THE UNIVERSITY *of* EDINBURGH
School of Mathematics

Bayesian Cluster Analysis

Sara Wade
University of Edinburgh

Joint work with Zoubin Ghahramani (University of Cambridge & Uber)

O'Bayes Conference, University of Warwick
28 June - 2 July, 2019

Outline

Outline

Mixture models[1]

The data are conditionally iid with density

$$f(y|P) = \int K(y|\theta)dP(\theta),$$

for parametric density $K(y|\theta)$, $y \in \mathcal{Y}$, $\theta \in \Theta$, and P is a p.m. on Θ.

[1] Frühwirth-Schnatter, Celeux, Robert (2019)

Mixture models[1]

The data are conditionally iid with density

$$f(y|P) = \int K(y|\theta) dP(\theta),$$

for parametric density $K(y|\theta)$, $y \in \mathcal{Y}$, $\theta \in \Theta$, and P is a p.m. on Θ.

In Bayesian setting: prior on P s.t.

$$P = \sum_{j=1}^{J} w_j \delta_{\theta_j} \text{ a.s.} \quad \rightarrow f(y|P) = \sum_{j=1}^{J} w_j K(y|\theta_j).$$

This induces a random partition: data points belong to the same cluster if they were generated from the same mixture component.

[1] Frühwirth-Schnatter, Celeux, Robert (2019)

How to choose J?

1. Bayes Factors/information criterion.
2. Mixtures of finite mixtures (MFM): prior on J.
3. Overfitted mixtures.
4. Infinite mixtures.

Mixtures of finite mixtures

Inference:

- Computational issue: dimension of the parameter space changes with J - 'one of the things we don't know is the number of things we don't know'.
- Addressed through, e.g. reversible-jump MCMC (Richardson and Green (1997); Miller and Harrison (2018)).

Positive asymptotics: e.g. Nobile (1994), Guha, Ho, and Nguyen (2019).

Overfitted (sparse) mixtures

- Set J to be larger than the 'true/expected' number of clusters.
- Define: k_N to be the number of clusters in the sample $(y_1, \ldots, y_N)$.

Overfitted (sparse) mixtures

- Set J to be larger than the 'true/expected' number of clusters.
- Define: k_N to be the number of clusters in the sample $(y_1, \ldots, y_N)$.

Will extra components automatically be emptied?

Overfitted (sparse) mixtures

- Set J to be larger than the 'true/expected' number of clusters.
- Define: k_N to be the number of clusters in the sample $(y_1, \ldots, y_N)$.

Will extra components automatically be emptied? yes!

Rousseau and Mengersen (2011): Assume:

$$(w_1, \ldots, w_J) \sim \text{Dir}(\alpha_1, \ldots, \alpha_J),$$

overfitted mixture converges to the truth if $\alpha_j < d/2$, with d denoting the dimension of θ (in this called, called sparse mixtures).

Infinite mixtures

Set $J = \infty$ and define k_N to be the number of clusters in the sample $(y_1, \ldots, y_N)$.

Specify a prior on P s.t. $P = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$: e.g. Dirichlet process, Pitman-Yor (Lijoi and Prünster (2010))

Inference: collapsed Gibbs (Neal (2000)); slice sampling (Kalli, Griffin, and Walker (2011)); finite-dimensional approximations (Ishwaran and James (2001)).

Mixed asymptotics: e.g. Ghosal and van der Vaart (2007); Miller and Harrison (2013); Rajkowski (2019).

Outline

Summarising the posterior of the partition

These approaches (MFM, overfitted, infinite) provide a model-based framework for Bayesian cluster analysis, returning a posterior over a large partition space, expressing belief and uncertainty in the clustering given the data.

How can we summarize this posterior?

Summarising the posterior of the partition

These approaches (MFM, overfitted, infinite) provide a model-based framework for Bayesian cluster analysis, returning a posterior over a large partition space, expressing belief and uncertainty in the clustering given the data.

How can we summarize this posterior?

In typical Bayesian analysis, the posterior of a univariate parameter θ is summarized by a point estimate (ex. $\mathrm{E}[\theta|\mathcal{D}]$) with 95% credible interval. Can we extend these ideas to Bayesian cluster analysis?

Summarising the posterior of the partition

- Developing summary tools is complicated by the discrete and unordered nature and huge dimension of the partition space:

  Dimension of the entire partition space is $B_N = \sum_{k=1}^{N} S_{N,k}$, for example

$$B_{10} = 115975 \text{ and } B_{20} = 51724158235372.$$

Summarising the posterior of the partition

- Developing summary tools is complicated by the discrete and unordered nature and huge dimension of the partition space:

  Dimension of the entire partition space is $B_N = \sum_{k=1}^{N} S_{N,k}$, for example

  $$B_{10} = 115975 \text{ and } B_{20} = 51724158235372.$$

- MCMC is often used for posterior inference $\rightarrow$ How can we summarise the MCMC draws of clusterings?
- Summary measures such as mean and standard deviation are not applicable.
- Mode is unreliable due to huge dimension of space.

Summarising the posterior of the partition

Aim: develop summary tools for the posterior:

1. Point estimate of the clustering,
2. Representation of the uncertainty.

Point estimation

Point estimates used in practice:

- Last iteration of the MCMC- subject to high variability.
- Posterior mode - may be unrepresententive, computational drawbacks.
- Ad-hoc methods: posterior similarity matrix is used as input in hierarchical or partitioning algorithms.
- Decision theoretic approach: minimize the posterior expected loss.

Decision theoretic approach

Notation: clustering $\mathbf{c} = (c_1, \ldots, c_N)$ where $c_n = j$ if y_n is in cluster j.

- **Loss function:** $L(\mathbf{c}, \widehat{\mathbf{c}}) \geq 0$ is the loss of estimating the "true" $\mathbf{c}$ with $\widehat{\mathbf{c}}$.
- We seek $\mathbf{c}^*$ which minimizes

$$\mathsf{E}[L(\mathbf{c}, \widehat{\mathbf{c}})|\mathcal{D}].$$

- Note: $\mathbf{c}^*$ corresponds to the posterior mode under the $0-1$ loss
  - $0-1$ loss does not take into account similarity between two clusterings.
- Need an appropriate loss function.

Loss functions

Binder's loss: (Binder (1978)) [2]

$$\mathrm{B}(\mathbf{c}, \widehat{\mathbf{c}}) = \sum_{n<n'} a\mathbf{1}(c_n = c_{n'}, \widehat{c}_n \neq \widehat{c}_{n'}) + b\mathbf{1}(c_n \neq c_{n'}, \widehat{c}_n = \widehat{c}_{n'}),$$

If $a = b$, and the point estimate c^* is that which minimizes

$$\sum_{n<n'} \left| P(c_n = c_{n'} | \mathcal{D}) - \mathbf{1}(\widehat{c}_n = \widehat{c}_{n'}) \right| .$$

[2] used in Lau and Green (2007) and Dahl (2006).

