8233	0.93	0.96
4	12515	0.90	0.95	4	10317	0.91	0.96

Corrupted Brain Image

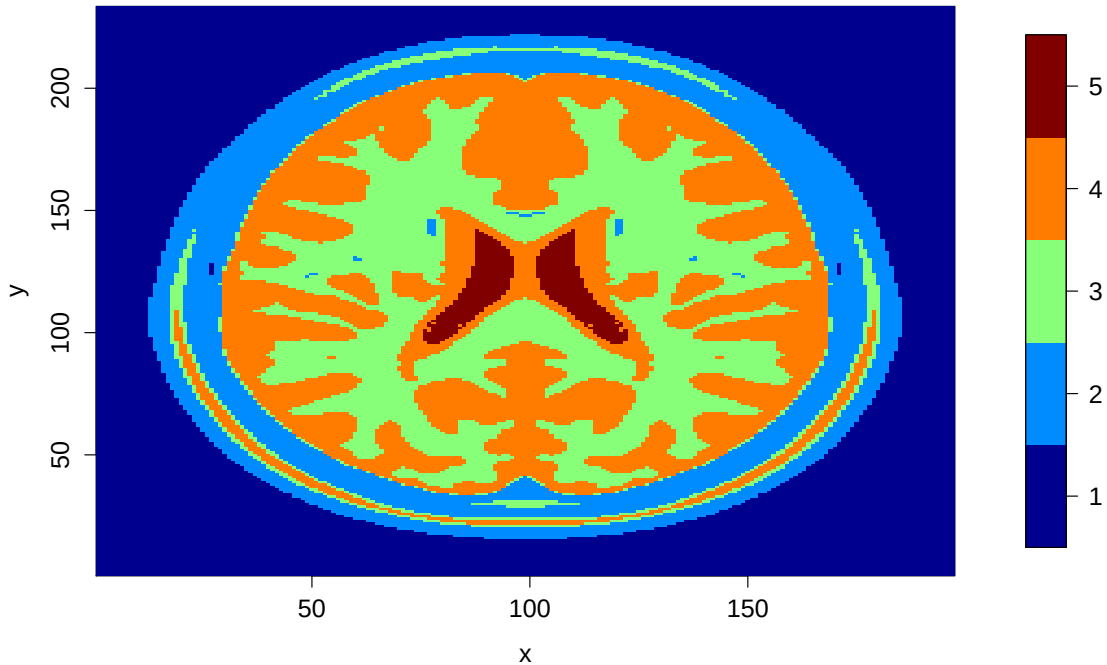

Figure 18: A corruption of the DEMP-merged image with $> 20\%$ of the pixels mutated at random.

