QUT Centre for Data Science

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Bayesian Modelling and Analysis of Challenging Data

Kerrie Mengersen
School of Mathematical Sciences
QUT

PC Mahalanobis Lecture Series
January 2021

Programme of Lectures

January 27th:

- Lecture 1: 10-1045am IST (230pm-3:15pm AEST)
  Identifying the Intrinsic Dimension of High-Dimensional Data

- Lecture 2: 11-11:45am IST (3:30pm-4:15pm AEST)
  Finding Patterns in Highly Structured Spatio-Temporal Data

January 29th:

- Lecture 3: 10-1045am IST (230pm-3:15pm AEST)
  Describing Systems of Data

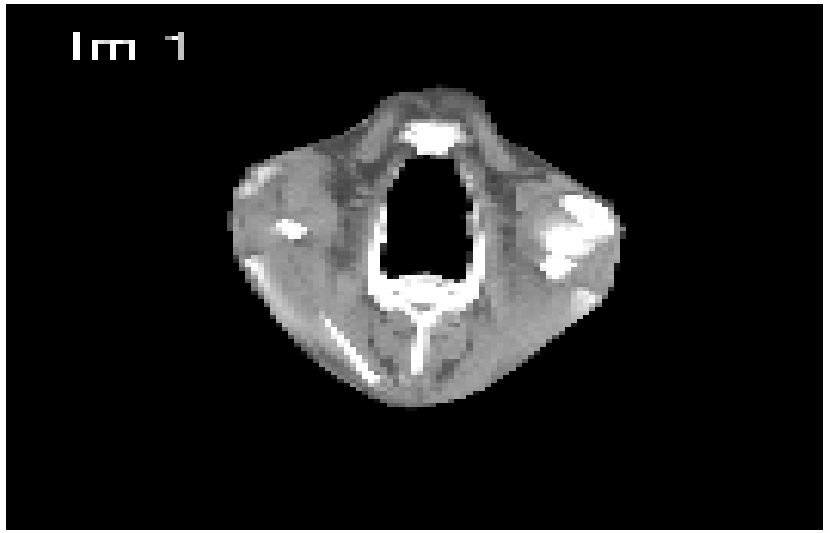
- Lecture 4: 11-11:45am IST (3:30pm-4:15pm AEST)
  Making New Sources of Data Trustworthy

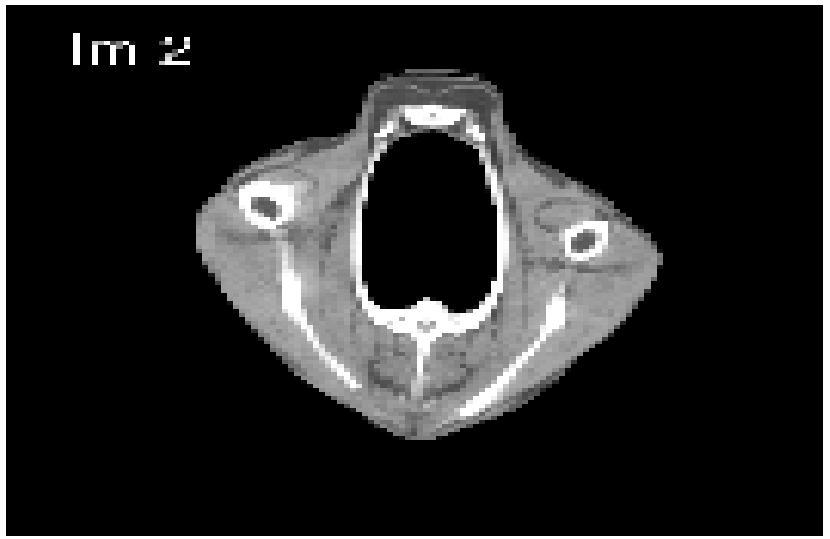
Bayesian Modelling and Analysis of Challenging Data

Lecture 2:
Finding patterns in highly structured spatio-temporal data

Clair Alston, Insha Ullah,
Edgar Santos-Fernandez, Erin Peterson,
Marcela Cespedes, Paul Wu,
Daniel Kennedy, Judith Rousseau

Case Study 1a: CAT scans of sheep

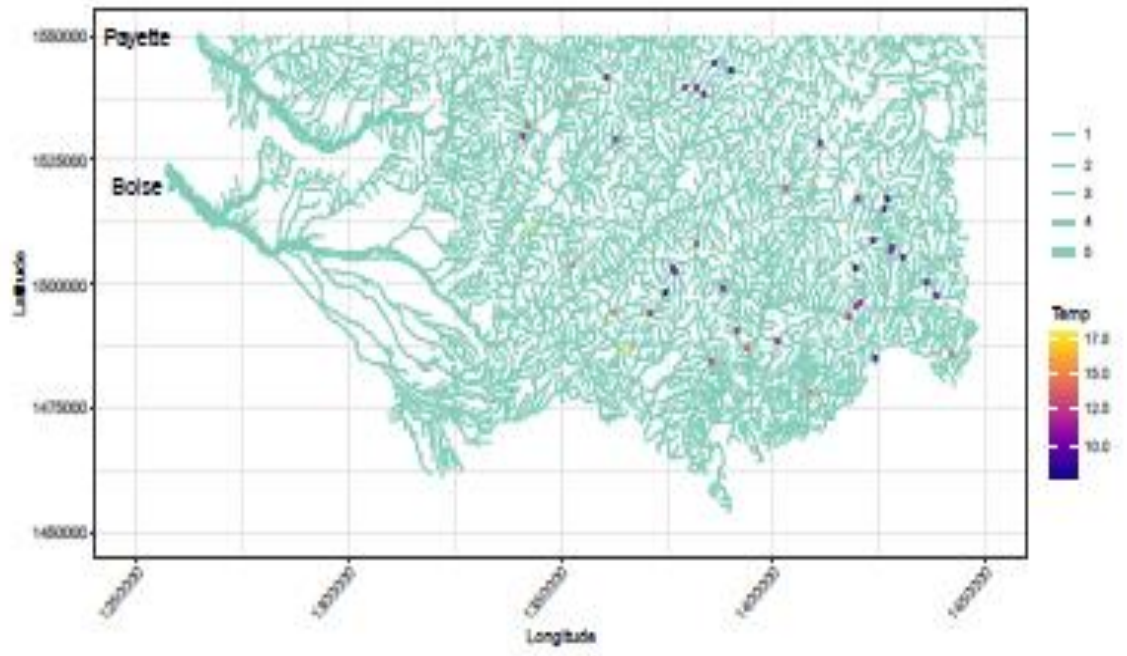


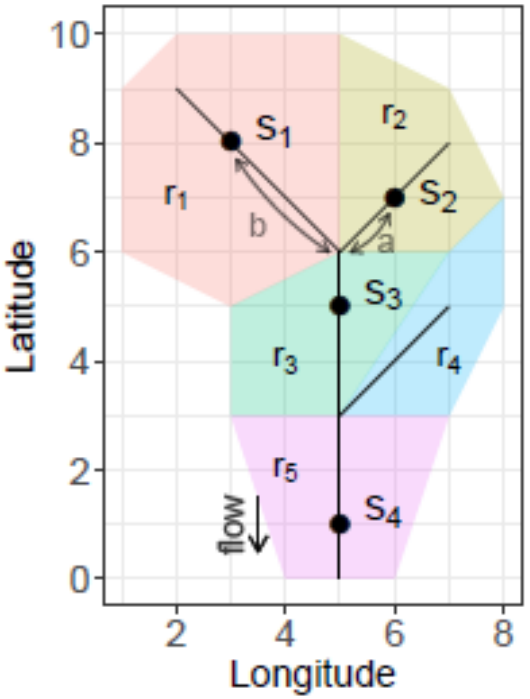


Case Study 1b: Feature detection in satellite data

- Fire ant surveillance program in Brisbane region since 2001.
- 17700 identified locations
- Want to focus eradication program on high risk areas.

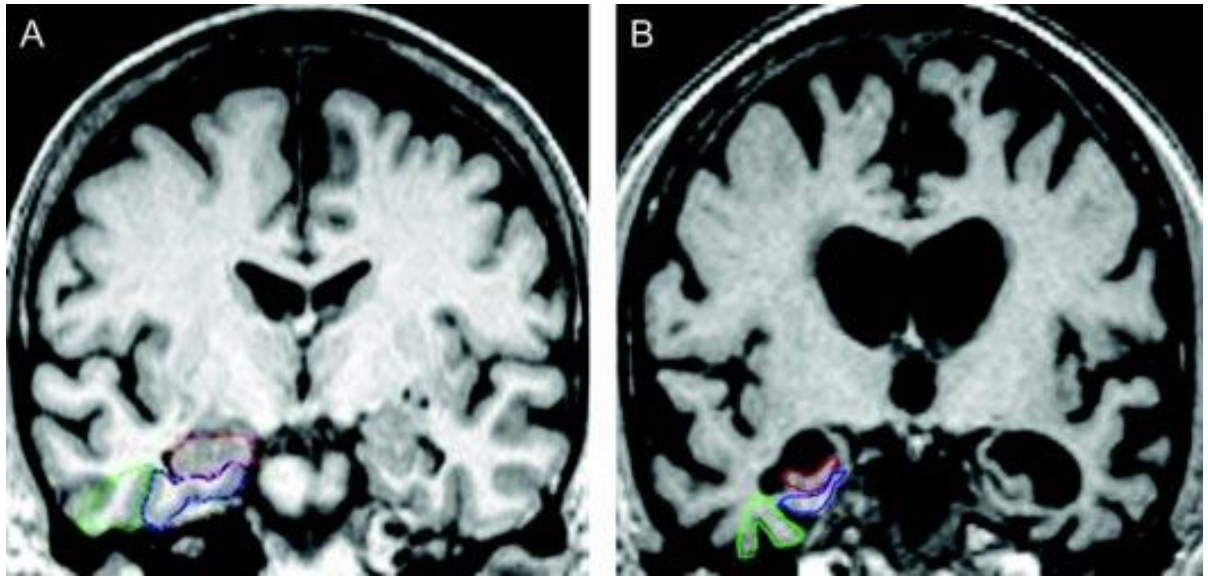
Case Study 2a: Stream Networks – Allowing for different spatial neighbourhoods





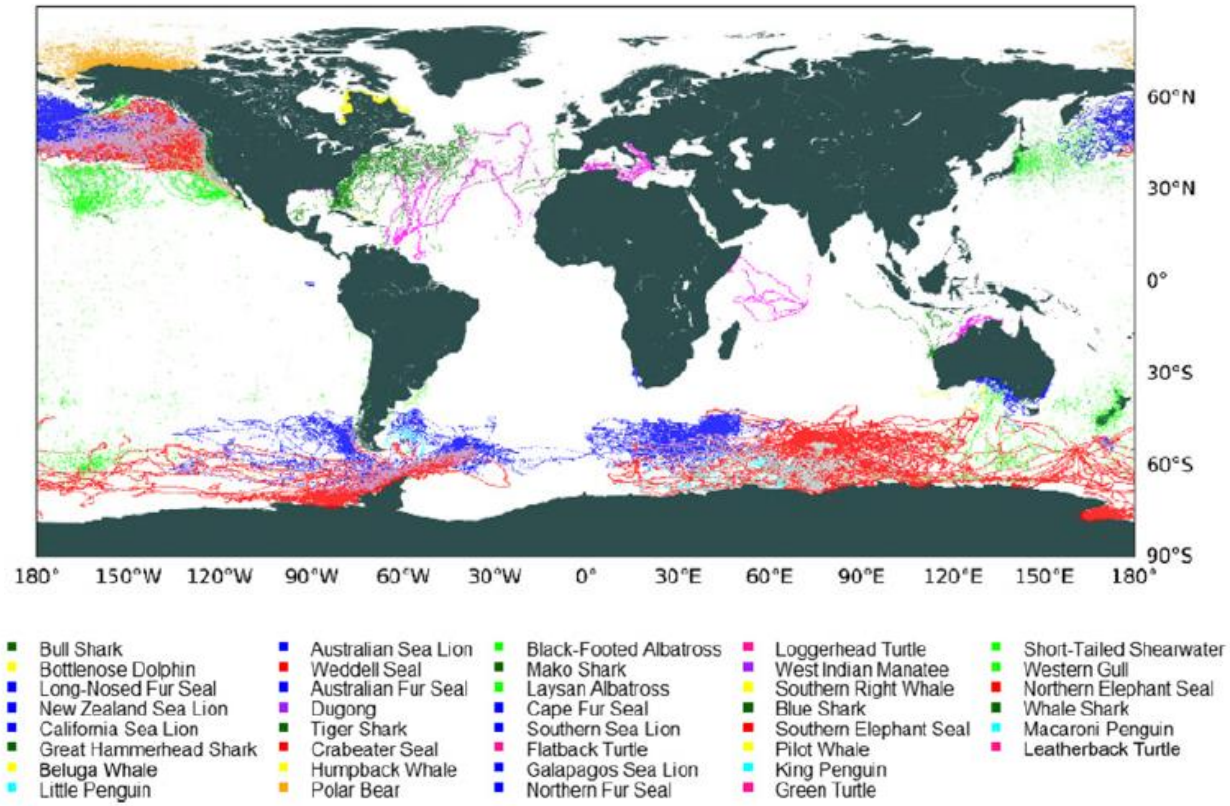
Santos-Fernandez *et al.* 2021
(In preparation)

Case Study 2b: MRI scans of brains – Estimating the spatial neighbourhood

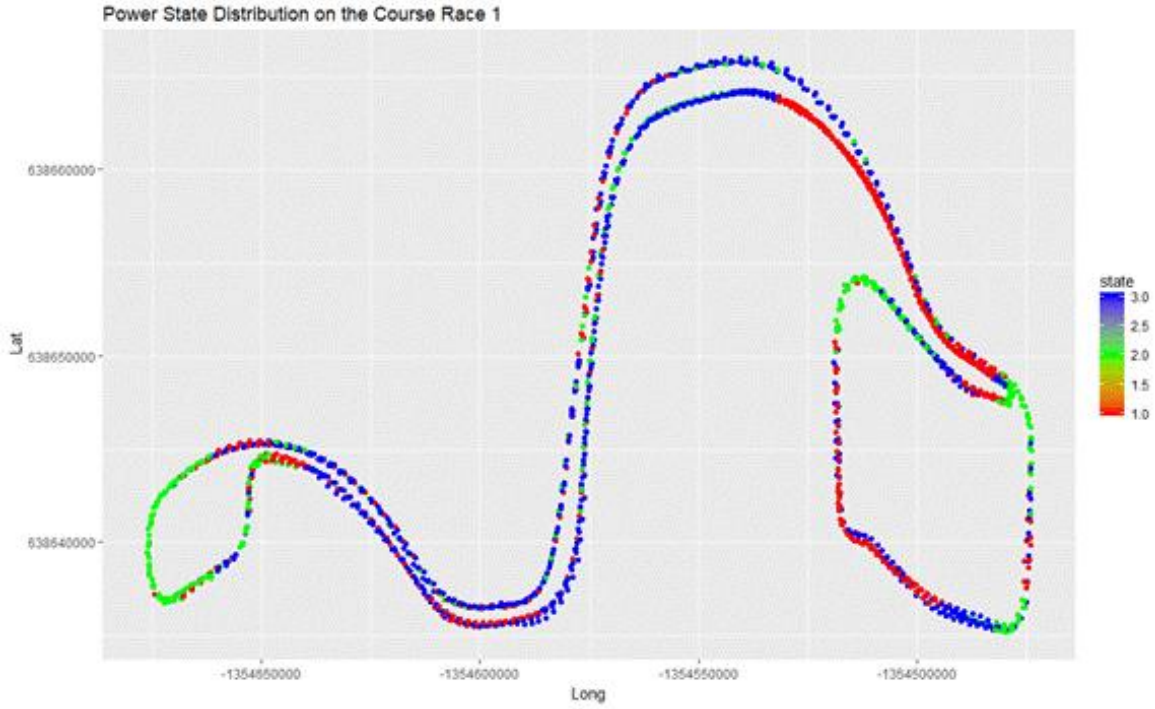


https://www.radiologyinfo.org/en/info.cfm?pg=alzheimers

Case Study 3: Hidden Markov Models



Sequeira *et al.*, PNAS, 2018



Wu *et al.*, QUT + Queensland Academy of Sport

Case Study 1a: Spatial analysis of images

- Spatial mixture models (Alston *et al.)*
- Spatial dynamic factor models (Strickland *et al.*)
- Mixture PCA approaches (Ullah *et al.*)
- Regression trees (Holloway *et al.*)

Bayesian Mixture Models

Aim: cluster real-valued observations in $Y = (y_1, \dots, y_n)$

each element is a p-dimensional realisation made independently over *n* objects.

$$f(y|\theta_1, \dots, \theta_K) = \sum_{k=1}^{K} \pi_k N_p(y|\theta_k) \qquad \theta_k = (\mu_k, \Sigma_k) \qquad 0 \le \pi_k \le 1, \sum_{k=1}^{K} \pi_k = 1$$

Define latent indicator $z_i, i = 1, \dots, n$, s.t. prior probability of assigning a particular observation y_i to a cluster k is

$$p(z_i = k|\pi) = \pi_k$$

Parametric or nonparametric priors on θ and π.