Loss functions

N-**invariant Binder's loss:**

$$\tilde{\mathsf{B}}(\mathbf{c}, \widehat{\mathbf{c}}) = \frac{2}{N^2}\mathsf{B}(\mathbf{c}, \widehat{\mathbf{c}})$$
$$= \sum_{i=1}^{k_N}\left(\frac{n_{i+}}{N}\right)^2 + \sum_{j=1}^{\widehat{k}_N}\left(\frac{n_{+j}}{N}\right)^2 - 2\sum_{i=1}^{k_N}\sum_{j=1}^{\widehat{k}_N}\left(\frac{n_{ij}}{N}\right)^2,$$

where $n_{ij} = \sum_n \mathbf{1}(c_n = i, \hat{c}_n = j)$; $n_{i+} = \sum_j n_{ij}$; $n_{+j} = \sum_i n_{ij}$.

Loss functions

Variation of information: (Meilă (2007)) constructed from information theory

$$\mathsf{VI}(\mathbf{c}, \widehat{\mathbf{c}}) = \mathsf{H}(\mathbf{c}) + \mathsf{H}(\widehat{\mathbf{c}}) - 2\mathsf{I}(\mathbf{c}, \widehat{\mathbf{c}})$$

$$= -\sum_{i=1}^{k_N} \frac{n_{i+}}{N} \log\left(\frac{n_{i+}}{N}\right) - \sum_{j=1}^{\widehat{k}_N} \frac{n_{+j}}{N} \log\left(\frac{n_{+j}}{N}\right) - 2\sum_{i=1}^{k_N}\sum_{j=1}^{\widehat{k}_N} \frac{n_{ij}}{N} \log\left(\frac{n_{ij}N}{n_{i+}n_{+j}}\right)$$

$$= \sum_{i=1}^{k_N} \frac{n_{i+}}{N} \log\left(\frac{n_{i+}}{N}\right) + \sum_{j=1}^{\widehat{k}_N} \frac{n_{+j}}{N} \log\left(\frac{n_{+j}}{N}\right) - 2\sum_{i=1}^{k_N}\sum_{j=1}^{\widehat{k}_N} \frac{n_{ij}}{N} \log\left(\frac{n_{ij}}{N}\right).$$

Properties of $\tilde{\mathrm{B}}$ and VI

Both $\tilde{\mathrm{B}}$ and VI are metrics and "aligned" with the lattice of partitions.

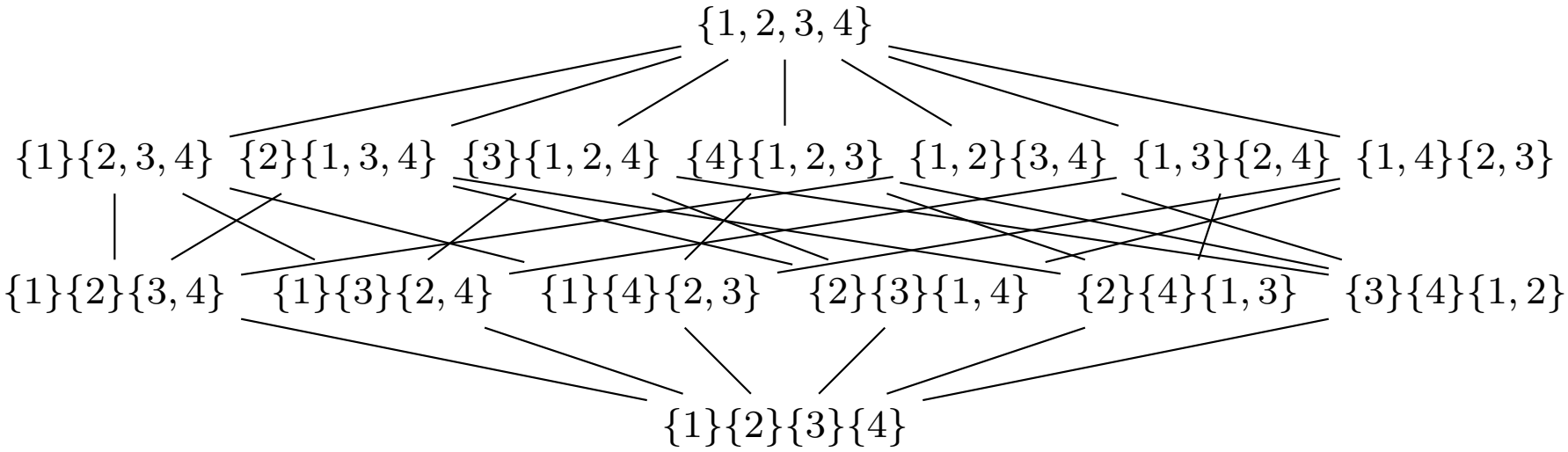


Figure: Hasse diagram for the lattice of partitions with a sample of size $N = 4$.

Properties of $\tilde{B}$ and VI

Property (Meilă (2007))

Both VI and $\tilde{B}$ are metrics on the space of partitions that satisfy,

i) *if* $\mathbf{c} \geq \widehat{\mathbf{c}} \geq \widehat{\widehat{\mathbf{c}}}$, *then*

$$d(\mathbf{c}, \widehat{\widehat{\mathbf{c}}}) = d(\mathbf{c}, \widehat{\mathbf{c}}) + d(\widehat{\mathbf{c}}, \widehat{\widehat{\mathbf{c}}}).$$

ii) *for any* $\mathbf{c}, \widehat{\mathbf{c}}$,

$$d(\mathbf{c}, \widehat{\mathbf{c}}) = d(\mathbf{c}, \widehat{\mathbf{c}} \wedge \mathbf{c}) + d(\widehat{\mathbf{c}}, \widehat{\mathbf{c}} \wedge \mathbf{c}).$$

Comparison of $\tilde{\mathrm{B}}$ and VI

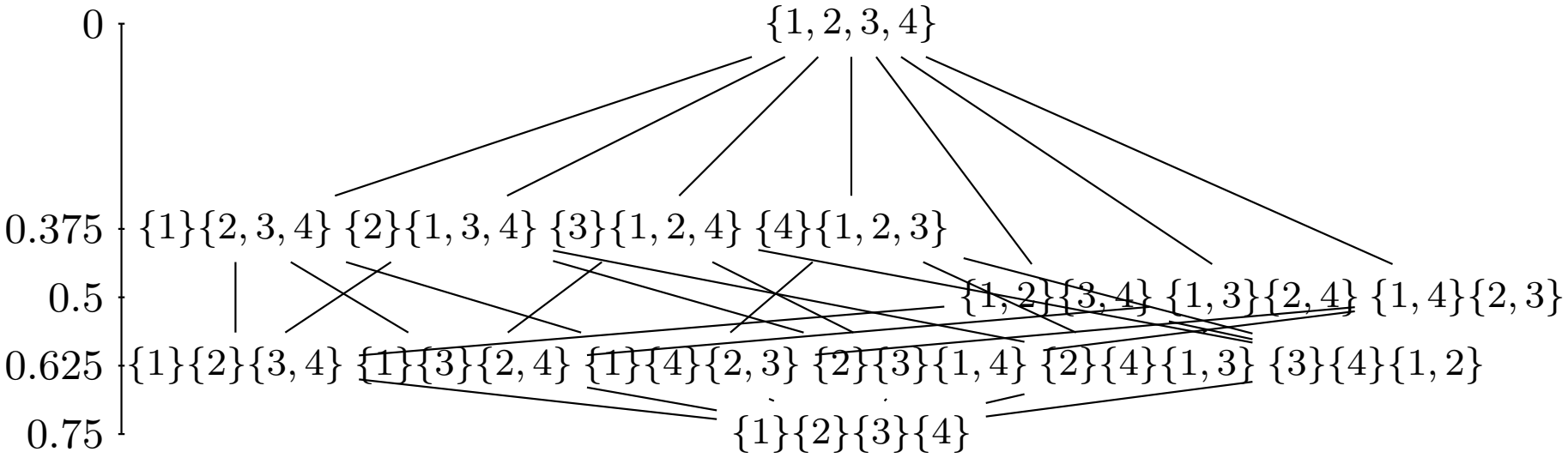


Figure: Hasse diagram stretched by $\tilde{\mathrm{B}}$ with a sample of size $N = 4$.