Corrupted Brain Image—2

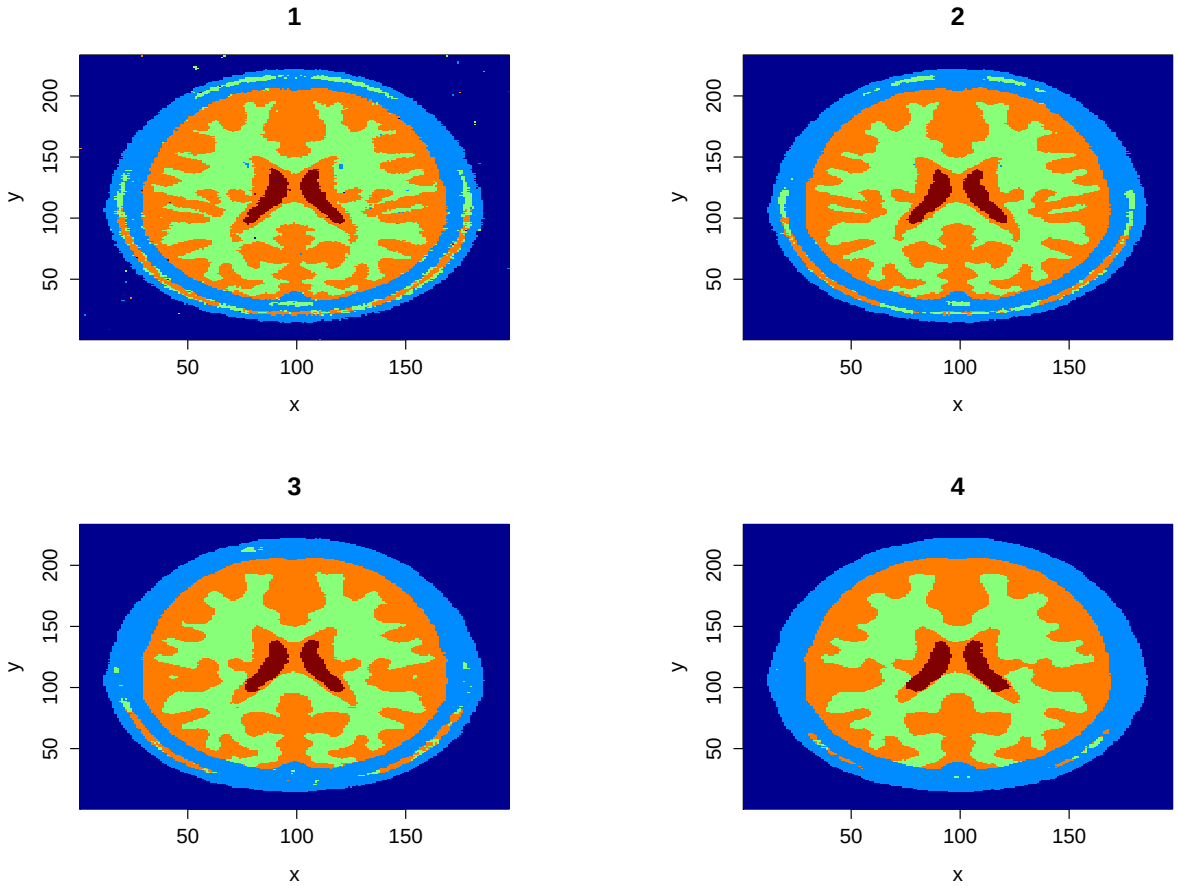

Figure 19: Clustering via MAP rule using MRF defined by Moore neighborhoods of sizes $d = 1, 2, 3, 4$, and average numbers of observations in each class.