Examples – parametric representation

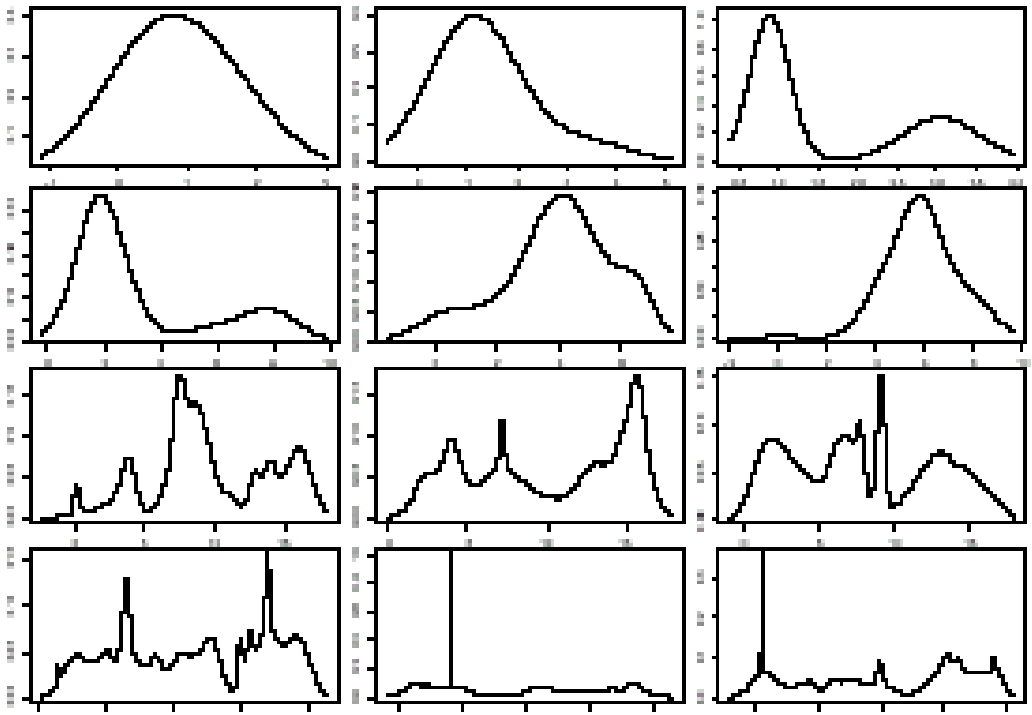
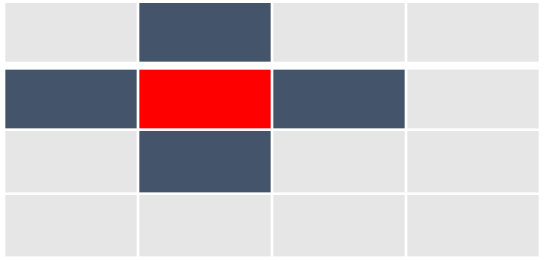


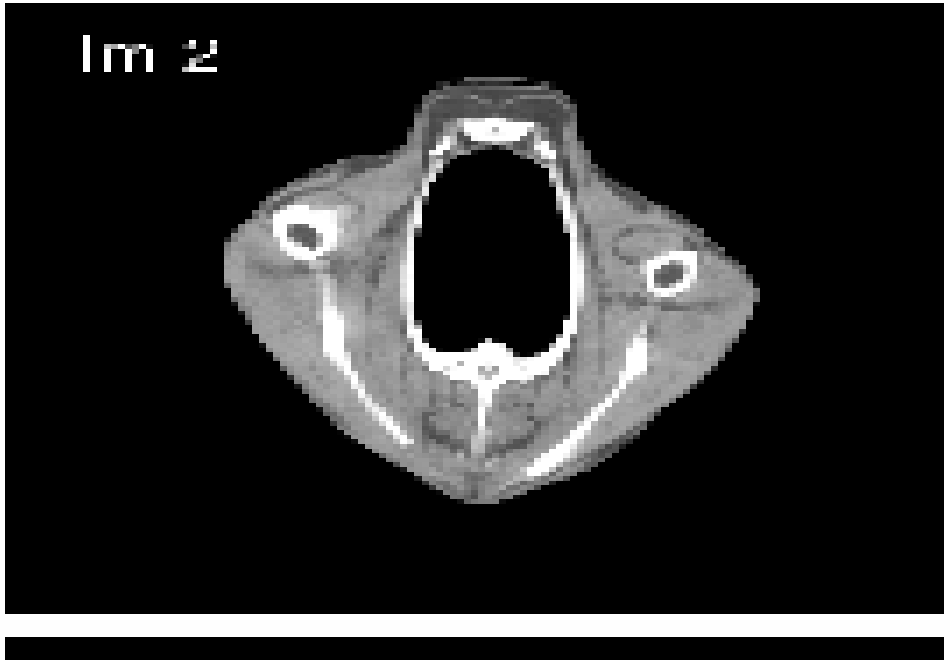
FIGURE 1. Some normal mixture densities for $K = 2$ *(first row)*, $K = 5$ *(second row)*, $K = 25$ *(third row)* and $K = 50$ *(last row)*.

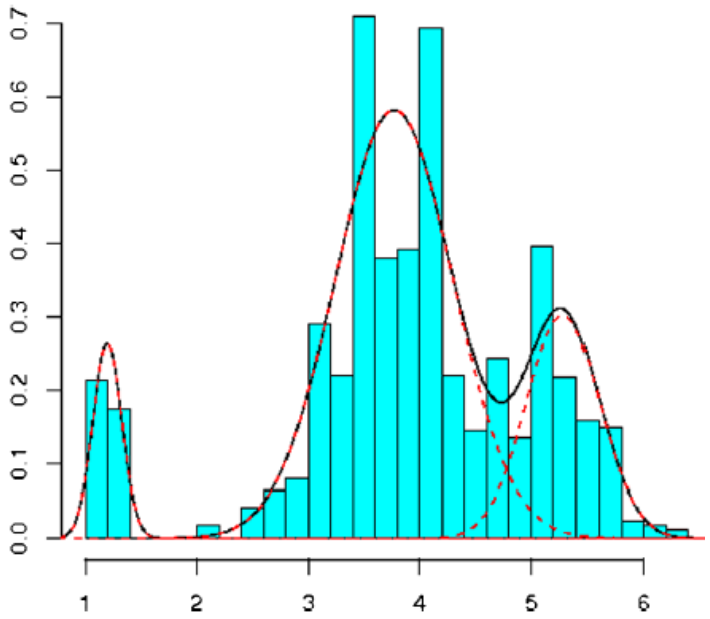
Spatial Mixture Models

$$\mu_k = X\beta + u_k + e_k$$

$$u_k \sim N(\bar{u}_{\sim k}, \frac{\sigma_u^2}{n_k}) \qquad e_k \sim N(0, \frac{\sigma_e^2}{n_k})$$

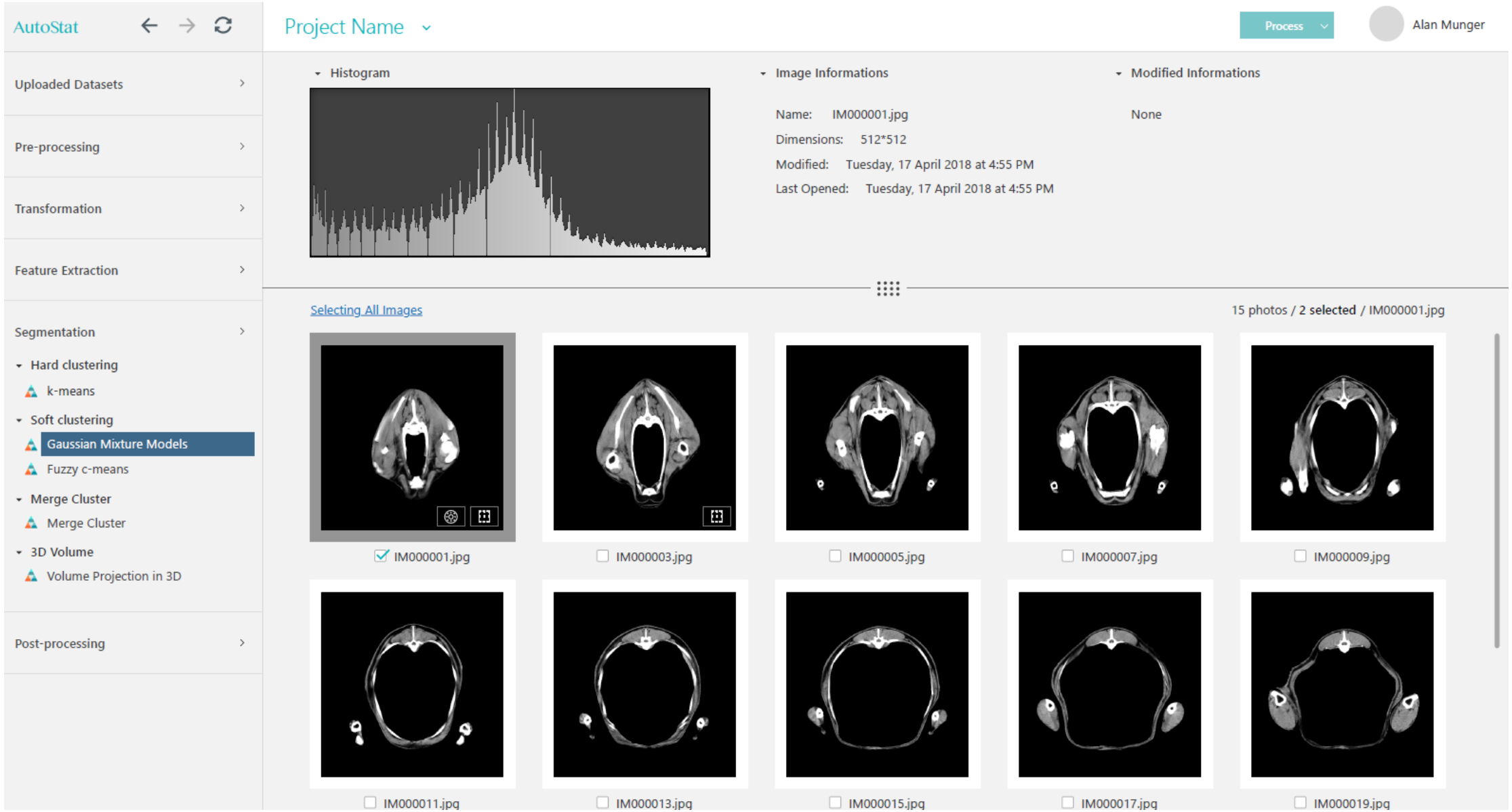






Bayesian Analysis via AutoStat

https://autostat.com.au/



AutoStat

Project Name

Process

Alan Munger

Uploaded Datasets

Pre-processing

Transformation

Feature Extraction

Segmentation

- Hard clustering
  - k-means
- Soft clustering
  - Gaussian Mixture Models
  - Fuzzy c-means
- Merge Cluster
  - Merge Cluster
- 3D Volume
  - Volume Projection in 3D

Post-processing

Image Visualisation | Summary

Add Transforms | 1 Gaussian Mixture Models

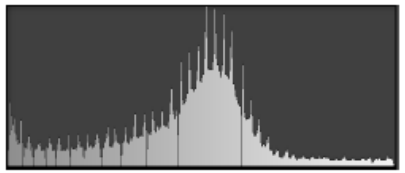
Apply

Gaussian Mixture Models

Auto AIC

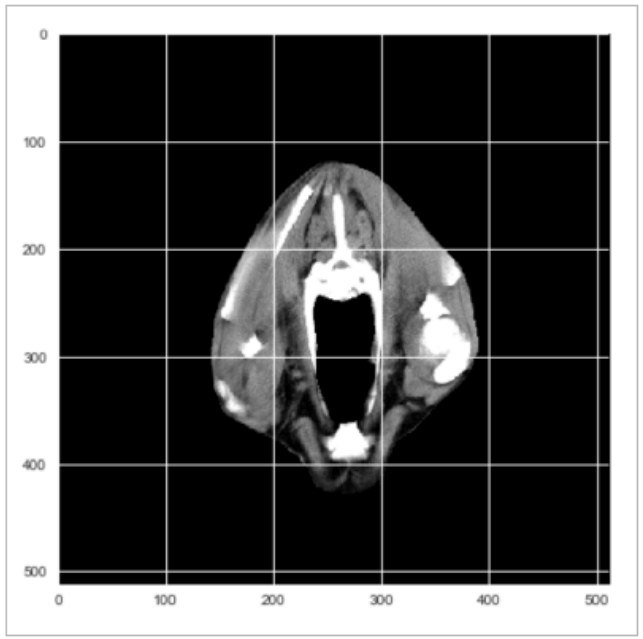
Manual Classify Numbers: 4

Histogram



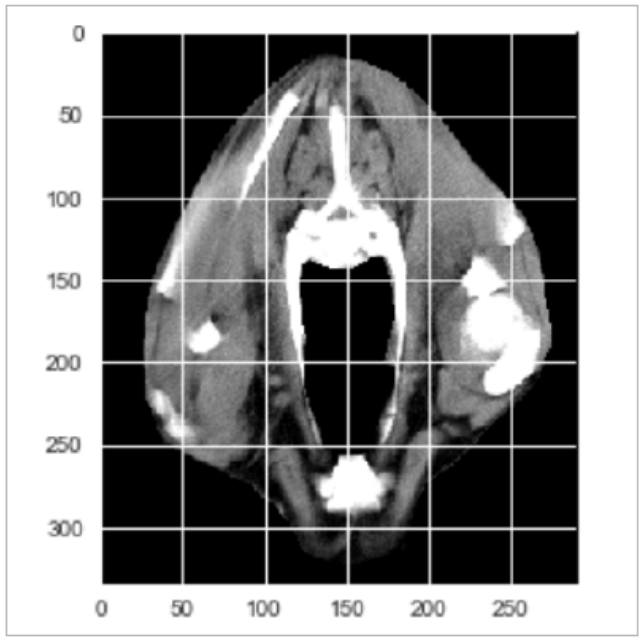
← Back to image library

Original Image

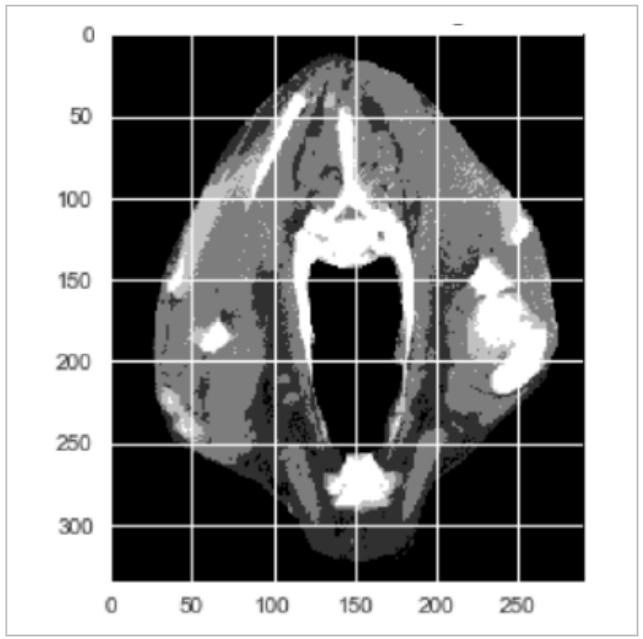


IM000001.jpg

Cropped Image



GMM Clustered Image



View all clusters

AutoStat

Project Name

Process

Alan Munger

Uploaded Datasets

Pre-processing

Transformation

Feature Extraction

Segmentation

- Hard clustering
  - k-means
- Soft clustering
  - Gaussian Mixture Models
  - Fuzzy c-means
- Merge Cluster
  - Merge Cluster
- 3D Volume
  - Volume Projection in 3D