Comparison of $\tilde{\mathrm{B}}$ and VI

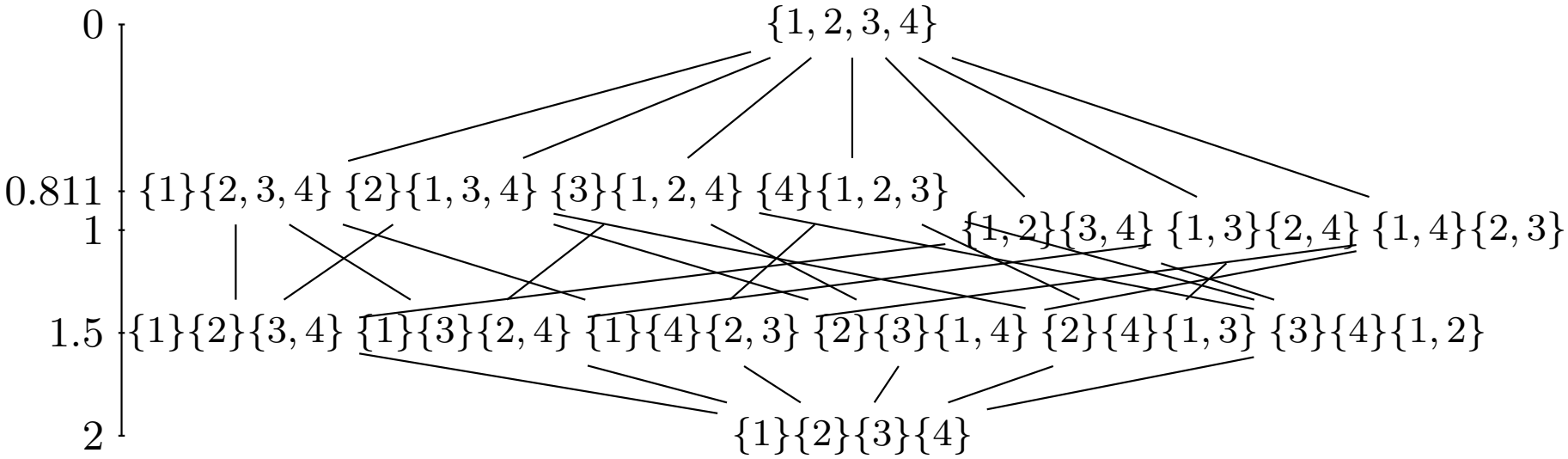


Figure: Hasse diagram stretched by VI with a sample of size $N = 4$. Note $2 - \frac{3}{4}\log(3) \approx 0.811$

Comparison of $\tilde{B}$ and VI

Property

Suppose N is divisible by k, and let $\mathbf{c}_k$ denote a partition with k clusters of equal size N/k.

$$\tilde{B}(\mathbf{1}, \mathbf{c}_k) = 1 - \frac{1}{k} > \frac{1}{k} - \frac{1}{N} = \tilde{B}(\mathbf{0}, \mathbf{c}_k).$$

$$VI(\mathbf{1}, \mathbf{c}_k) = \log(k) \leq \log(N) - \log(k) = VI(\mathbf{0}, \mathbf{c}_k), \quad \text{for } k \leq \sqrt{N},$$

and

$$VI(\mathbf{1}, \mathbf{c}_k) = \log(k) \geq \log(N) - \log(k) = VI(\mathbf{0}, \mathbf{c}_k), \quad \text{for } k \geq \sqrt{N}.$$

Comparison of $\tilde{B}$ and VI

Property

Suppose N is an even and square integer. Then, the partitions with two clusters of sizes $n = \frac{1}{2}(N - \sqrt{N})$ and $N - n$ are equally distance from $\mathbf{1}$ and $\mathbf{0}$ under $\tilde{B}$.

Point estimate under VI

Under the VI, the optimal partition $\mathbf{c}^*$ is

$$\begin{aligned}\underset{\widehat{\mathbf{c}}}{\operatorname{argmin}} \sum_{n=1}^{N} \log(1 + \sum_{n' \neq n} \mathbf{1}(\widehat{c}_{n'} = \widehat{c}_n)) \\ - 2 \sum_{n=1}^{N} \mathsf{E}[\log(1 + \sum_{n' \neq n} \mathbf{1}(c_{n'} = c_n)\mathbf{1}(\widehat{c}_{n'} = \widehat{c}_n)) | \mathcal{D}],\end{aligned} \tag{1}$$

To reduce computations, we can swap the log and expectation:

$$\begin{aligned}\underset{\widehat{\mathbf{c}}}{\operatorname{argmin}} \sum_{n=1}^{N} \log(1 + \sum_{n' \neq n} \mathbf{1}(\widehat{c}_{n'} = \widehat{c}_n)) \\ - 2 \sum_{n=1}^{N} \log(1 + \sum_{n' \neq n} P(c_{n'} = c_n | \mathcal{D})\mathbf{1}(\widehat{c}_{n'} = \widehat{c}_n)).\end{aligned} \tag{2}$$

Note: from Jensen's inequality, (2) provides a lower bound to (1).