Corrupted Brain Image—3

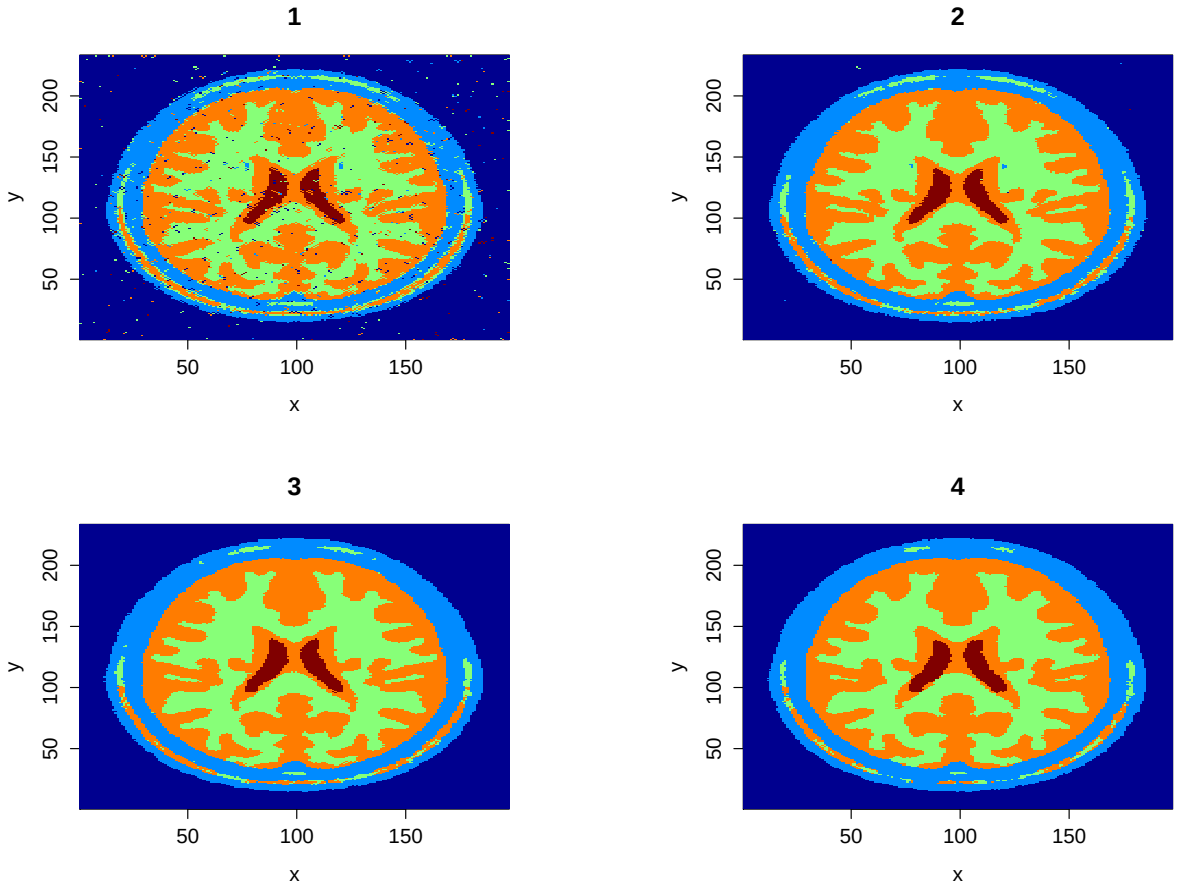

Figure 20: Clustering via MAP rule using MRF defined by von Neumann neighborhoods of sizes $d = 1, 2, 3, 4$, and average numbers of observations in each class.

Corrupted Brain Image—4

Table 6: PLIC values for each of the fitted MRFs, as well as adjusted Rand indices (cf. [Hubert and Arabie, 1985]) and agreement with initial (uncorrupted) clustering, for each of the clusterings via the MAP rule. Boldface indicates selection via PLIC rule.

Moore				von Neumann			
d	PLIC value	ARI	Agree	d	PLIC value	ARI	Agree
1	69341	0.90	0.95	1	72739	0.85	0.93
2	69436	0.90	0.94	**2**	**68869**	**0.91**	**0.95**
3	71152	0.87	0.93	3	69386	0.90	0.95
4	73644	0.84	0.91	4	70550	0.88	0.94

Extensions

- Note that we have defined the neighborhood N_z to not include the point z, in order to be able to predict C_z without observing a category at the location. If we instead observe an unreliable signal at the location and wish to estimate the true underlying signal, with the unreliable signal as information, then we can use neighborhoods that include z.
  - Moore smoothing neighborhood: $N_z = \{(i,j) : \max\{|x-i|, |y-j|\} \leq d\}$.
- Model-based clustering and MRF smoothing can be iterated in order to refine clustering via the ICM (iterative conditional modes) method (cf. [Besag, 1986, McLachlan, 1992, Ch. 13]).
- MRF-based clustering and smoothing can be extended to three-dimensional spatial data (e.g. [Nguyen et al., 2014]).

Extensions

- Note that we have defined the neighborhood $N_{\boldsymbol{z}}$ to not include the point $\boldsymbol{z}$, in order to be able to predict $C_{\boldsymbol{z}}$ without observing a category at the location. If we instead observe an unreliable signal at the location and wish to estimate the true underlying signal, with the unreliable signal as information, then we can use neighborhoods that include $\boldsymbol{z}$.
  - Moore smoothing neighborhood: $N_{\boldsymbol{z}} = \{(i,j) : \max\{|x-i|, |y-j|\} \leq d\}$.
- Model-based clustering and MRF smoothing can be iterated in order to refine clustering via the ICM (iterative conditional modes) method (cf. [Besag, 1986, McLachlan, 1992, Ch. 13]).
- MRF-based clustering and smoothing can be extended to three-dimensional spatial data (e.g. [Nguyen et al., 2014]).

Extensions

- Note that we have defined the neighborhood N_z to not include the point z, in order to be able to predict C_z without observing a category at the location. If we instead observe an unreliable signal at the location and wish to estimate the true underlying signal, with the unreliable signal as information, then we can use neighborhoods that include z.
  - Moore smoothing neighborhood: $N_z = \{(i,j) : \max\{|x-i|, |y-j|\} \leq d\}$.
- Model-based clustering and MRF smoothing can be iterated in order to refine clustering via the ICM (iterative conditional modes) method (cf. [Besag, 1986, McLachlan, 1992, Ch. 13]).
- MRF-based clustering and smoothing can be extended to three-dimensional spatial data (e.g. [Nguyen et al., 2014]).