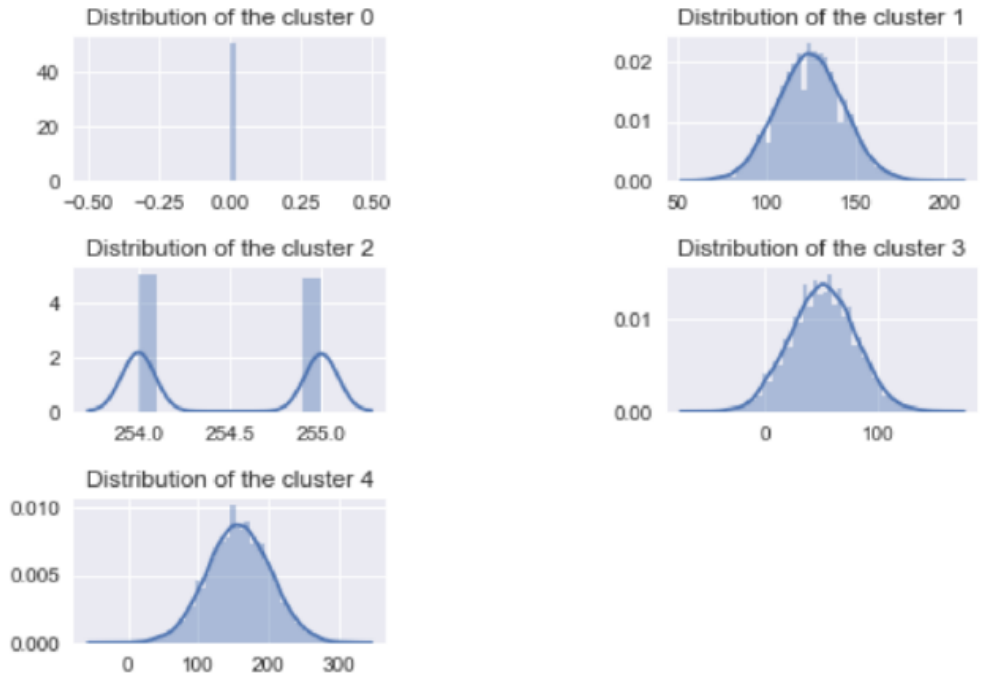
Post-processing

Image Visualisation | Summary

Add Transforms | 1 GMM | Apply

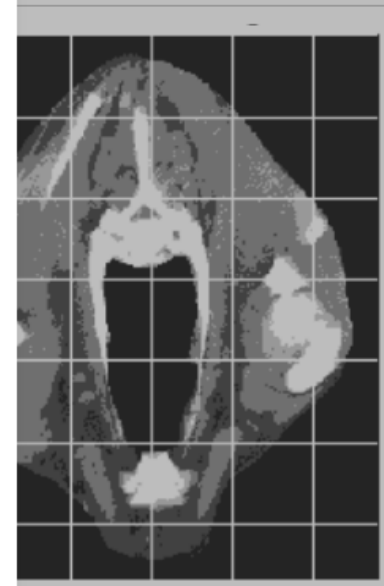
GMM | Histogram

Clusters



Close

GMM Clustered Image



IM000001.jpg

View all clusters

AutoStat

Project Name

Process

Alan Munger

Image Visualisation | Summary

Add Transforms

1 Gaussian Mixture Models — 2 Merge Cluster — 3 Volume Projection in 3D

Apply

Volume Projection in 3D

Auto Component Numbers: 4

Manual Merge Components: IM000001, IM000003...

Uploaded Datasets

Pre-processing

Transformation

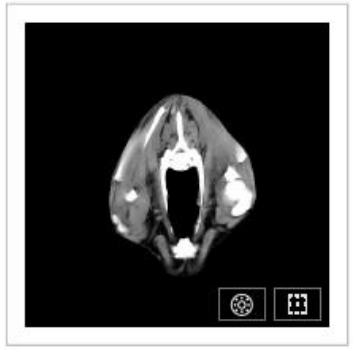
Feature Extraction

Segmentation

- Hard clustering
  - k-means
- Soft clustering
  - Gaussian Mixture Models
  - Fuzzy c-means
- Merge Cluster
  - Merge Cluster
- 3D Volume
  - Volume Projection in 3D

Post-processing

← Back to image library



IM000001.jpg

IM000003.jpg

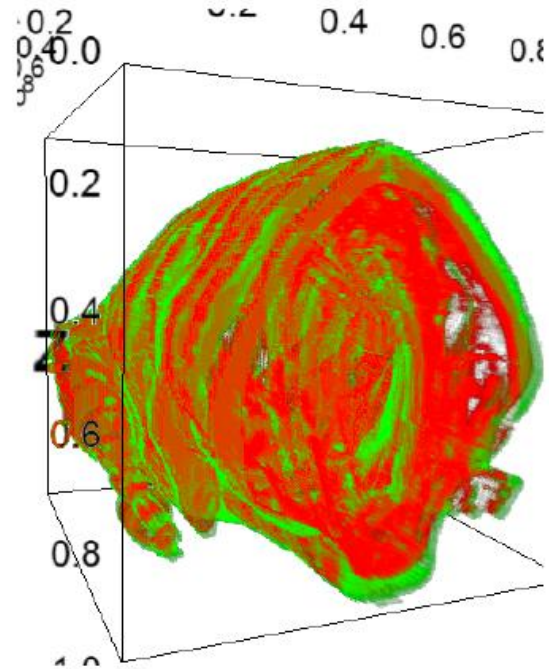
IM000005.jpg

IM000007.jpg

All Masks

Levels: 1.00

Opacities: 0.05



Case Study 1b: Scaling up Bayesian mixture models

- Computing approaches:
  - Graphics processing units
  - Parallel programming
- Algorithmic approaches:
  - Variational Bayes
  - Approximate Bayesian computation
- Sampling based approaches:
  - Huang & Gelman (2005) - partition the data at random and perform MCMC independently on each subset to draw samples from the posterior given the data subset, using methods based on normal approximation and importance re-sampling to make consensus posteriors.
  - Scott et al. (2016) - use similar approach with a different rule for combining posterior draws.
  - Manolopoulou et al. (2010) - improve inference about the parameters of the component of interest in the mixture model by analysing an initial sub-sample to guide selection from targeted components in a sequential manner using Sequential Monte Carlo sampling. (Need a good initial random sample.)

These sampling approaches can be problematic in a massive dataset: a low probability component of interest is likely to escape the initial random sample, which will lead to unreliable inference.

Stratified sampling approach

In satellite imagery, most of the data are replications.
Eg: all water pixels should appear similar while pixels from the land covered with the same crop should produce similar observations.

Thus, inference based on a stratified random sample of the data should be representative of the whole image.

- Use k-means to first label the data.
- Use these labels to obtain a stratified random sample (hence enforcing representation from each sub-population).
- Stratify using nonparametric mixture models.

✓Makes use of the strengths of two clustering methods: the computationally less demanding method of k-means clustering and the more sophisticated DPGMMs, which not only account for correlations between variables, but also learn K in a data-driven fashion.

Proposed approach

- Obtain multiple samples from data using stratified random sampling to enforce adequate representation in each sample from sub-populations that may exist in data.

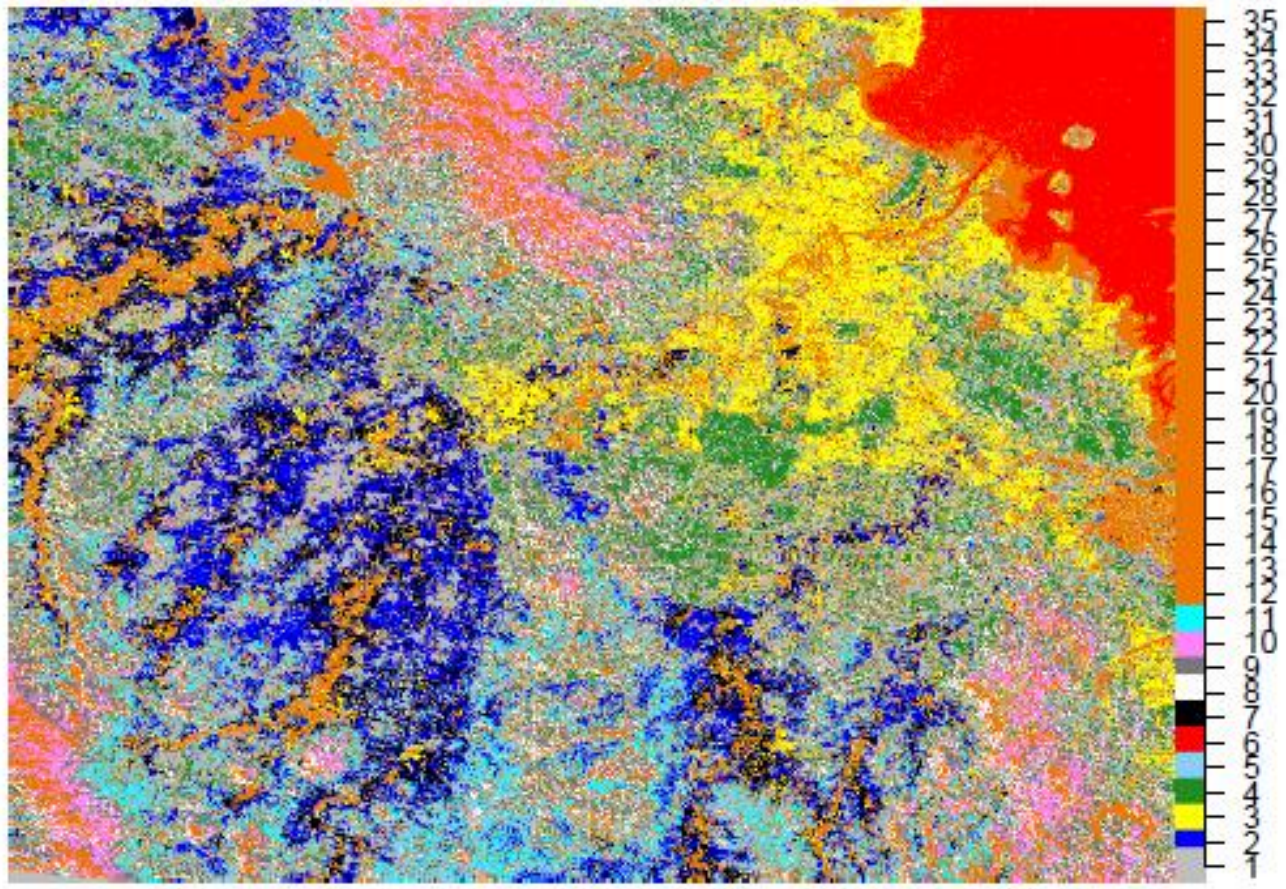
- Fit the mixture model to each sample dataset independently to obtain posterior estimates.

- Obtain label correspondence across multiple estimates:
  - find multivariate component densities of a chosen reference partition
  - pool multiple posterior estimates to form a consensus posterior inference.

- Infer labels for pixels in the entire image using the conditional posterior distribution given pooled estimates - substantially reducing the computational time and memory requirement.

Results

Cluster analysis of satellite image in 2003

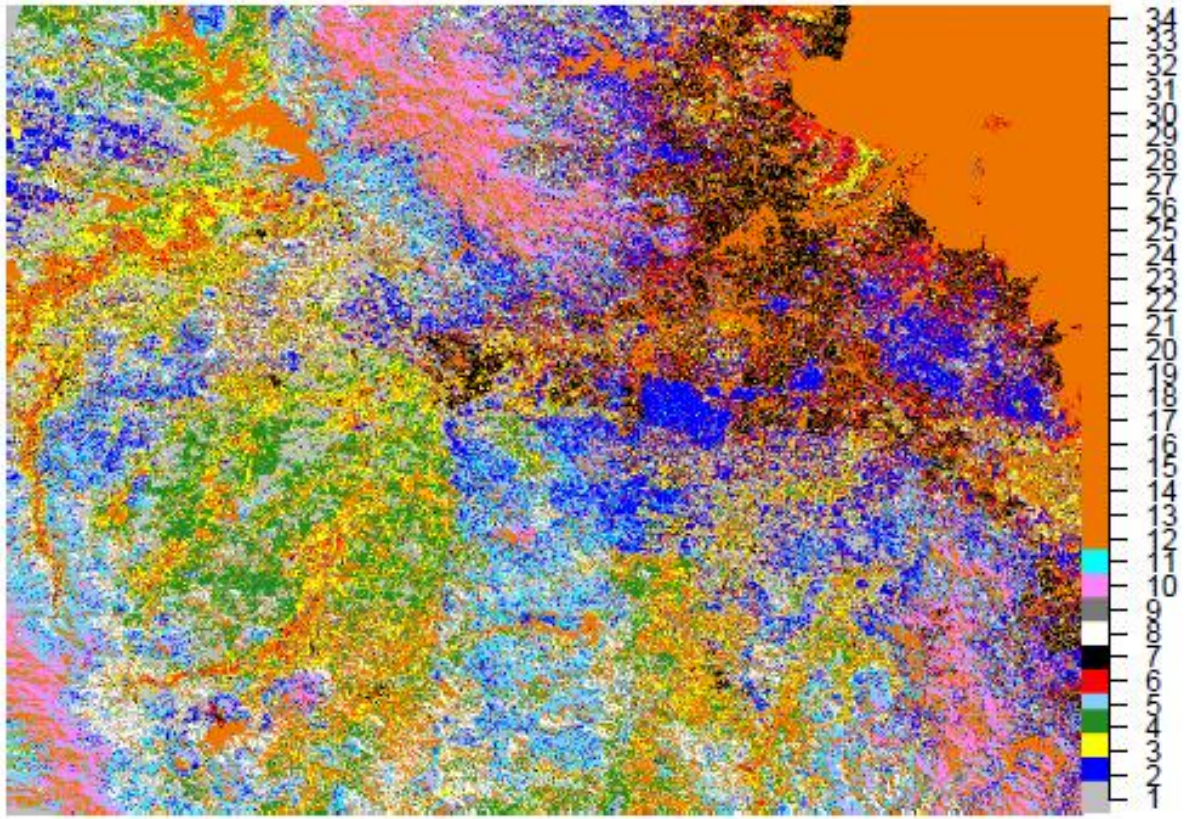


% fireants detected in each cluster

C.No	C.Size	2001	2002	2003	2004	20
1	17.48	13.9	11.2	16.2	16.7	2(
2	9.77	1.1	1.9	3.3	10.7	3
3	8.75	53.1	47.7	50.8	23.4	5:
4	8.06	1.0	1.9	1.1	1.0	(
5	6.02	1.7	1.3	1.4	5.7	1
6	5.58	0.0	0.0	0.0	0.0	(
7	5.49	1.3	2.1	2.9	27.1	(
8	5.37	0.0	0.3	0.5	0.0	(
9	4.65	0.1	0.0	0.3	0.3	(
10	4.16	0.0	0.0	0.0	0.8	(
11	4.04	0.3	0.6	1.0	2.7	(
12	3.78	0.0	0.0	0.0	0.0	(