Greedy search algorithm

Greedy search algorithm based on the Hasse diagram:

- For $i = 1, \ldots I$
  - Find the L closest partitions that cover $\widehat{\mathbf{c}}$ and the L closest partitions that $\widehat{\mathbf{c}}$ covers.
  - Compute $\mathsf{E}[L(\mathbf{c}, \widehat{\widehat{\mathbf{c}}})|\mathcal{D}]$ for all $2L$ partitions and select the partition $\mathbf{c}'$ with minimal $\mathsf{E}[L(\mathbf{c}, \mathbf{c}')|\mathcal{D}]$.
  - If $\mathsf{E}[L(\mathbf{c}, \mathbf{c}')|\mathcal{D}] < \mathsf{E}[L(\mathbf{c}, \widehat{\mathbf{c}})|\mathcal{D}]$, set $\widehat{\mathbf{c}} = \mathbf{c}'$. Otherwise, STOP.
- end

Credible balls

Credible ball around $\mathbf{c}^*$ of level $1-\alpha$, $\alpha \in [0,1]$:

$$B_{\epsilon^*}(\mathbf{c}^*) = \{\mathbf{c} : d(\mathbf{c}^*, \mathbf{c}) \leq \epsilon^*\},$$

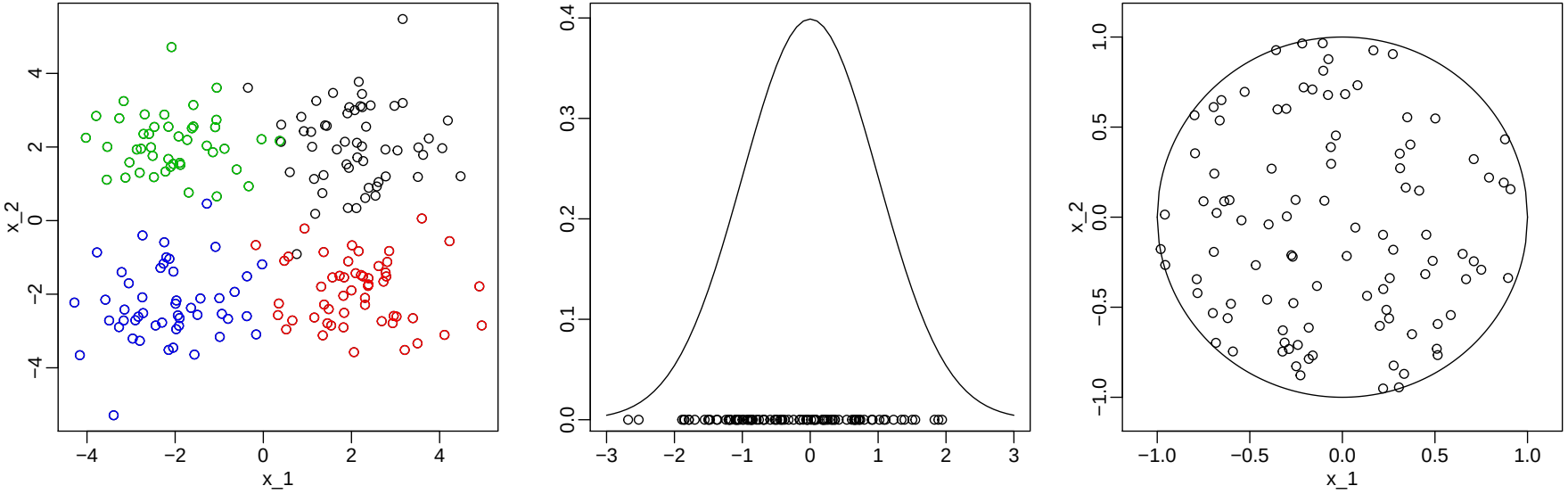
where ϵ^* is the smallest $\epsilon \geq 0$ such that

$$P(B_\epsilon(\mathbf{c}^*)|\mathcal{D}) \geq 1-\alpha.$$

- We can estimate ϵ^* from MCMC.
- To summarize the credible ball, we report the *vertical* and *horizontal bounds*:
  - **Vertical upper bounds:** partitions with the smallest number of clusters which are most distant from $\mathbf{c}^*$.
  - **Vertical lower bounds:** partitions with the largest number of clusters which are most distant from $\mathbf{c}^*$.
  - **Horizontal bounds:** partitions which are most distant from $\mathbf{c}^*$.

Outline

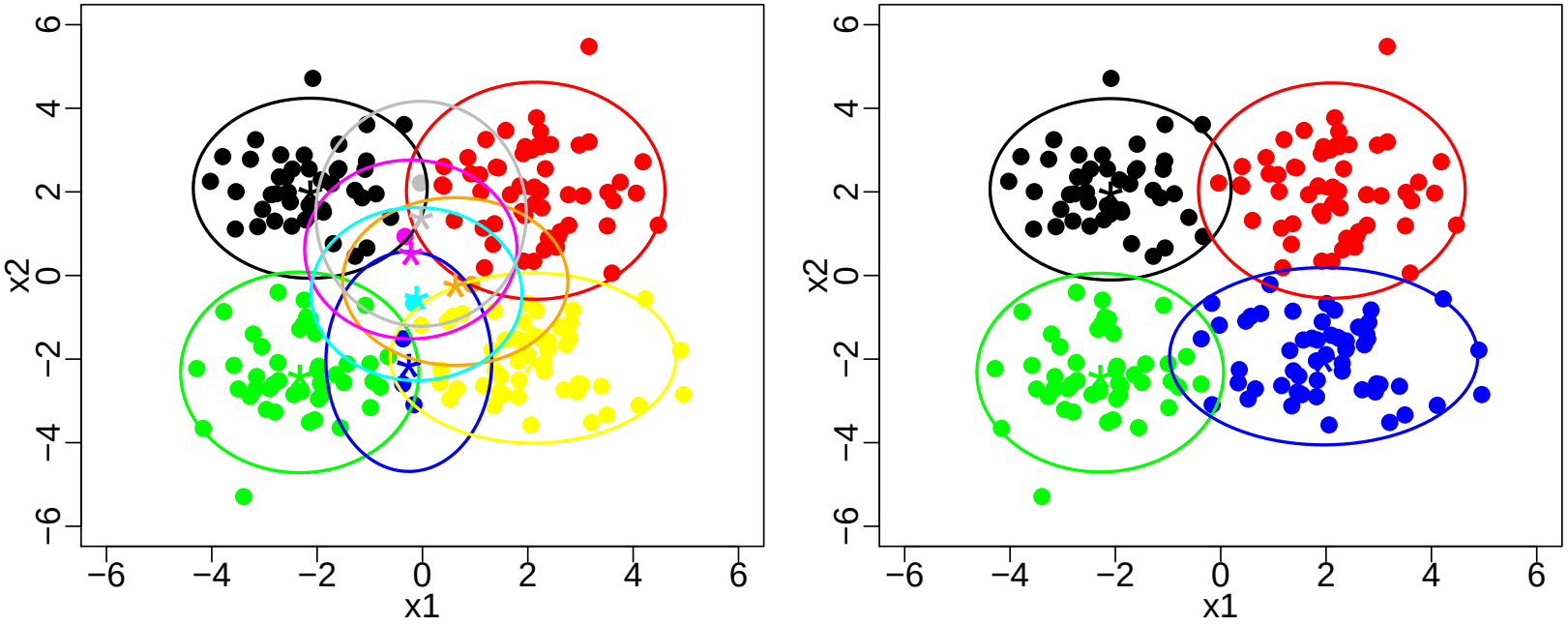
Simulated Examples



(a) Ex 1: 4 clusters (b) Ex Miller: 1 clusters (c) Ex Rajkowski: 1 clusters

Figure: Simulated scenarios: 1) mixture of 4 bivariate Gaussians (W. and Ghahramani (2018)); 2) standard normal (Miller and Harrison (2013)); 3) uniform within the unit circle (Rajkowski (2019)). Perform MCMC inference for a DPM.