Next Lecture

- Introduce autoregressive-moving average models.
- Discuss covariance stationarity and invertibility.
- Introduce autocorrelation functions and correlograms.
- Consider maximum likelihood estimation for autoregressive models.
- Present asymptotic results regarding the MLE
- Introduce and demonstrate extremum estimation
- Present a method for model selection between different AR models.

Next Lecture

- Introduce autoregressive-moving average models.
- Discuss covariance stationarity and invertibility.
- Introduce autocorrelation functions and correlograms.
- Consider maximum likelihood estimation for autoregressive models.
- Present asymptotic results regarding the MLE
- Introduce and demonstrate extremum estimation
- Present a method for model selection between different AR models.

Next Lecture

- Introduce autoregressive-moving average models.
- Discuss covariance stationarity and invertibility.
- Introduce autocorrelation functions and correlograms.
- Consider maximum likelihood estimation for autoregressive models.
- Present asymptotic results regarding the MLE
- Introduce and demonstrate extremum estimation
- Present a method for model selection between different AR models.

Next Lecture

- Introduce autoregressive-moving average models.
- Discuss covariance stationarity and invertibility.
- Introduce autocorrelation functions and correlograms.
- Consider maximum likelihood estimation for autoregressive models.
- Present asymptotic results regarding the MLE
- Introduce and demonstrate extremum estimation
- Present a method for model selection between different AR models.

Next Lecture

- Introduce autoregressive-moving average models.
- Discuss covariance stationarity and invertibility.
- Introduce autocorrelation functions and correlograms.
- Consider maximum likelihood estimation for autoregressive models.
- Present asymptotic results regarding the MLE
- Introduce and demonstrate extremum estimation
- Present a method for model selection between different AR models.

Next Lecture

- Introduce autoregressive-moving average models.
- Discuss covariance stationarity and invertibility.
- Introduce autocorrelation functions and correlograms.
- Consider maximum likelihood estimation for autoregressive models.
- Present asymptotic results regarding the MLE
- Introduce and demonstrate extremum estimation
- Present a method for model selection between different AR models.

Next Lecture

- Introduce autoregressive-moving average models.
- Discuss covariance stationarity and invertibility.
- Introduce autocorrelation functions and correlograms.
- Consider maximum likelihood estimation for autoregressive models.
- Present asymptotic results regarding the MLE
- Introduce and demonstrate extremum estimation
- Present a method for model selection between different AR models.

Questions?

tinyurl.com/hiendnguyen

References I

Besag, J. (1974).

Spatial interaction and the statistical analysis of lattice systems.
Journal of the Royal Statistical Society Series B, 36:192–236.

Besag, J. (1986).

On the statistical analysis of dirty pictures.
Journal of the Royal Statistical Society Series B, 48:259–302.

Bohning, D. (1992).

Multinomial logistic regression algorithm.
Annals of the Institute of Mathematical Statistics, 44:197–200.

Bohning, D. and Lindsay, B. R. (1988).

Monotonicity of quadratic-approximation algorithms.
Annals of the Institute of Mathematical Statistics, 40:641–663.

Geman, S. and Graffigne, C. (1986).

Markov random field image models and their applications to computer vision.
In *Proceedings of the International Congress of Mathematicians*, pages 1496–1517.

Hennig, C. (2000).

Identifiability of models for clusterwise linear regression.
Journal of Classification, 17:273–296.

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

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

Ji, C. and Seymour, L. (1996).

A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood.
Annals of Applied Probability, 6:423–443.

References II

McLachlan, G. J. (1992).

Discriminant Analysis And Statistical Pattern Recognition.
Wiley, New York.

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, 33:1735–1748.

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.

Stanford, C. D. and Raftery, A. E. (2002).

Approximate Bayes factor for image segmentation: the pseudolikelihood information criterion (PLIC).
IEEE Transactions on Pattern Analysis and Machine Intelligence, 24:1517–1520.