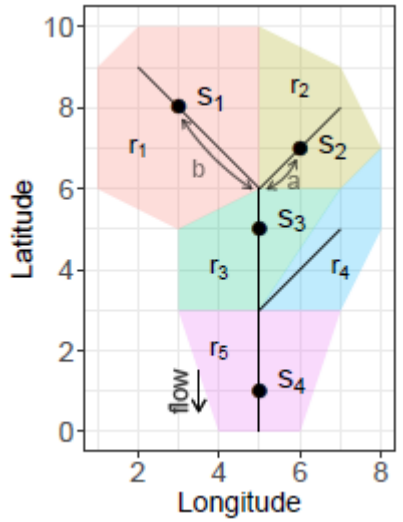
Results

Cluster analysis of satellite image in 2011



C.No	C.Size	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
1	12.05	2.5	3.9	6.0	14.8	6.9	8.9	3.6	3.8	2.4	14.7	3.2	1.8	4.5
2	9.59	1.4	1.6	2.4	4.5	3.1	1.4	0.0	0.2	0.0	0.1	0.1	0.3	0.5
3	7.93	9.0	10.2	9.0	30.1	21.3	7.4	27.9	24.2	40.1	24.2	49.2	29.8	30.9
4	7.40	0.2	0.4	0.1	3.8	0.0	0.0	4.7	0.5	4.1	27.7	11.3	0.9	3.9
5	6.01	0.1	0.3	0.8	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.8
6	5.28	10.3	9.8	11.0	4.8	7.7	7.9	1.9	5.6	7.0	2.3	2.0	1.2	2.5
7	5.27	27.6	27.4	28.1	12.7	13.9	23.4	13.6	16.6	14.2	5.0	10.1	23.6	20.5
8	4.85	0.2	0.3	0.1	0.8	0.0	2.1	1.6	0.4	0.2	0.5	0.5	0.6	0.7
9	4.79	6.7	7.4	9.3	7.2	15.3	23.1	3.2	4.3	13.5	5.3	5.1	6.3	6.3
10	4.72	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	4.36	0.1	0.1	0.1	0.3	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.1	0.1
12	4.10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	3.65	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	3.51	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	3.38	31.6	29.6	21.0	10.5	14.3	18.7	27.2	31.5	6.2	3.1	6.5	9.2	9.4
16	3.12	0.2	0.0	0.4	1.3	1.5	0.0	0.7	0.0	0.2	1.6	0.9	0.1	0.6
17	2.81	0.1	1.1	0.2	4.5	0.0	0.0	4.5	1.5	3.8	11.1	2.2	3.4	1.5
18	1.66	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	1.43	8.7	5.9	9.6	1.8	7.7	5.7	1.9	1.4	3.2	0.9	2.2	2.2	1.7
20	1.25	0.1	0.1	0.2	0.0	0.8	0.8	0.4	0.6	1.5	2.3	0.9	0.2	0.6
21	0.99	0.3	0.4	0.1	0.8	0.0	0.0	0.0	0.7	0.5	0.1	0.5	0.7	0.9
22	0.48	0.2	0.0	0.1	0.0	0.0	0.0	0.0	1.0	0.0	0.1	0.0	0.0	0.0
23	0.44	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
24	0.27	0.2	0.6	0.4	1.0	3.8	0.0	6.6	2.4	2.4	0.4	3.5	10.9	9.9
25	0.24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26	0.24	0.2	0.6	0.3	0.0	3.7	0.5	2.2	4.8	0.9	0.2	1.6	8.5	4.6
27	0.12	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
28	0.04	0.5	0.4	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
29	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.2	0.1	0.0	0.1
30	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
31	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
32	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
33	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
34	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total incursions		1788	701	928	387	130	365	547	965	664	5690	2866	1272	1414

Case Study 2a. Accounting for different spatial neighbourhoods



Covariance structures:

- Tail up
  - Covariance depends on upstream points

$$C_{TU}(s_i, s_j) = C_u W_{ij} \text{ if flow-connected; 0 otherwise}$$

W_{ij} : spatial weights defined by the branching structure of the network

Choice of unweighted tail-up covariance functions C_u, e.g., exponential, sill, spherical

- Tail down
  - Covariance depends on both upstream and downstream points

- Mixture covariance

Case Study 2a. Spatio-temporal models

Three main approaches:

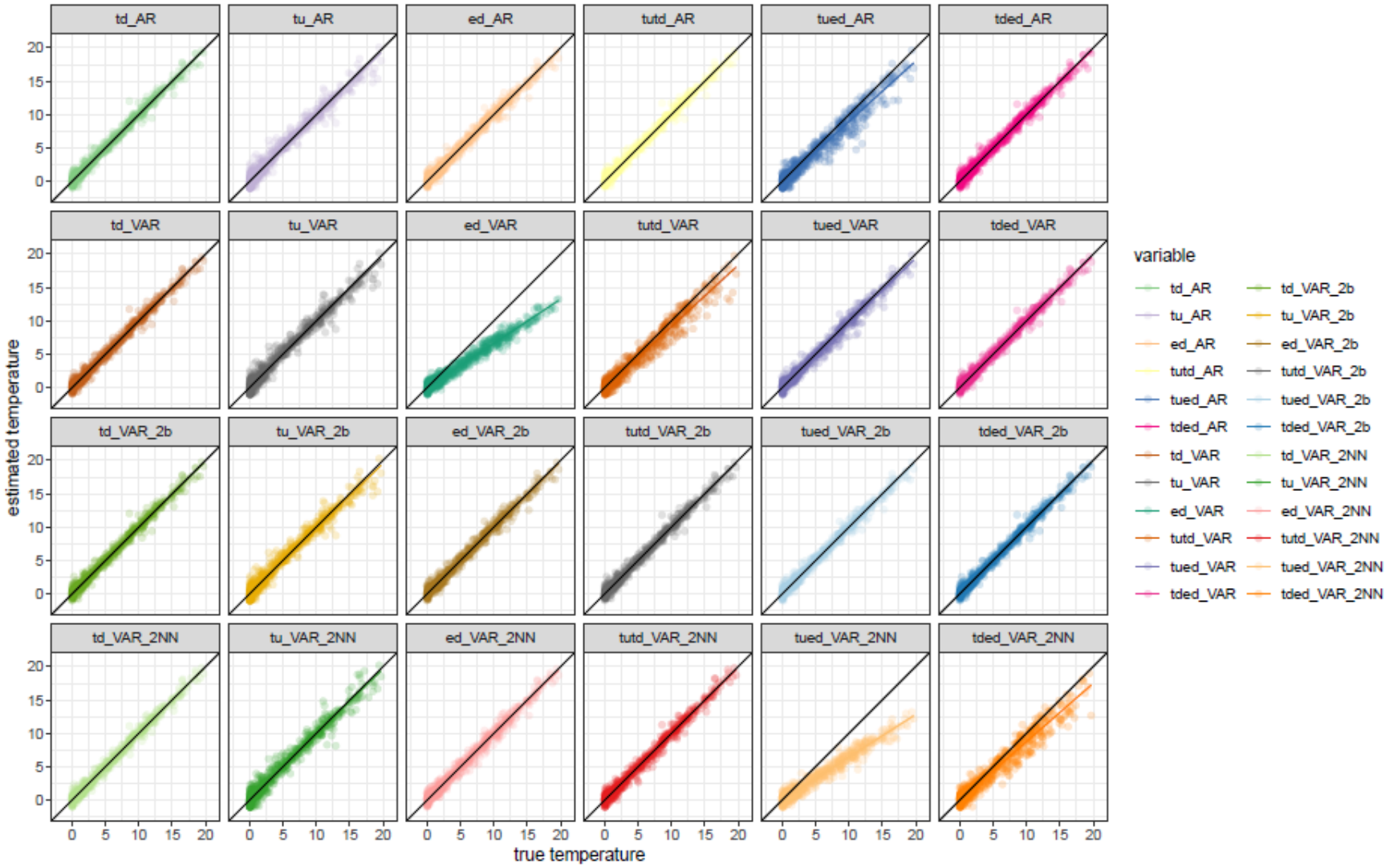
1. Model with the full space-time covariance function: full information, computationally intensive
2. Separable space-time model: lose interactions, computationally easier
3. Dynamical model that evolves the spatial process: can be shown to be equivalent to (2) and computationally more efficient in Bayesian context

$$\{Y_{(s,t)}\} = \prod_{t=1}^{T} [Y_{(s,t)} | Y_{(s,t-1)}]$$

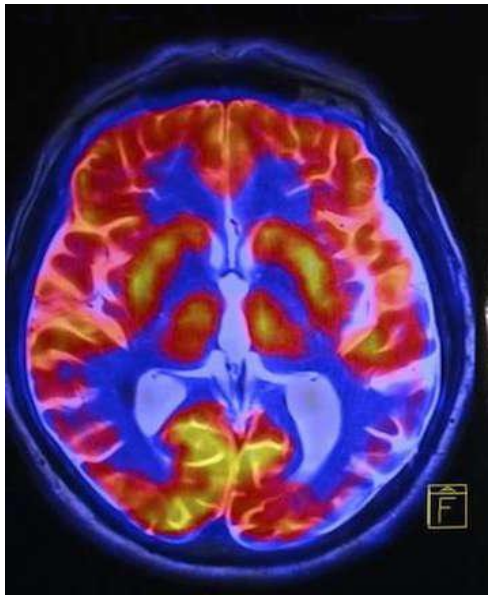
$$[Y_{(s,t)} | Y_{(s,t-1)}] = \mathcal{N}(\mu_{(s,t:t-1)}, \Sigma_{(s)} + \sigma^2 I)$$

$$\mu_{(s,t:t-1)} = X'_{(s,t)}\beta + \Phi_1(Y_{(s,t-1)} - X'_{(s,t-1)}\beta)$$

Indicative Results



Case Study 2b. Estimating the spatial neighbourhood



Y_{irk} : cortical thickness of region $k = 1{:}K$
for participant $i = 1{:}I$
who has $r = 1{:}R_i$ replicates

$$y_{irk}|b_{ik}, \beta, \sigma^2 \sim N(x_i\beta + b_{ik}, \sigma^2)$$

$$\boldsymbol{b}_i|\sigma_s^2, W \sim MVN(\mathbf{0}, \sigma_s^2 \mathrm{Q})$$

$$Q^{-1} = \rho(D_w - W) + (1-\rho)I$$

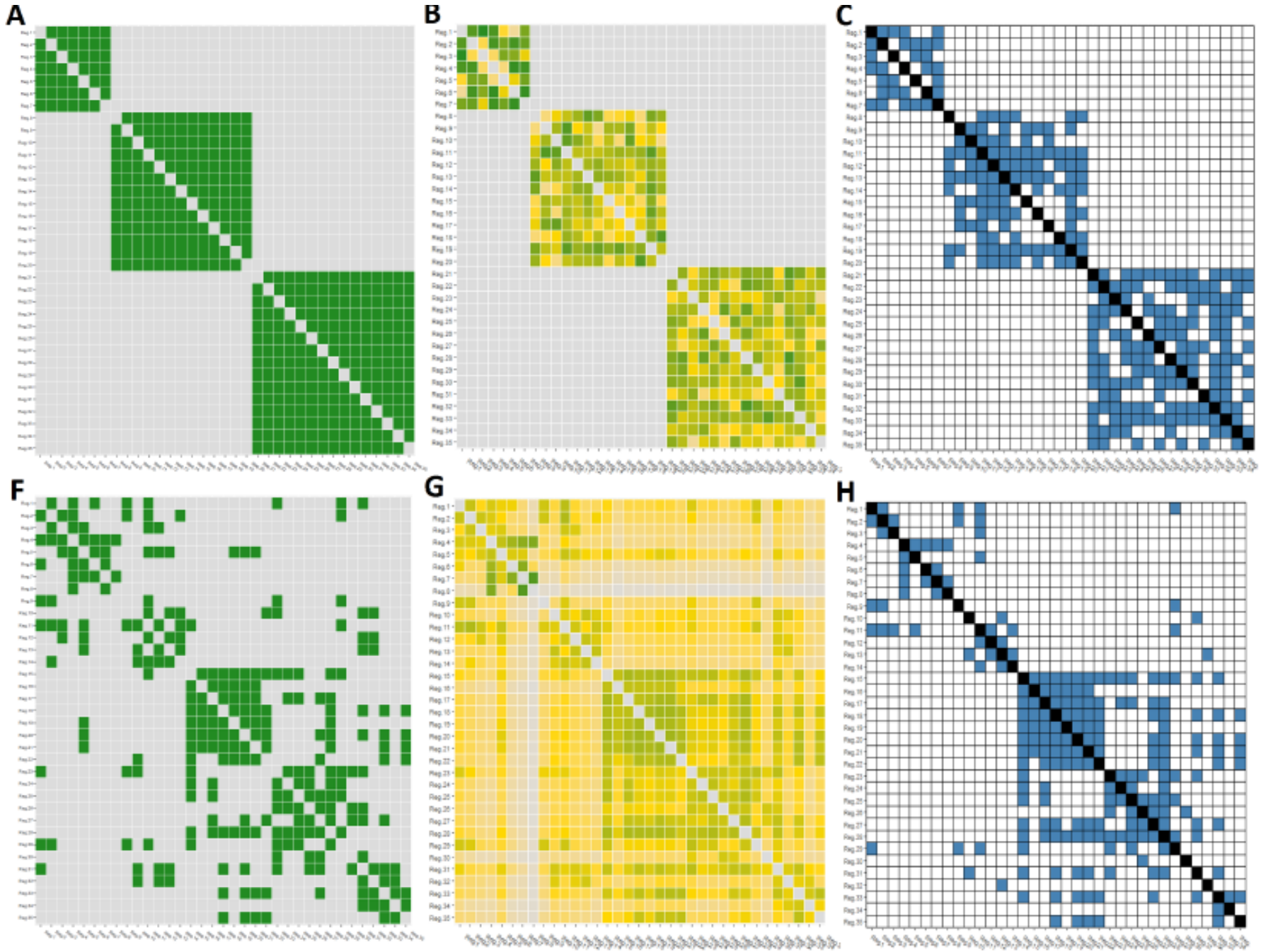
D : diagonal matrix with elements given by row sums (or number of neighbours) $\sum_{j=1}^{K} w_{jk}$
W : zero-diagonal, binary symmetric matrix, $w_{jk} = 1$ if regions j and k are neighbours, else = 0
ρ : determines global level of spatial correlation

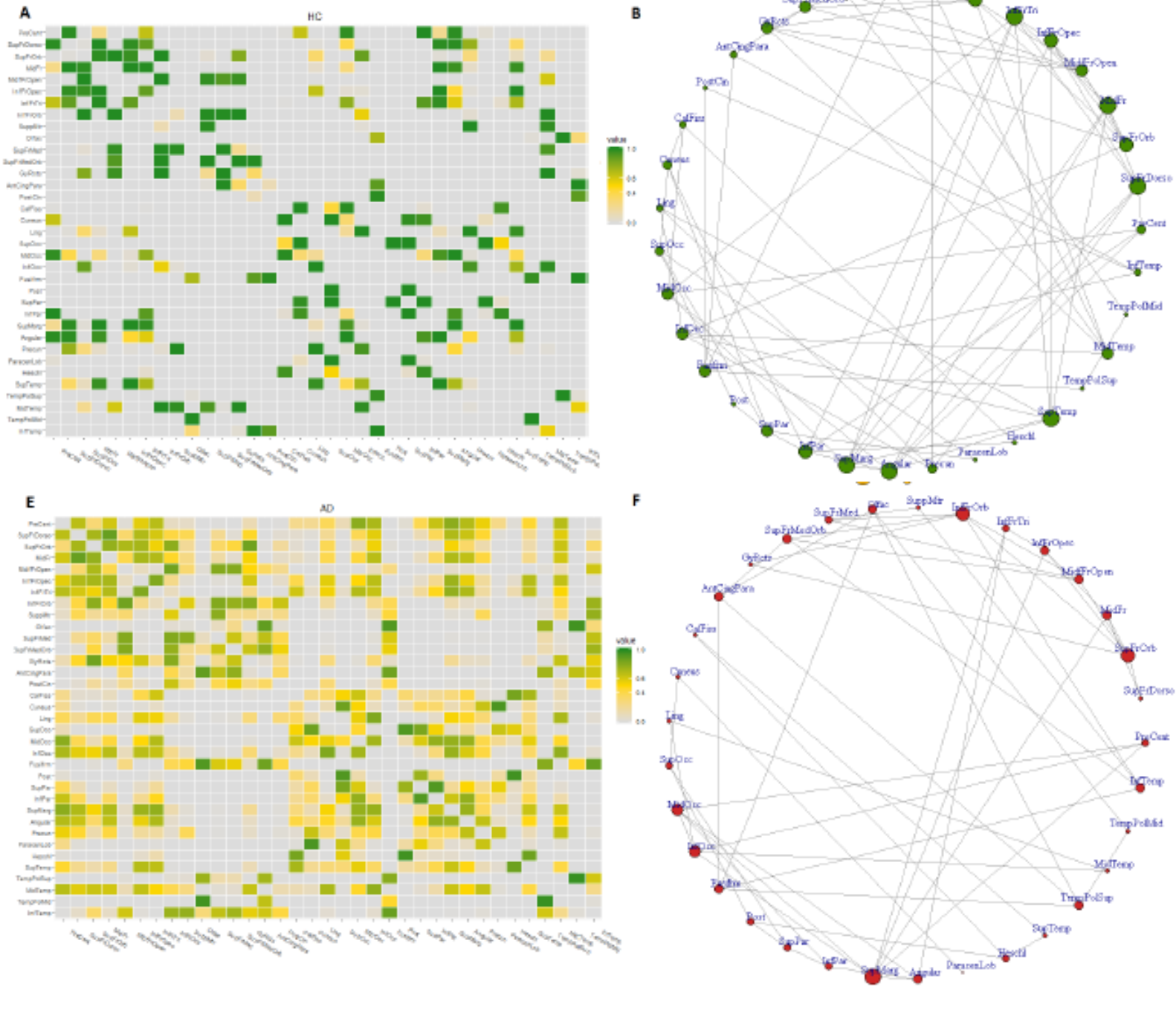
Estimating the neighbourhood

$$p(W, \sigma^2, \sigma_s^2, \beta | \mathbf{y}, X) \propto \left[\prod_{i=1}^{I} \prod_{r=1}^{R_i} \prod_{k=1}^{K} p(y_{irk} | b_{ik}, \sigma^2, \beta, \mathbf{x}_i) \right] \left[\prod_{i=1}^{I} p(\mathbf{b}_i | \sigma_s^2, W) \right] p(\beta) p(\sigma^2) p(\sigma_s^2) p(W)$$

$$p(W | \sigma_s^2, \mathbf{b}) \propto \left[\prod_{i=1}^{I} p(\mathbf{b}_i | \sigma_s^2, W) \right] p(W)$$