Simulated Example: $N = 200$



(a) $\tilde{\mathrm{B}}$: 9 clusters

(b) VI: 4 clusters

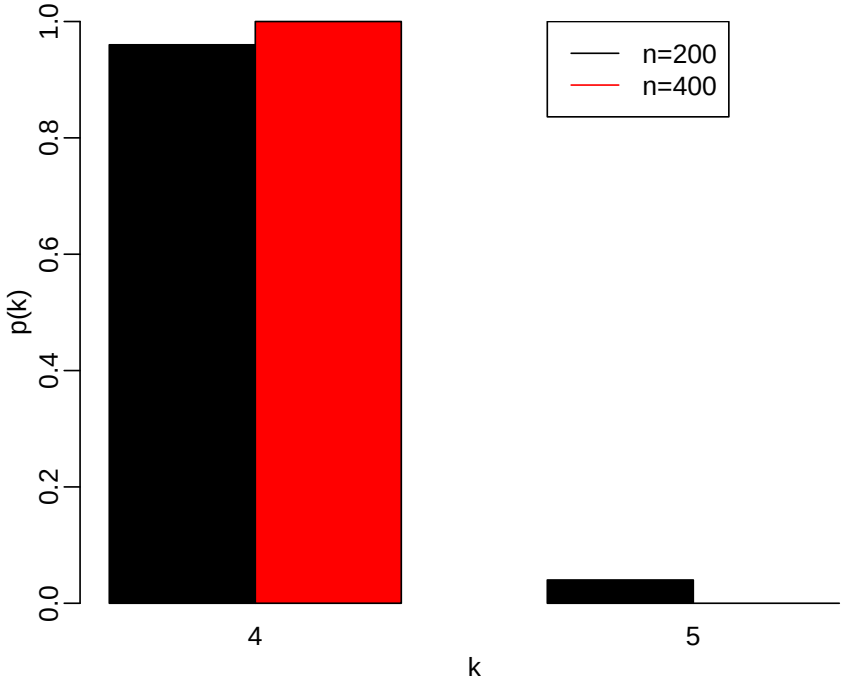
Figure: Point estimate of the clustering from a DPM.

Simulated Example: increasing N

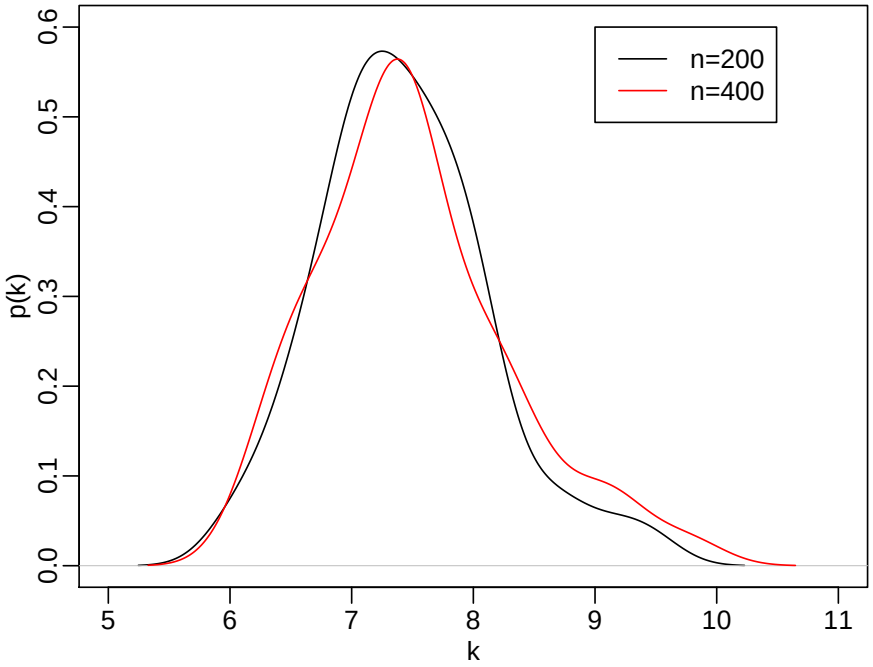
Ex 1	Loss	k_N^*	p_I	$\tilde{\mathrm{B}}(\mathbf{c}_t, \mathbf{c}^*)$	$\mathrm{VI}(\mathbf{c}_t, \mathbf{c}^*)$
$N = 200$:	$\tilde{\mathrm{B}}$	9	0.065	0.045	0.643
	VI	**4**	**0.045**	**0.044**	**0.569**
$N = 400$:	$\tilde{\mathrm{B}}$	17	0.0775	0.052	0.769
	VI	**4**	**0.045**	**0.044**	**0.54**
$N = 800$:	$\tilde{\mathrm{B}}$	24	0.0775	0.061	0.903
	VI	**4**	**0.0587**	**0.056**	**0.742**
$N = 1600$:	$\tilde{\mathrm{B}}$	41	0.0581	**0.044**	0.719
	VI	**4**	**0.0294**	0.045	**0.629**

Table: Comparison in terms of 1) number of clusters k_N^*; 2) proportion of data points incorrectly classified p_I; 3) $\tilde{\mathrm{B}}$ between the optimal and true clusterings; and 4) VI between the optimal and true clusterings.

Simulated Example: 100 repetitions



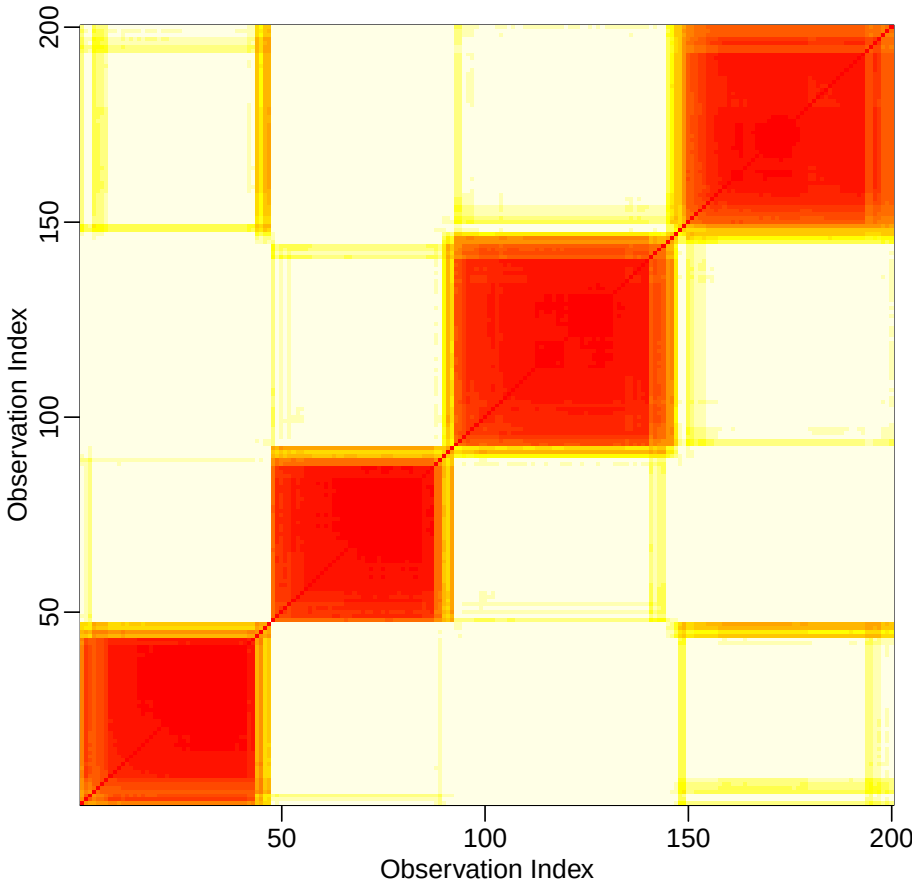
(a) VI estimate of k



(b) Posterior mean of k

Figure: (a) Distribution of $\widehat{k}$ from the VI clustering estimate and (b) (kernel) estimated density of the posterior mean k, across simulations.