Computation: MCMC

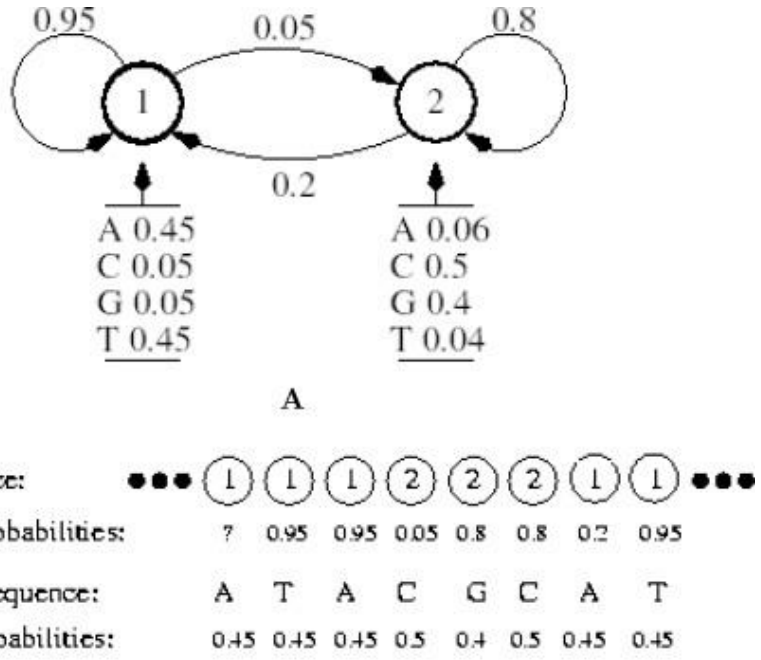

A
B
C
F
G
H


A
HC
value
B
E
AD
value
F

Case Study 3: Finite Space Hidden Markov Model (HMM)

HMMs:

- arise when observations from a mixture of distributions depend on an unobserved (hidden) Markov chain
- provide a framework for identifying and modelling homogeneous sub-sequences in data which display heterogeneity
- applications in DNA segmentation, economic analyses, conservation, sport, etc



Choo et al., 2004

HMM setup

Observed time series $Y_t = \{y_1, \dots, y_n\}$ depends on a single realisation of the underlying stochastic process determined by the unobserved states $X_t = \{x_1, \dots, x_n\}$

$$\forall t \leq n: [Y_t | X_t = x] \sim g_{\gamma_x}, \qquad \gamma \epsilon \Gamma \subset \mathbb{R}^d, \qquad X_t \in \boldsymbol{X} = \{1, \dots, K\}$$

$(X_t)_{t \geq 1}$ is a Markov chain with K states and transition matrix $Q = (q_{i,j})_{1 \leq i,j \leq K}$

The Markov chain is also associated with a stationary distribution which contains the long term state probabilities μ_Q satisfying $\mu_Q Q = \mu_Q$

Examples

Sim 1: $K^* = 3$ states, one well separated, two overlap

Sim 2: $K^* = 3$ states, different state means and transition probabilities

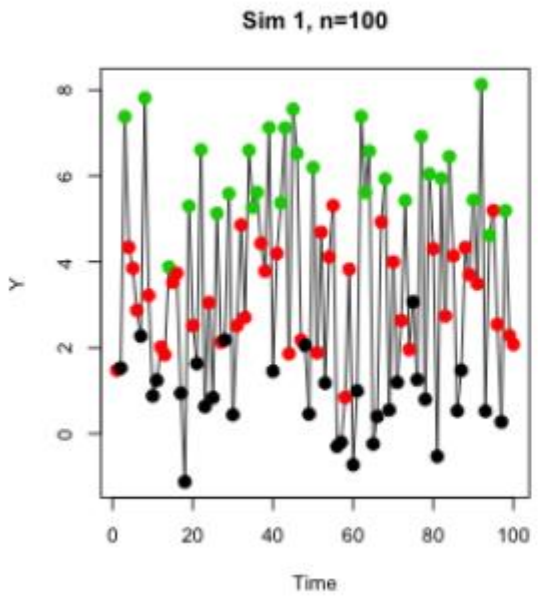
Sim 3: $K^* = 5$ states, well separated, equally spaced means, mainly large values on the diagonal of Q^* (sticky)
All have known state specific variances equal to 1.

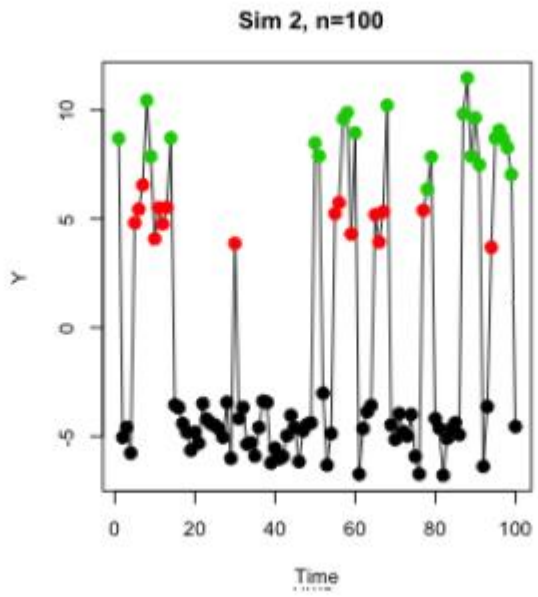
$$\textbf{Sim 1}\ \gamma^*_{S1} = (1, 3, 6),\ \mu^*_{S1} = (0.33, 0.38, 0.29),\ \text{and}\ Q^*_{S1} = \begin{bmatrix} 0.2 & 0.3 & 0.5 \\ 0.5 & 0.25 & 0.25 \\ 0.25 & 0.65 & 0.1 \end{bmatrix}$$

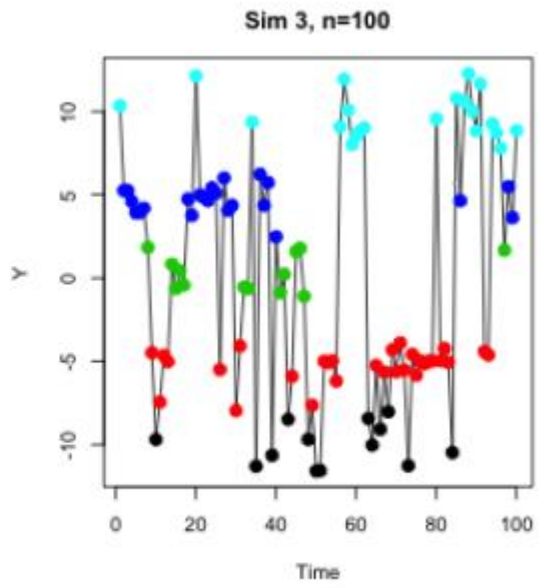
$$\textbf{Sim 2}\ \gamma^*_{S2} = (-5, 5, 9),\ \mu^*_{S2} = (0.56, 0.18, 0.26),\ \text{and}\ Q^*_{S2} = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.4 & 0.4 \\ 0.3 & 0.2 & 0.5 \end{bmatrix},$$

$$\textbf{Sim 3}\ \gamma^*_{S3} = (-10, -5, 0, 5, 10),\ \mu^*_{S3} = (0.11, 0.24, 0.20, 0.22, 0.22),$$
$$\text{and}\ Q^*_{S3} = \begin{bmatrix} 0.2 & 0.3 & 0.1 & 0.2 & 0.2 \\ 0.1 & 0.6 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.6 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.6 \end{bmatrix}.$$

Examples - realisations







Estimation: no. states known

Estimation of Q and $(X_t)_{t\geq 1}$ is straightforward (Fruhwirth-Schnatter, 2008; Nagaraja, 2006; Scott, 2002).

In a Bayesian context, MCMC estimation is particularly straightforward with conjugate priors and a data augmentation approach (Chib, 1996; Taylor et al., 2012):

- Update $(Q|X_t, Y_t)$
- Update $(\gamma_i|X_t, Y_t)$
- Estimate posterior allocations $(((X_t)_{t\geq 1}|\gamma_i, i \leq K, Q, (Y_t)_{t\geq 1})$

Estimation: no. states unknown

Non-identifiability when more states are included in a HMM than are supported by the observations

Frequentist setting: order estimation is difficult:

- likelihood ratio statistic is unbounded even for simple case of comparing models with $K = 1$ and $K = 2$ states (Gassiat & Kerribin, 2000)
- Solution based on implementing heavy penalties in a maximum likelihood setting (Gassiat & Boucheron, 2003)
- Penalised marginal pseudo-likelihood to obtain weakly consistent estimators for K (Gassiat, 2002)

Bayesian setting:

- reversible jump Markov chain Monte Carlo (Richarson & Green, 1997; Boys & Henderson, 2004)
- variational Bayes methods (McGrory & Titterington, 2009)
- sequential inference methods (Chopin, 2001)
- Bayes factors (Han & Carlin, 2001; Friel & Pettitt, 2008)
- nonparametric methods (Beal et al., 2002; Ding & Ou, 2010; Teh et al., 2006; Fox et al., 2008)

Overfitting HMMs

Some understanding for finite mixtures:

Asymptotically, the prior on the mixture weights determines the posterior behaviour of extra groups (Rousseau & M, 2011) and can induce posterior emptying of superfluous components (van Havre et al., 2015).

- If $\max(\alpha_j) < d/2$, where $d = \dim(\theta)$, then asymptotically $f(\theta|Y)$ concentrates on the subset of parameters for which $f_\theta = f_0$, so $K - K_0$ components have weight 0.
- If $\min(\alpha_j) > d/2$, then 2 or more components will tend to merge with non-neglectable weights each. This will lead to less stable behaviour.
- If $\min(\alpha_j) \leq d/2 \leq \max(\alpha_j)$, then the situation varies depending on the α's and on the difference between K and K_0

Overfitting HMMs

Very limited understanding of asymptotic posterior convergence for overfitted HMMs

- Gassiat & Rousseau (2012): for parametric, finite space HMMs, the dependence between the states leads the neighbourhoods of the true parameter values to contain transition matrices that lead to non-ergodic Markov chains, corresponding to areas of poor Markov behaviour

- no theoretical results for posterior emptying.

Indications of posterior emptying of overfitted HMMs

- Variational Bayes methods: depend on this through ‘state-removal’ phenomena, but no underlying theory (McGrory & Titterington, 2009)
- Particle filter model: considers only states which appear in the posterior sequence of hidden states, like ignoring emptying groups (Chopin et al., 2001)

Example: DNA segmentation (Boys & Henderson, 2000, 2004; Nur et al., 2009)

- Use common prior for *i*th row of $Q \sim D(\alpha_{i1}, \ldots, \alpha_{iK})$
- Assume we are unlikely to detect short segments, except if searching for a state with known parameters or if many short segments from a particular state
- Instead of the standard choice $\alpha_{ij} = \alpha \; \forall i, j \leq K$, consider alternative prior: $Q_{i.} \sim D(\alpha, \alpha, \ldots, d, \alpha, \ldots, \alpha)$, s.t. d is the ith element and is larger than the exchangeable off-diagonal elements, so $\mathrm{E}(Q_{k,k}) \to 1$
- **Effective**, but little supporting theory

Simple example

True model: standard normal distribution, equivalent to a HMM with single state *K = 1*. For *t = 1,…,1000:*

$$y_t \sim \text{Normal}(0, 1)$$

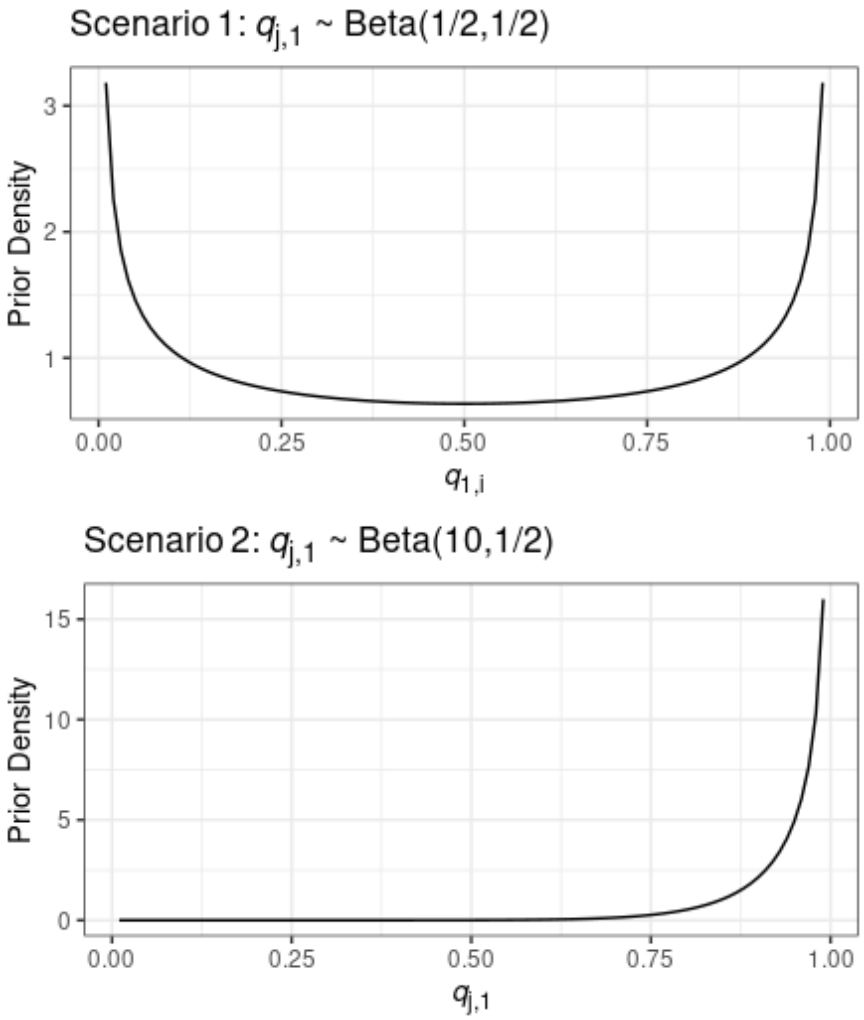
Overfitting model: a HMM with *K = 2* states and normal emission distributions:

$$y_t | X_t = j \sim \text{Normal}(\mu_j, 1)$$

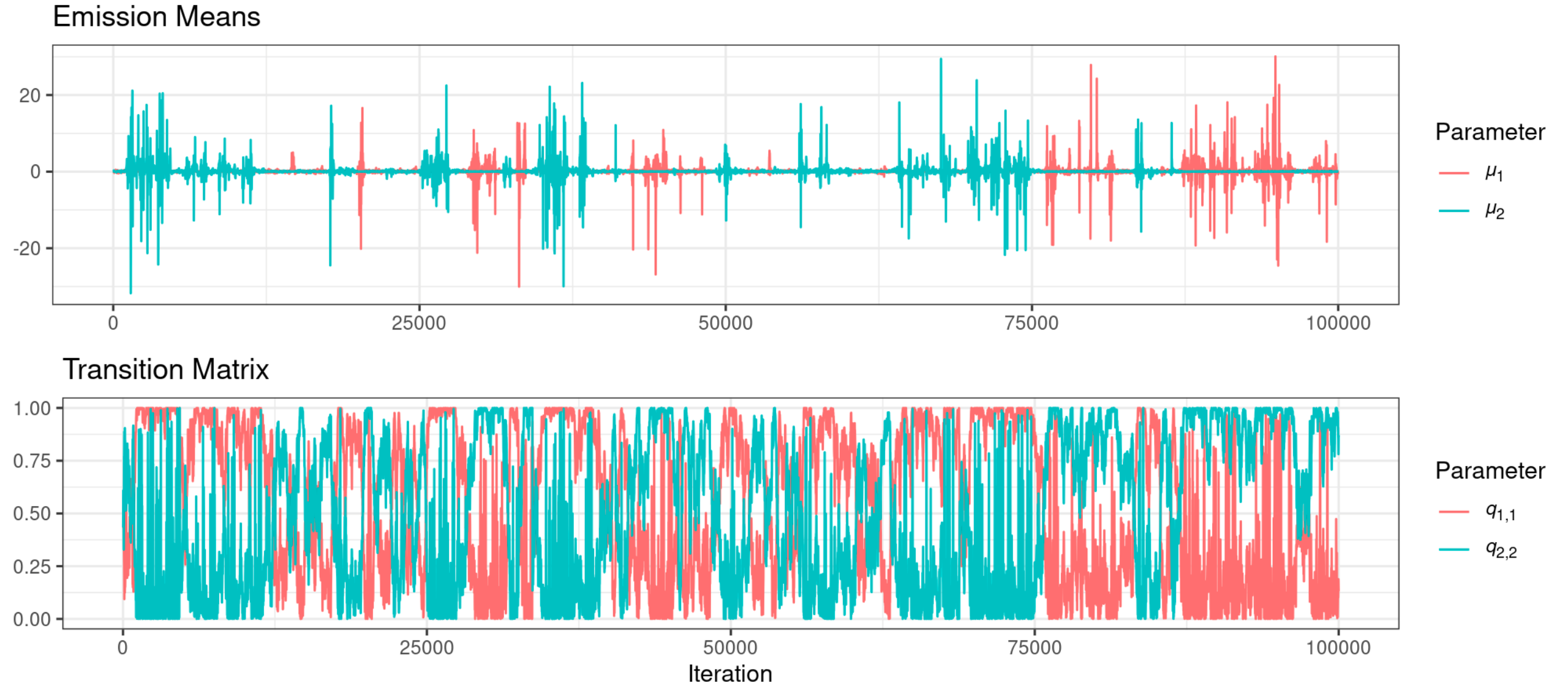
Priors:

$$\mu_1, \mu_2 \sim \text{Normal}\left(\bar{y}, 10^2\right)$$

$$q_{1,1}, q_{2,1} \sim \text{Beta}(\alpha_1, \alpha_2)$$



Scenario 1


Emission Means
20
0
-20
0
25000
50000
75000
100000
Parameter
μ_1
μ_2
Transition Matrix
1.00
0.75
0.50
0.25
0.00
0
25000
50000
75000
100000
Iteration
Parameter
$q_{1,1}$
$q_{2,2}$

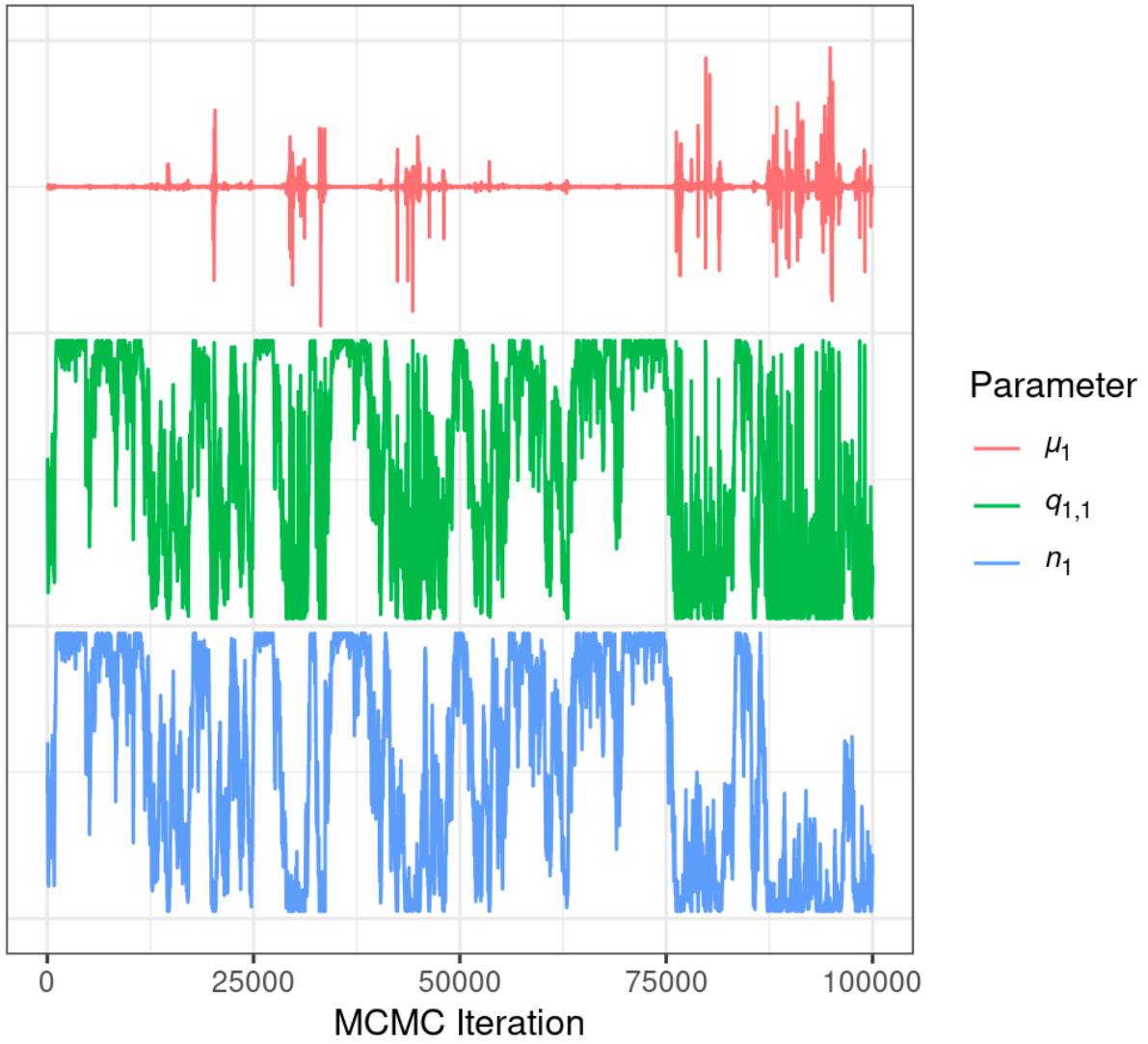
Scenario 1

n_1: Number of observations in state 1.

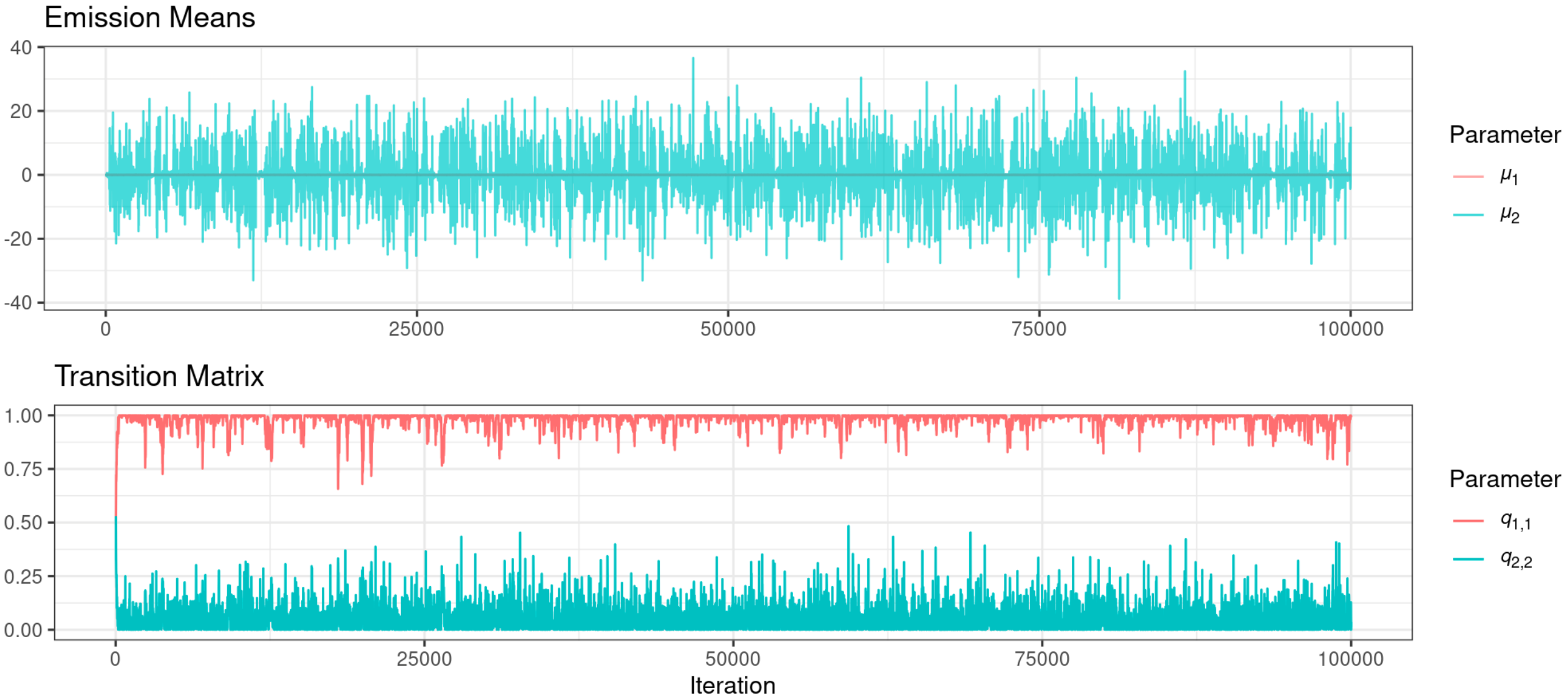
MCMC chain shows very slow mixing in transition matrix parameters.

“Sparks” in emission mean parameters occur when the state empties. i.e. a draw from the prior.

Similar behaviour seen in $q_{1,1}$ and $q_{2,2}$.



Scenario 2


Emission Means
40
20
0
-20
-40
0
25000
50000
75000
100000
Parameter
μ_1
μ_2
Transition Matrix
1.00
0.75
0.50
0.25
0.00
0
25000
50000
75000
100000
Iteration
Parameter
$q_{1,1}$
$q_{2,2}$

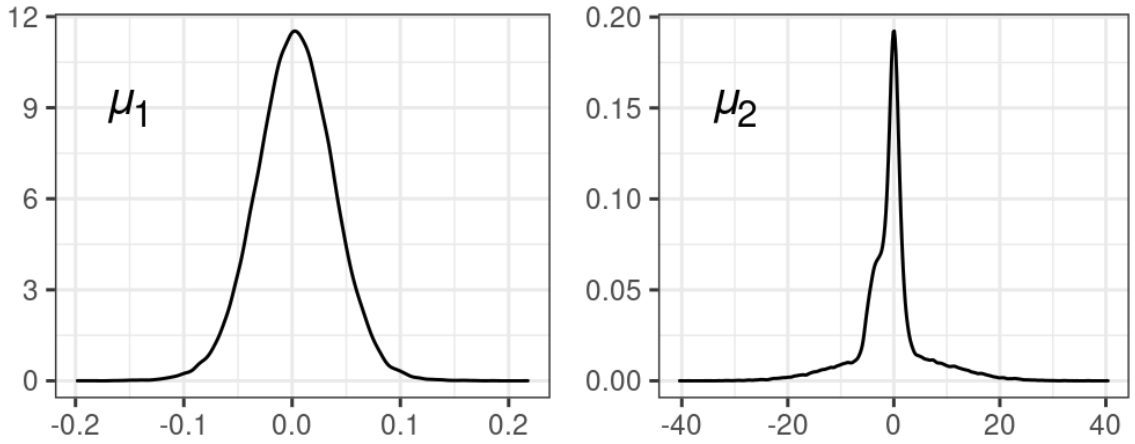
Scenario 2

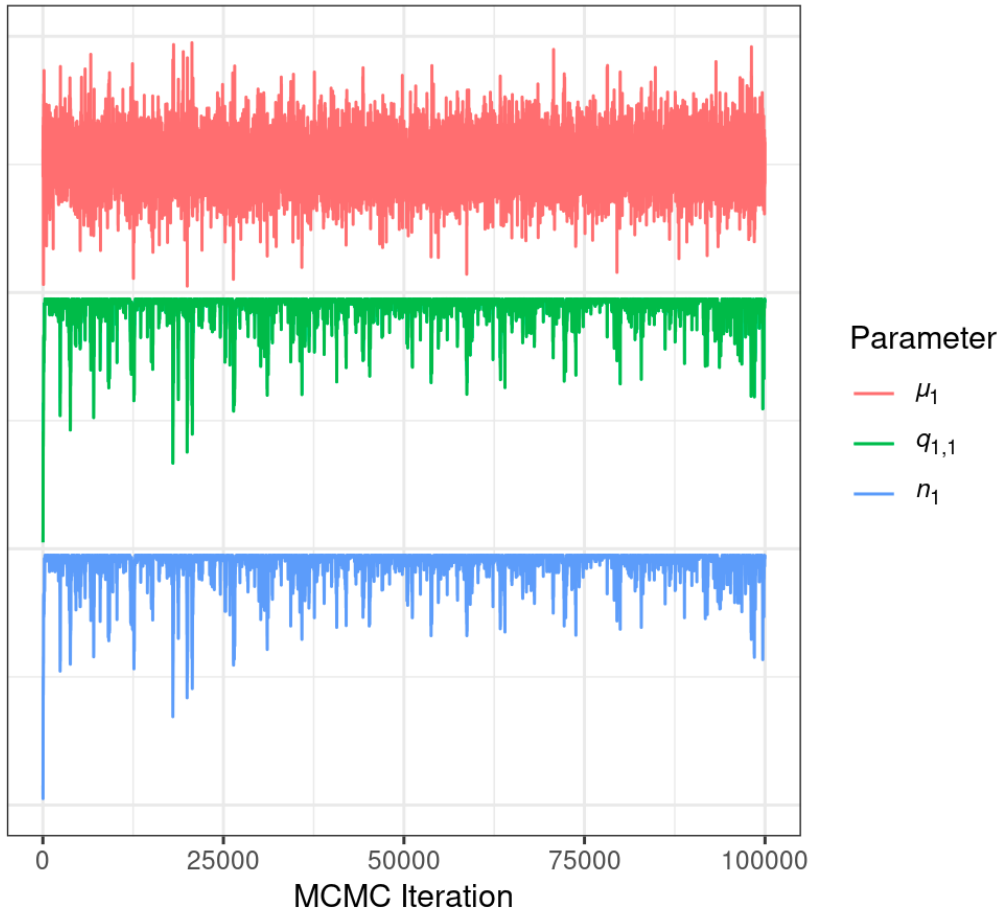
n_1: Number of observations in state 1.

MCMC chain shows better mixing than Scenario 1.

Data are mostly emptied from state 2.

Posterior densities show approximate normality for μ_1 but not μ_2.





General problem setup

- Set μ as a prior initial distribution for the hidden Markov chain $X_t = \{x_1, \dots, x_n\}$
- Conditional distribution $[y_t | x_t = j] \sim G_{\gamma_j}$, absolutely continuous w.r.t. fixed measure γ with density g_{γ_j}
- Set $\theta = (Q, \gamma_1, \dots, \gamma_k) \in \mathcal{Q}_K \times \Gamma^K = \Theta_K$ where

$$\mathcal{Q}_K = \left\{ Q = \left(q_{i,j}\right)_{i,j \leq K}; \sum_{i=1}^{K} q_{i,j} = 1, q_{i,j} \geq 0 \ \forall i, j \right\}$$

- Denote by μ_Q the stationary distribution associated to Q, i.e. the probability distribution on $\{1, \dots, K\}$ satisfying $\mu_Q = \mu_Q Q$.