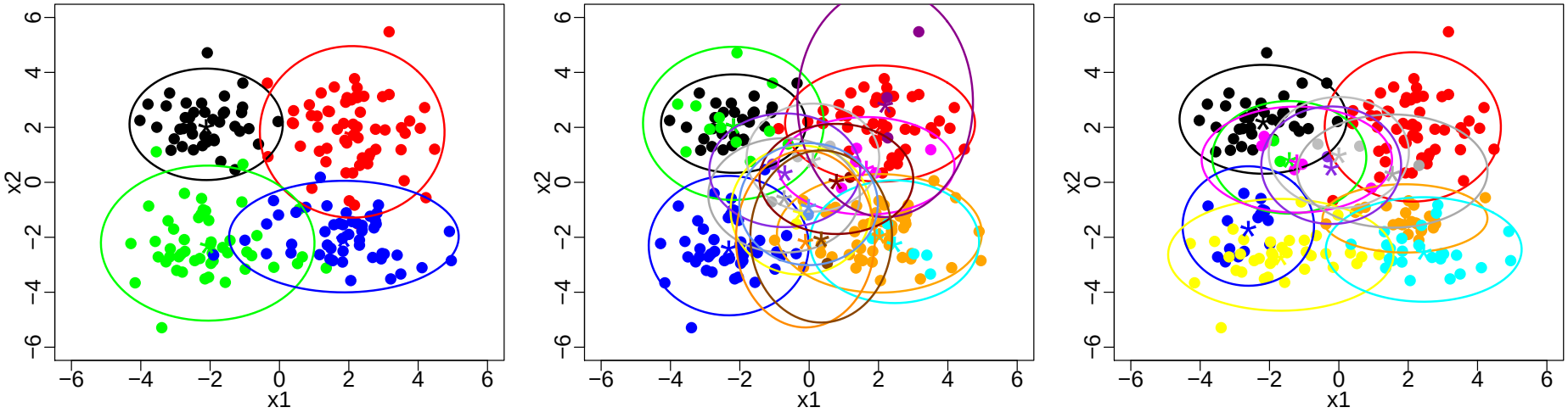
Simulated Example: $N = 200$



(a) Example 1

Figure: Posterior similarity matrix.

Simulated Example: $N = 200$



(a) VI upper vertical bound: 4 clusters

(b) VI lower vertical bound: 16 clusters

(c) VI horizontal bound: 7 clusters

Figure: 95% VI credible ball.

Simulated Example: 100 repetitions

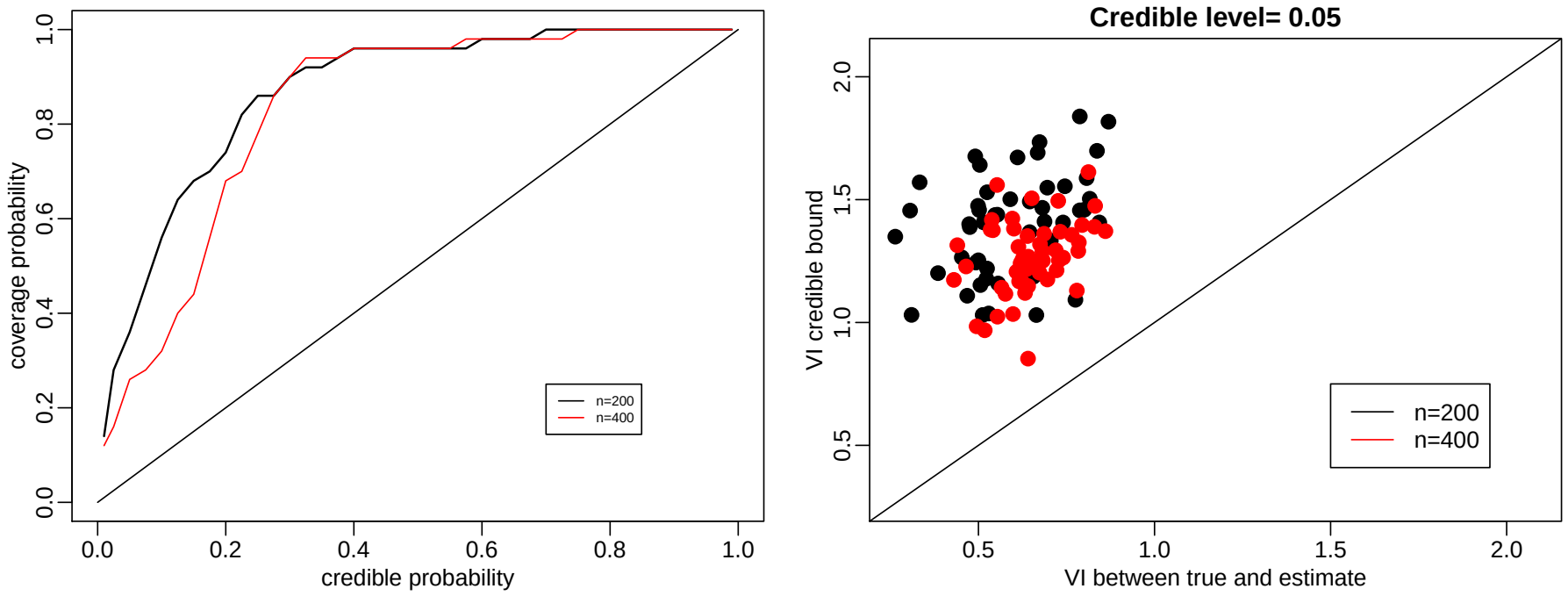
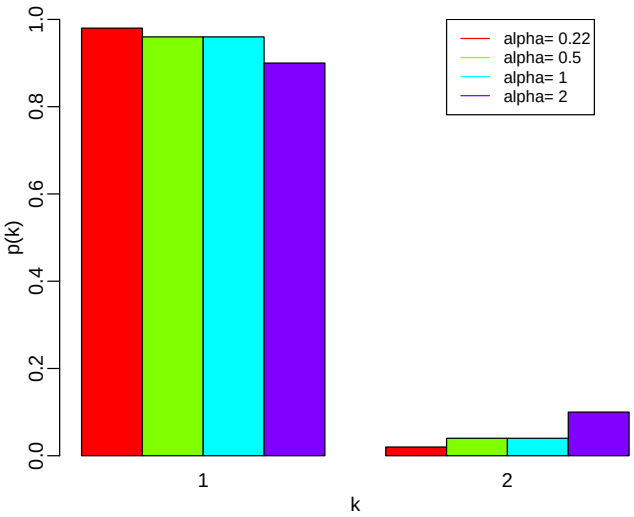
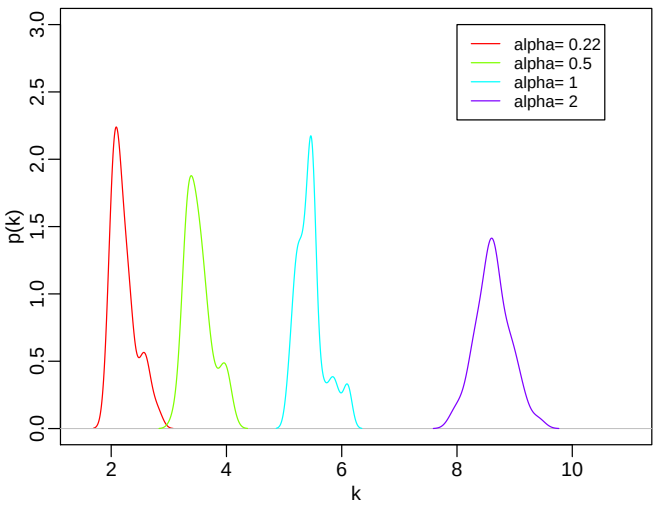


Figure: Coverage probability as a function of credible probability for the VI credible balls.

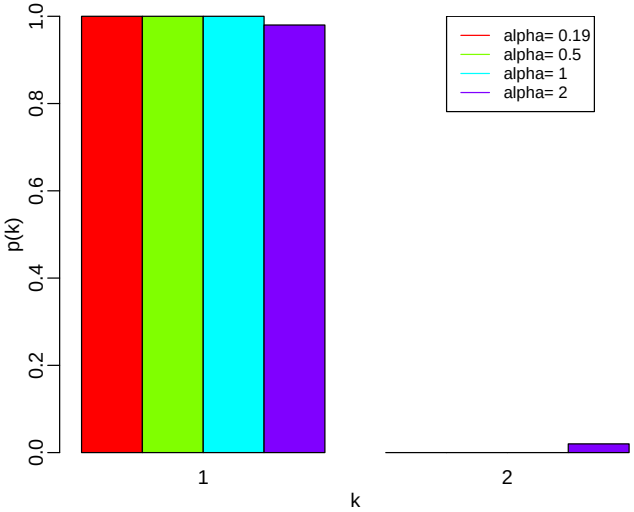
Miller & Harrison (2013) Example: 50 repetitions



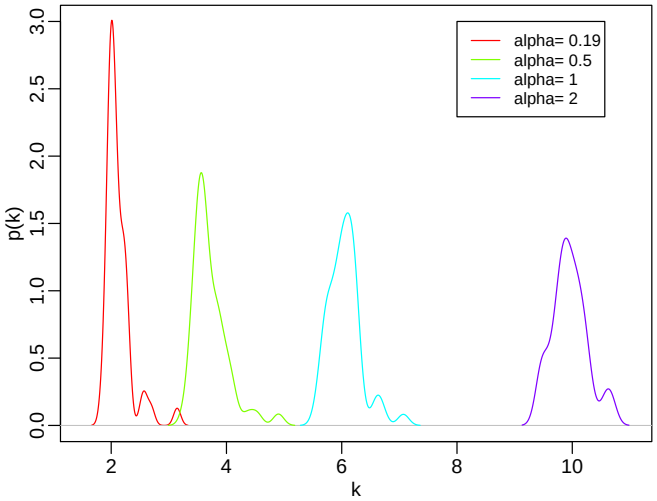
(a) VI partition: $N = 100$



(b) Post. mean of k: $N = 100$

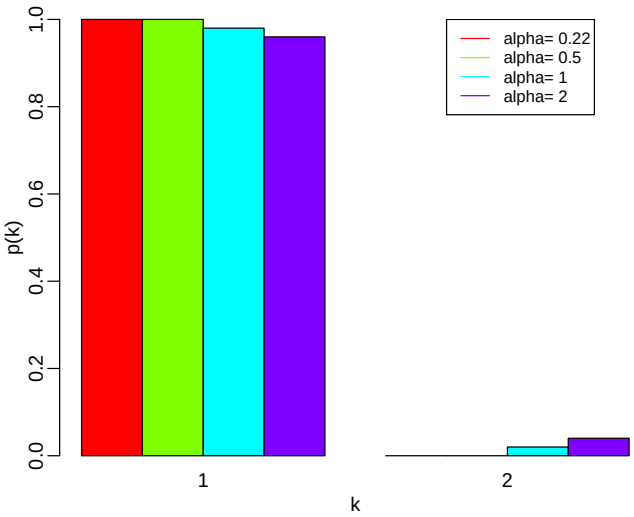


(c) VI partition: $N = 200$

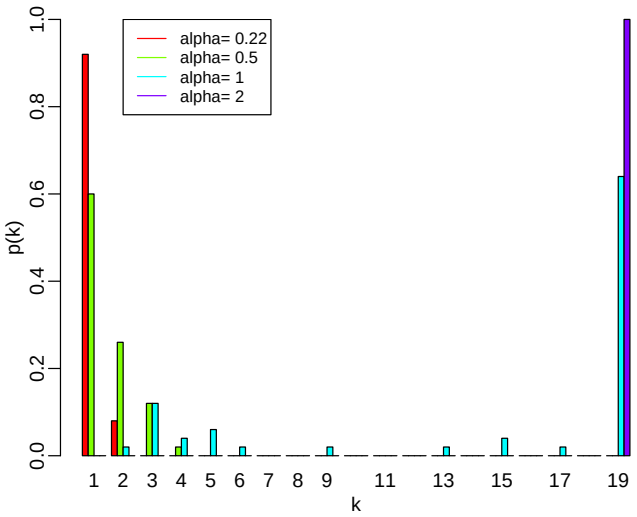


(d) Post. mean of k: $N = 200$

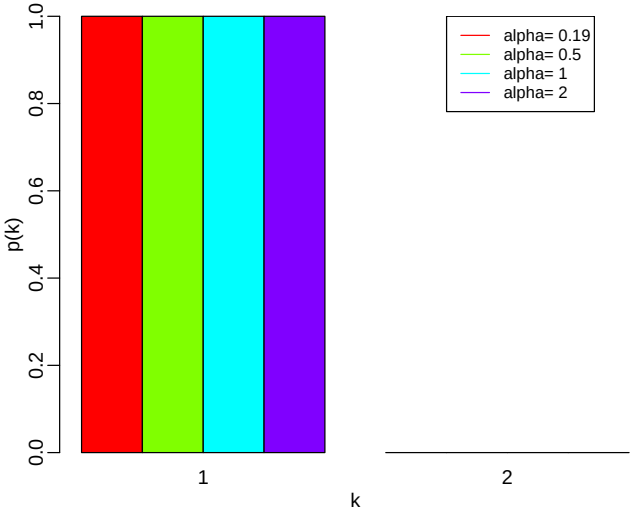
Miller & Harrison (2013) Example: 50 repetitions