General problem setup

- Recall for any Markov chain on a finite state-space with transition probability Q and stationary distribution μ_Q (one of them if Q admits more than one stationary distribution), it is possible to define $\rho_Q \geq 1$ s.t., for any m and any $i \leq K$

$$\sum_{j=1}^{k} |(Q^m)_{ij} - \mu_Q(j)| \leq \rho_Q^{-m}, \quad \rho_Q = \left(1 - \sum_{j=1}^{K} \min_{1 \leq i \leq K} q_{i,j}\right)^{-1}$$

- The complete likelihood conditional on $X_1 = x_1$ is

$$f_n(Y_t, X_t|\theta) = g_{\gamma_{X_1}}(y_1) \prod_{i=1}^{n-1} q_{x_i,x_{i+1}} g_{\gamma_{x_{i+1}}}(y_{i+1})$$

- The likelihood of Y_t conditional on $X_1 = x_1$, defining $x_{2:n} = (x_2, \ldots, x_n)$, is

$$f_n(Y_t|\theta, X_1) = \sum_{x_{2:n}} f_n(Y_t, X_t|\theta).$$

- We also write

$$f_n(Y_t|\theta, \mu) = \sum_{x_1=1}^{K} f_n(Y_t|\theta, X_1)\mu(X_1)$$

and

$$\ell_n(\theta, X_1) = \log f_n(Y_t|\theta, X_1) \text{ and } \ell_n(\theta, \mu) = \log f_n(Y_t|\theta, \mu)$$

Priors

Study the behaviour of posterior distributions associated with priors belonging to the following family:

(C1) Prior on Q: the rows Q^i are iid Dirichlet $D(\alpha_1, \dots, \alpha_K), \alpha_j > 0, j \leq K$

(C2) Independent prior on the γ's: $\gamma_j \overset{\text{iid}}{\longrightarrow} \sim \pi_\gamma$ with positive and continuous density on Γ.

Denote by π the relevant prior and $\pi(.|Y)$ the corresponding posterior distribution, so that

$$\pi(d\theta|Y) = \frac{f_n(Y|\theta,\mu)\pi(d\theta)}{\int_\Theta f_n(Y|\theta,\mu)\pi(d\theta)}$$

Denote by $F(h) = \int h(x)dF(x)$ for every probability measure F and integrable function h.

Asymptotic analysis

What is the asymptotic behaviour of $\pi(.|Y)$ under the different priors, when the true parameter corresponds to a HMM with $K_0 < K$ hidden states?

- In the K_0- parameter space $\Theta_{K_0}, \theta^* = \left(Q^*, \gamma_1^*, \dots \gamma_{K_0}^*\right)$
- The true model can be parametrized by infinitely many parameters in the K-parameter space Θ_K.
- In particular, any parameter of the form $\left(\gamma_1^*, \dots \gamma_{K_0}^*, \dots, \gamma_{K_0}^*\right)$ and Q with $q_{i,j} = q_{i,j}^*$ if $i \leq k^*, j \leq k^* - 1$, $\sum_{j=K_0}^{K} q_{i,j} = q_{i,K_0}$ and $q_{i_{,.}} = q_{K_0} \forall i \geq K_0 + 1$ leads to the same likelihood function $f_n(Y|\theta^*, \mu)$, for all μ.
- The parameters $\theta \, \epsilon \, \Theta_K$ defined by $Q = \begin{pmatrix} Q^* & 0 & \dots & 0 \\ R & 0 & \dots & 0 \end{pmatrix}$ where for all $i = K_0 + 1, \dots, K, R_{i,1} = 1$ and $R_{i,j} = 0$ if $j \geq 2$ and $\gamma_j = \gamma_j^*$ for all $j \leq k$ lead to the same likelihood function for all μ having support in $\{1, \dots, K_0\}$.

- Denote by $\Theta^* \subset \Theta_K$ the set of all θ s.t. either $f_n\left(Y\middle|\theta, \mu_Q\right) = f_n(Y|\theta^*, \mu_{Q^*})$ or $f_n(Y|\theta, \mu) = f_n(Y|\theta^*, \mu)$ for all n.

Asymptotic behaviour of the posterior distribution - Theorem

We want to find some sufficient conditions on the prior to ensure that the posterior distribution concentrates on the configuration where the extra states are emptied out when the number of observations goes to infinity.

Assume that the true model is a HMM on $K_0 < K$ *hidden states with true parameter* $\theta^* \in \Theta_{K_0}$.

Under some regularity assumptions on g_γ, *if there exists* $1 \leq p \leq K_0$ *such that*

$$\alpha_1 = \cdots = \alpha_p = \overline{\alpha} \quad \textit{and} \quad \alpha_{p+1} = \cdots = \alpha_K = \underline{\alpha} \quad \textit{satisfying}$$

$$p\bar{\alpha} + (K-p)\underline{\alpha} > \frac{(K_0(K_0-1+d) + \underline{\alpha}K(K-K_0))(K_0(d+K_0-1) + \underline{\alpha}(K_0+1)(K-K_0-1) + d/2)}{d/2 - \underline{\alpha}[(K-K_0)^2 - (K-2K_0-1)]}$$

$$d/2 > \underline{\alpha}((K-K_0)^2 - (K-2K_0-1))$$

then setting $A_1 = K(K-K_0)\alpha + K_0(K_0-1+d)$ *and* $A = A_1/(p\overline{\alpha} + (K-p)\alpha$, *for any* $M_n \to \infty$

$$\pi(\min_{\sigma \in \mathcal{S}_K} \sum_{j=K_0+1}^{K} p_{\sigma(j)} > M_n v_n | Y) = o_p(1), \quad \textit{and}$$

$$v_n = n^{-1/2[(1-A)B - A_1]/(d/2 + \underline{\alpha}(K-2K_0-1))} (\log n)^{B/(d+2\underline{\alpha}(K-2K_0-1))}$$

with $B = K_0(d+K_0-1) + \underline{\alpha}(K_0+1)(K-K_0-1) + d/2$.

Discussion of Theorem

1. Since K_0 is unknown, $\overline{\alpha}$ and $\underline{\alpha}$ have to be chosen conservatively, eg:
   $K_0 = K - 1$ for the lower bound on $\overline{\alpha}$; $K_0 = 1$ for the upper bound on $\underline{\alpha}$, so

$$d/2 > \underline{\alpha}(K^2 - 3K + 4)$$

$$p\bar{\alpha} + (K - p)\underline{\alpha} > \frac{((K-1)(K-2+d) + \underline{\alpha}K)((K-1)(d+K-2) + d/2)}{d/2 - \underline{\alpha}[K^2 - 3K + 4]}$$

2. Hence we can achieve posterior emptying of extra states in HMMs by binding the posterior distribution of the μ associated with extra states to be small yet remain non-zero, thereby retaining the ergodicity of the estimated Markov chain.
3. This behaviour is possible due to the structure of the prior, which is asymmetric with respect to the hyperparameter values (containing large values $\overline{\alpha}$ and smaller values $\underline{\alpha}$).
4. The prior constraints depend on *d*, the number of free parameters in each state, and the total number of states in the model *K*, but also on *p*, which defines the number of $\overline{\alpha}$ values included in the prior.
5. As $p > K$ must hold, *p* must be set to the smallest reasonable value
   (eg $p = 1$ corresponds to a noninformative setup on the number of components).

Large simulation example - setup

- HMM with Normally distributed state-specific distributions:

$$[Y_t|X_t = j] \sim \mathcal{N}(\gamma_j, 1) \quad X_t \in \mathbf{X} = \{1, \ldots, K\}$$

- Prior on the emission means: $\pi(\gamma) \sim N(\gamma_0 = \bar{Y}, \tau_0 = 100)$.
- Prior on each row of Q: $D(\overline{\alpha}, \underline{\alpha}, \ldots, \underline{\alpha})$.
- Posterior:

$$p(X_t, \gamma, Q|Y_t) \propto p(Y_t|X_t, \gamma, Q)p(X_t|\gamma, Q)\pi(\gamma)\pi(Q)$$

- Gibbs sampler run for M = 20,000 iterations; first 10,000 discarded.

- Sample of 10,000 observations simulated from a univariate Gaussian HMM with $K^* = 2$ states, transition matrix $Q^* = \begin{Bmatrix} 0.6 & 0.4 \\ 0.7 & 0.3 \end{Bmatrix}$, leading to stationary distribution $\mu^* = \{0.64, 0.36\}$.
- The individual states are distributed N($\gamma^* = \{-1, 3\}$, 1) [variance assumed known].

- Model with $K = 4$ states and 9 combinations of hyperparameters: $\overline{a}$ = Theory, K, 1 ; $\underline{a} = \overline{a}$, 0.01, 1/$n$.

Large simulation example - results

The structure of the prior on Q resulted in both merging and emptying of extra states depending on the choice of $\underline{\alpha}$.

- Choice of $\underline{\alpha} = \overline{\alpha}$ caused all states to be occupied regardless of the value of $\overline{\alpha}$, and estimation of 4 states in all cases.
- An asymmetric prior with $\overline{\alpha} > \underline{\alpha}$ caused extra states to be assigned small stationary distributions with few or no observations. This was most clearly observed under the theoretically given constraints.

Distribution of number of occupied states $P(K_A)$

$\bar{\alpha}$	$\underline{\alpha}$	$P(K_A = 2)$	$P(K_A = 3)$	$P(K_A = 4)$
172 (Theory)	$\bar{\alpha}$	0.0000	0.0000	1.0000
	0.01	0.9651	0.0348	0.0000
	1/n	1.0000	0.0000	0.0000
4 (K)	$\bar{\alpha}$	0.0000	0.0000	1.0000
	0.01	0.8539	0.1446	0.0015
	1/n	1.0000	0.0000	0.0000
1	$\bar{\alpha}$	0.0000	0.0000	1.0000
	0.01	0.8782	0.1130	0.0088
	1/n	1.0000	0.0000	0.0000

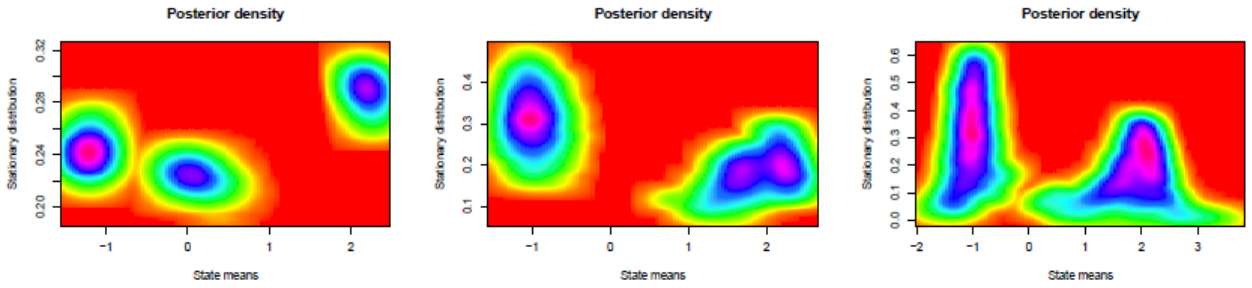
Large simulation example - results

Estimated bivariate density of posterior means (x-axis) and posterior stationary distribution (y-axis)

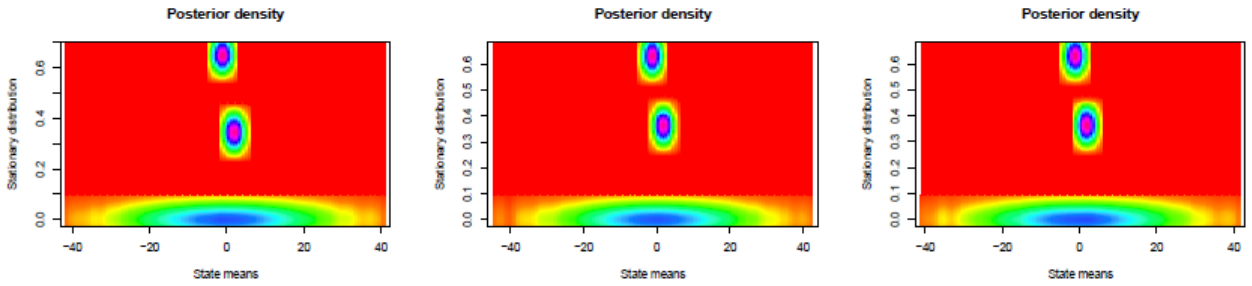
- For $\overline{\alpha} = 1$, extra states merged to some degree with the true states (2); similarly for $\overline{\alpha} = 4$.
- For the largest value of $\overline{\alpha}$, three modes were created; the extra states merged to create a spurious mode.
- For all values of $\overline{\alpha}$, a similar posterior space was created when $\underline{\alpha} = \frac{1}{n}$: two modes at the true parameter values, atop a 'pool' of samples representing MCMC draws from empty states (γ_k drawn directly from the prior).

Fitting HMM with K=4 states to HMM with $K^* = 2$ states; $q_{i,\cdot} \sim D(\overline{\alpha}_1, \underline{\alpha}_2, \alpha_3, \underline{\alpha}_4)$.

Columns: different values of $\overline{\alpha}$
Rows: different prior configs
1: overfitted posterior, $\overline{\alpha} = \underline{\alpha}$
2: $\underline{\alpha} = \frac{1}{n} = 0.0001$



(a) $q_{i,\cdot} \sim \mathcal{D}(172, 172, 172, 172)$ (b) $q_{i,\cdot} \sim \mathcal{D}(4, 4, 4, 4)$ (c) $q_{i,\cdot} \sim \mathcal{D}(1, 1, 1, 1)$



(d) $q_{i,\cdot} \sim \mathcal{D}(172, \frac{1}{n}, \frac{1}{n}, \frac{1}{n})$ (e) $q_{i,\cdot} \sim \mathcal{D}(4, \frac{1}{n}, \frac{1}{n}, \frac{1}{n})$ (f) $q_{i,\cdot} \sim \mathcal{D}(1, \frac{1}{n}, \frac{1}{n}, \frac{1}{n})$

Probability
== Low
==

==
==
== High

Applicability of asymptotic theory for smaller sample sizes

The influence of the hyperparameters can be unexpected and non-trivial, due in part to the intrinsic relationship between $\alpha_{i,j}$ and n.

- $\alpha_{i,j}$ can be interpreted as the prior number of transitions from state i to state j.
- Let $n_{i,j}$ be the number of transitions observed from state i to state j.

The posterior distribution of the transition matrix, given a Dirichlet prior on each row of Q, is

$$p(q_{i,.}|X_t) \sim \mathcal{D}(\bar{\alpha}_{i,1} + n_{i,1}, \underline{\alpha}_{i,2} + n_{i,2}, \cdots, \underline{\alpha}_{i,K} + n_{i,K})$$

- A choice of $\underline{\alpha} = 0.001$ and $K=2$ according to the Theorem leads by theory to $\overline{\alpha} > 3.02$.
- $K=3 \Rightarrow \overline{\alpha} > 36.32$; $K=5 \Rightarrow \overline{\alpha} > 543.38$; $K=10 \Rightarrow \overline{\alpha} > 15{,}498.38$ (not sharp)
- Hence an increasingly large sample size is required to overcome the given $\overline{\alpha}$.

Example – small sample case

HMM with $K^* = 2$ states, and true transition probabilities $q^*_{1,1} = 0.6$, $q^*_{1,2} = 0.4$, $q^*_{2,1} = 0.7$, $q^*_{2,2} = 0.3$

For an arbitrary sample size *n*, assuming the allocations are known, the transition frequencies are expected to be (approximately) $n_{i,j} = n \times q_{i,j}$ for *i,j* = (1,2) and 0 otherwise.