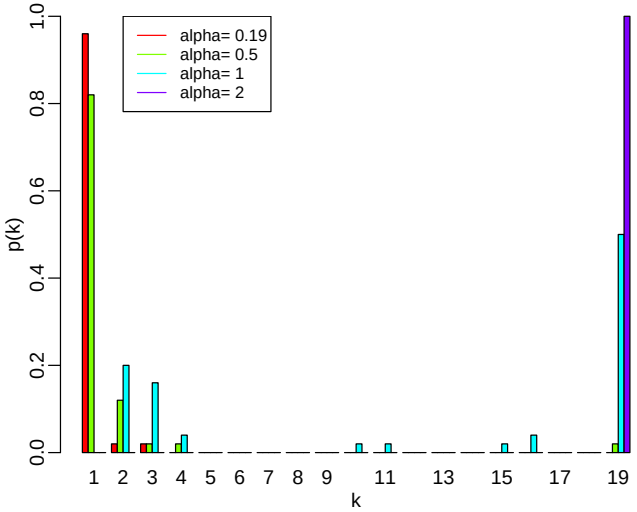
(e) MAP partition: $N = 100$



(f) Binder partition: $N = 100$

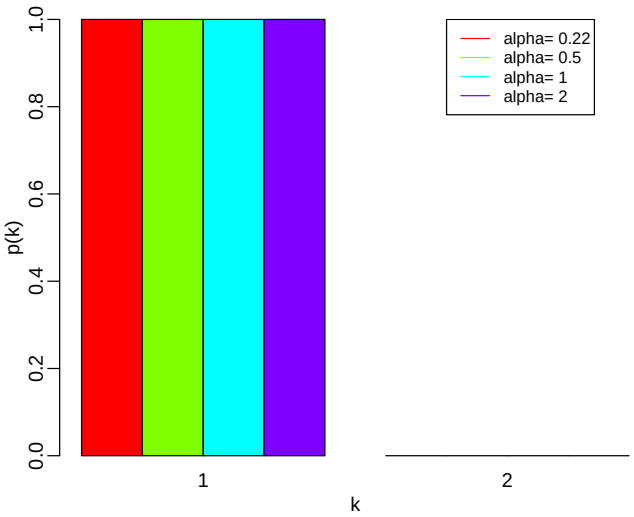


(g) MAP partition: $N = 200$

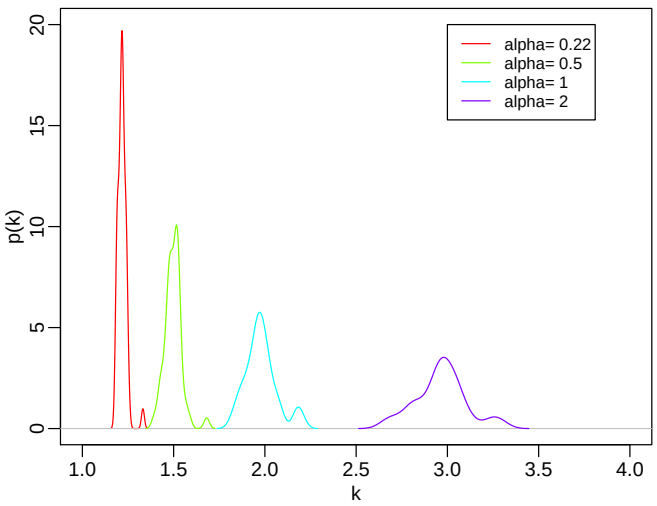


(h) Binder partition: $N = 200$

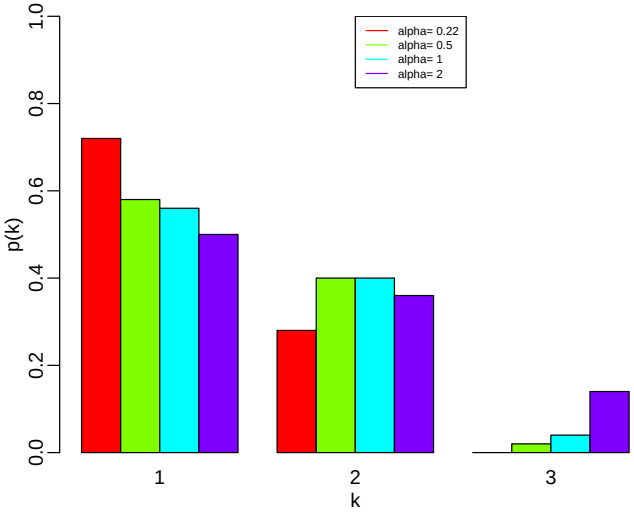
Rajkowski (2019) Example: 50 repetitions



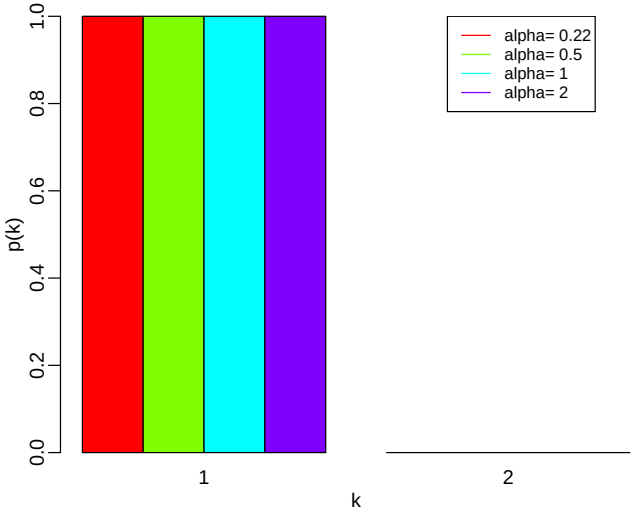
(i) VI partition: $N = 100$



(j) Post. mean of k: $N = 100$



(k) MAP partition: $N = 100$



(l) Binder partition: $N = 100$

Outline

Summary and Future Work

Summary:

- Developed summary tools for Bayesian cluster analysis:
  - point estimate through minimization of VI,
  - credible balls to describe uncertainty around estimate.
- Available in **R** package 'mcclust.ext' and **R** package 'sdols'. Alternative algorithms in Friel and Rastelli (2018) that are linear in N **R** package 'GreedyEPL'.
- Marginally focusing on the number of clusters may be misleading $\rightarrow$ estimate the clustering directly if it is the object of interest.

Future work:

- Extend to feature allocation analysis.
- Asymptotic study of the VI clustering estimate and coverage of the credible balls.
- Fast estimate of VI clustering estimate, avoiding MCMC.

Thanks!

https://projecteuclid.org/euclid.ba/1508378464
https://github.com/sarawade/mcclust.ext

We are grateful to all discussants for the their thoughtful comments and contributions.