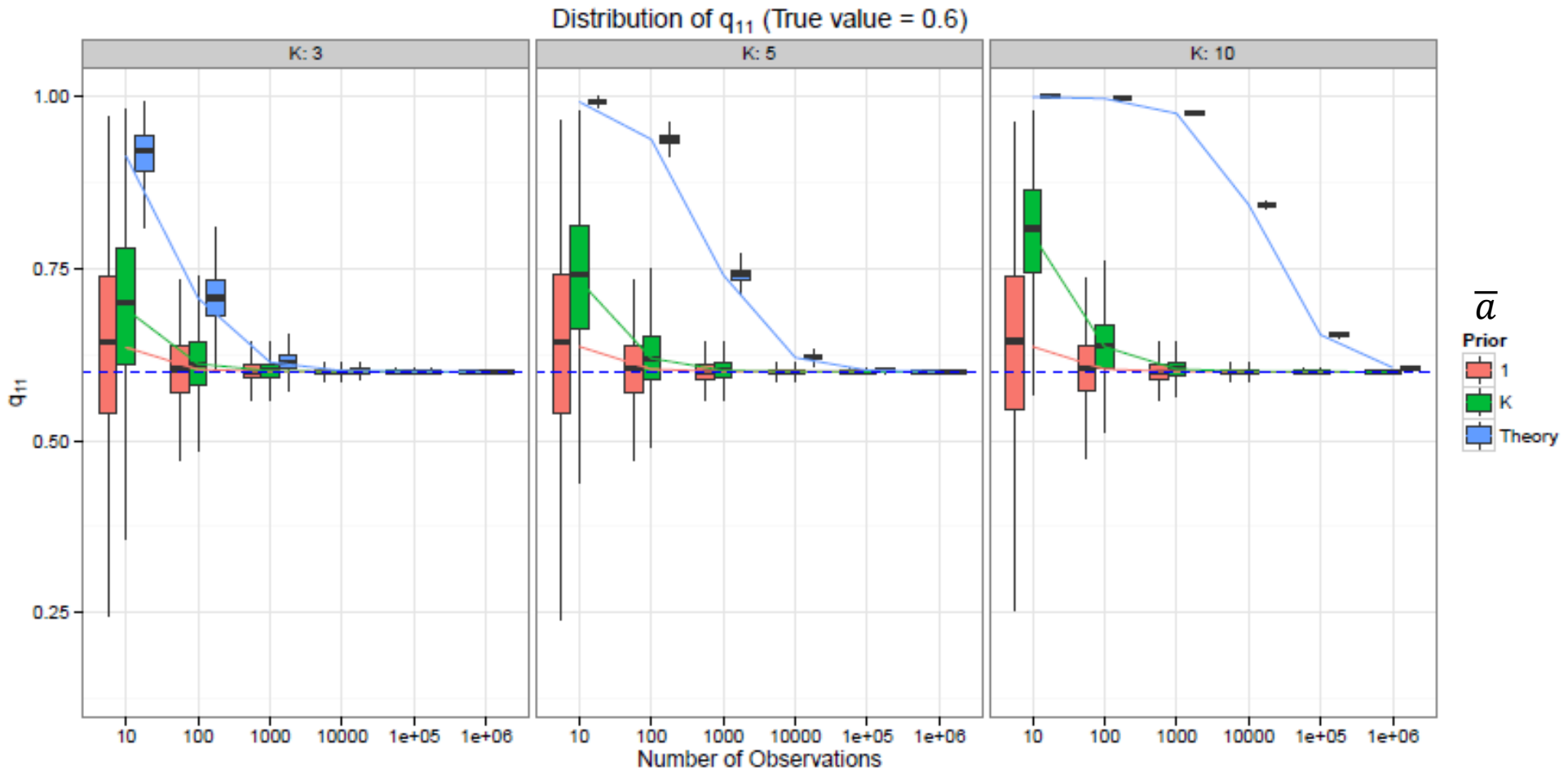
Explore the influence of the hyperparameter $\overline{\alpha}$ on the posterior transition probability $q_{1,1}$:

- Draw 10,000 samples from

$$p(q_{1,}|X_t) \sim \mathcal{D}(\bar{\alpha} + nq^*_{i,1}, \underline{\alpha}_2 + nq^*_{i,2}, \underline{\alpha}_3, \cdots, \underline{\alpha}_K), \text{ where } \underline{\alpha}_2 = \cdots = \underline{\alpha}_K = 0.001$$

- Choosing $\overline{\alpha}$ according to the asymptotic bound had a very strong influence on the estimated distribution of $q_{1,1}$
- When *K=3*, needed ~1,000 observations for the true value of 0.5 to be within the 25th and 75th quantiles of the posterior distribution of $q_{i,j.}$
- For a model overfitted with *K=10* states, needed ~1,000,000 observations.

Example – small sample case



Alternative prior configurations

There are many ways a transition matrix can be written.

Consider the structure of the prior in terms of the position of the large $\overline{\alpha}$ w.r.t. the smaller values $\underline{\alpha}$

- Column prior $\pi_c(Q)$ [used for the asymptotic results]:
  - the prior on each row of *Q* is of the same form, with $\overline{\alpha}$ in the first position, thus favouring the configuration which empties *a priori* the last *K*-*p* states.
- Diagonal prior $\pi_d(Q)$ [used in a nonparametric setting by Nur et al. 2009; Boys & Henderson, 2004]:
  - place $\overline{\alpha}$ on the diagonal of the prior on Q, since the Markov chain is expected to be more likely to remain in the same state than transition to another state *a priori.*
- Mixture prior $\pi_m(Q)$ [leverage the benefits of both approaches]:
  - $0.5\pi_c(Q) + 0.5\pi_d(Q)$

Inducing mixing via prior parallel tempering

Asymmetrical hyperparameters can induce poor MCMC mixing algorithm:
multimodal posteriors, empty states, large areas of low probability separating supported nodes.

Prior Parallel Tempering (van Havre *et al.*, 2015):

- Set up J parallel samplers with slightly different hyperparameters on Q
- Allow $\underline{\alpha}^{j}$ to increase slowly until it matches $\overline{\alpha}$ in the parallel samplers
- Allow the samplers to share information by swapping posterior samples when samples are close, via a Metropolis Hastings acceptance step
- Track the acceptance ratio to ensure acceptable mixing between samplers

Small sample size simulation – setup

Sim 1: $K^* = 3$ states, one well separated, two overlap

Sim 2: $K^* = 3$ states, different state means and transition probabilities

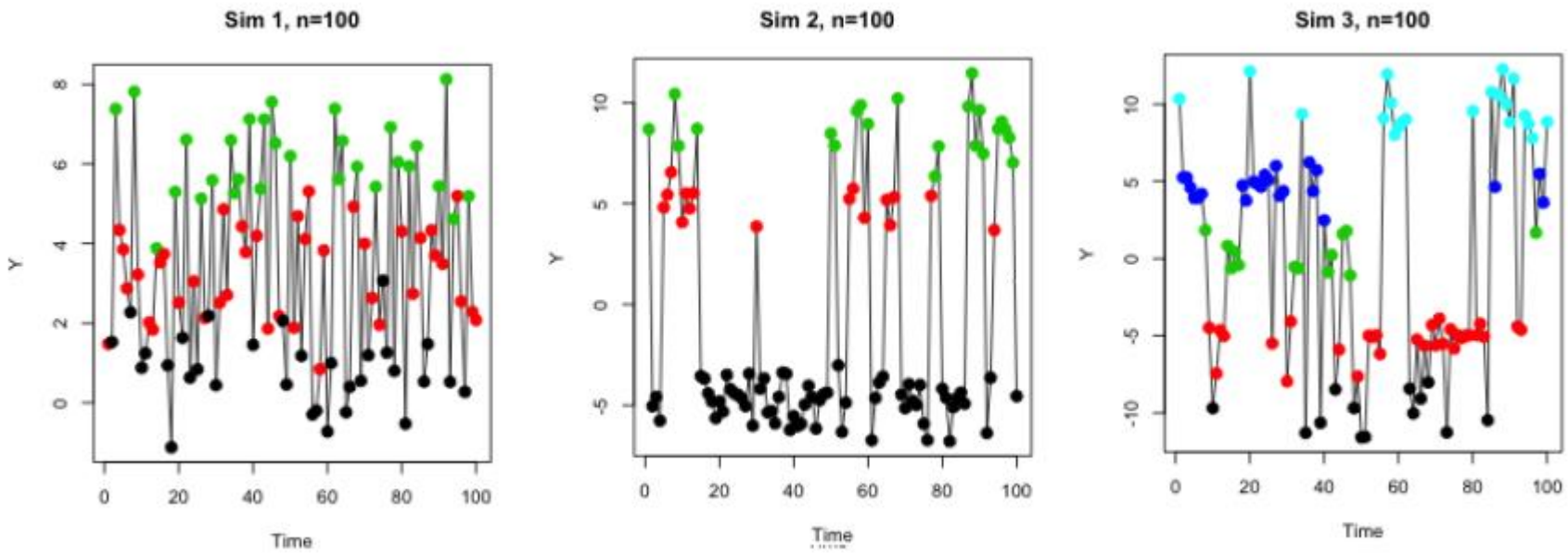
Sim 3: $K^* = 5$ states, well separated, equally spaced means, mainly large values on the diagonal of Q^* (sticky)
All have known state specific variances equal to 1.

$$\textbf{Sim 1}\ \gamma^*_{S1} = (1, 3, 6),\ \mu^*_{S1} = (0.33, 0.38, 0.29),\ \text{and}\ Q^*_{S1} = \begin{bmatrix} 0.2 & 0.3 & 0.5 \\ 0.5 & 0.25 & 0.25 \\ 0.25 & 0.65 & 0.1 \end{bmatrix}$$

$$\textbf{Sim 2}\ \gamma^*_{S2} = (-5, 5, 9),\ \mu^*_{S2} = (0.56, 0.18, 0.26),\ \text{and}\ Q^*_{S2} = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.4 & 0.4 \\ 0.3 & 0.2 & 0.5 \end{bmatrix},$$

$$\textbf{Sim 3}\ \gamma^*_{S3} = (-10, -5, 0, 5, 10),\ \mu^*_{S3} = (0.11, 0.24, 0.20, 0.22, 0.22),$$
$$\text{and}\ Q^*_{S3} = \begin{bmatrix} 0.2 & 0.3 & 0.1 & 0.2 & 0.2 \\ 0.1 & 0.6 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.6 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.6 \end{bmatrix}.$$

Small sample size simulation - realisations



Small sample size simulation – results

Evaluation:

n = 100, 500
K = 10 states fitted under the three prior configurations
hyperparameters: $\overline{\alpha} = (n, K, 1)$, $\underline{\alpha} = \left(\frac{1}{n}, \frac{1}{10n}\right)$

Gibbs sampler with 20 PPT chains for 20,000 iterations, 10000 burn-in

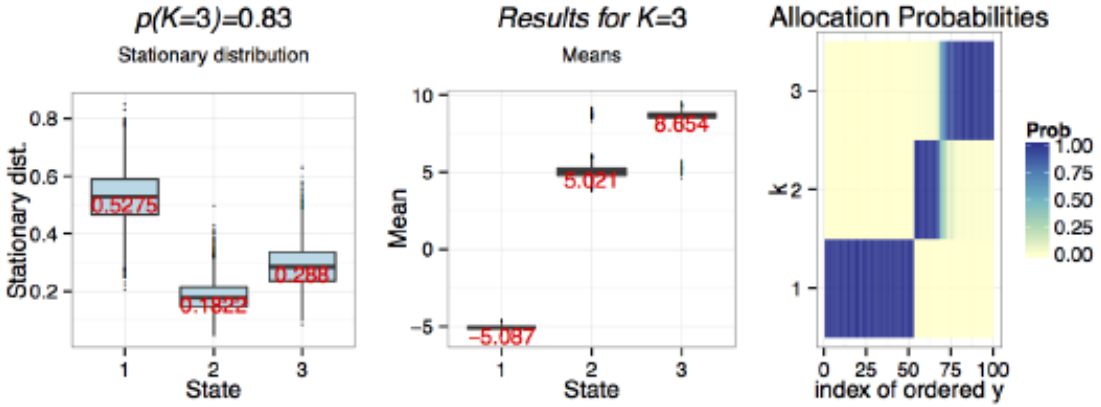
Results for Sim 1 ($K^* = 3$): proportion of replicates which contains 2,3,4 occupied states

			Prior type: π_c		π_d			π_m	
n	$\bar{\alpha}$	$\underline{\alpha}$	$P(\hat{K}_A = 2)$	$P(\hat{K}_A = 3)$	$P(\hat{K}_A = 2)$	$P(\hat{K}_A = 3)$	$P(\hat{K}_A = 4)$	$P(\hat{K}_A = 2)$	$P(\hat{K}_A = 3)$
100	n	$1/n$	0.921	0.079	1.000	-	-	1.000	-
100		$1/10n$	0.941	0.059	1.000	-	-	1.000	-
100	10	$1/n$	0.842	0.158	1.000	-	-	0.969	0.031
100		$1/10n$	0.947	0.053	1.000	-	-	1.000	-
100	1	$1/n$	0.686	0.286	0.921	0.079	-	0.767	0.233
100		$1/10n$	1.000	-	1.000	-	-	1.000	-
500	n	$1/n$	0.889	0.111	1.000	-	-	1.000	-
500		$1/10n$	0.889	0.111	1.000	-	-	1.000	-
500	10	$1/n$	0.333	0.667	0.889	0.111	-	0.667	0.333
500		$1/10n$	0.714	0.286	1.000	-	-	0.667	0.333
500	1	$1/n$	-	1.000	0.500	0.500	-	-	1.000
500		$1/10n$	-	1.000	0.500	0.333	0.167	-	1.000

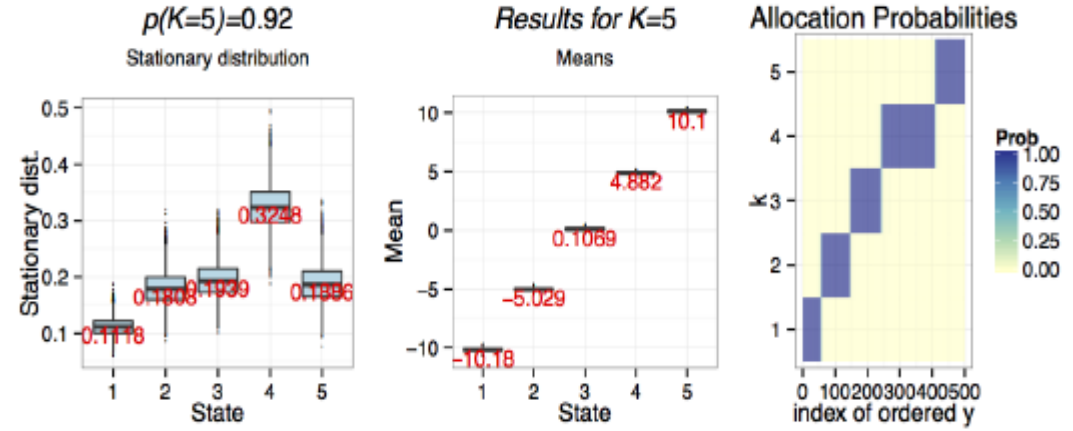
Small sample size simulation – results

Results for Sim 2 and Sim 3: column prior only

Sim	K_A	$P(K_A)$	%
2 (n=100)	3	0.83	0.96
2 (n=100)	4	0.15	0.53
3 (n=100)	5	0.39	0.99
3 (n=100)	6	0.45	0.77
3 (n=100)	7	0.14	0.17
3 (n=200)	5	0.86	0.99
3 (n=200)	6	0.13	0.76
3 (n=500)	5	0.92	0.98
3 (n=500)	6	0.08	0.29



(a) Sim 2, n=100.



(d) Sim 3, n=500

Back to the simple study

True model: standard normal distribution, which is equivalent to a HMM with single state *K = 1*. For *t = 1,…,1000:*

$$y_t \sim \text{Normal}(0, 1)$$

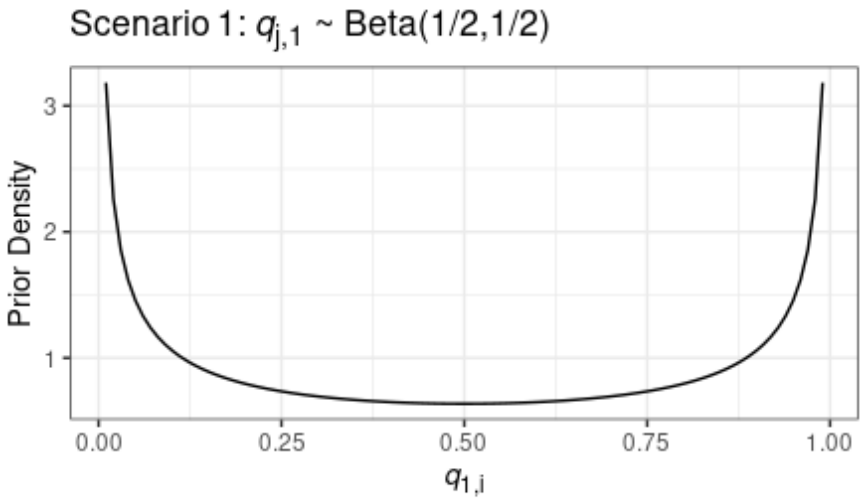
Overfitted model: HMM with *K = 2* states and normal emission distributions:

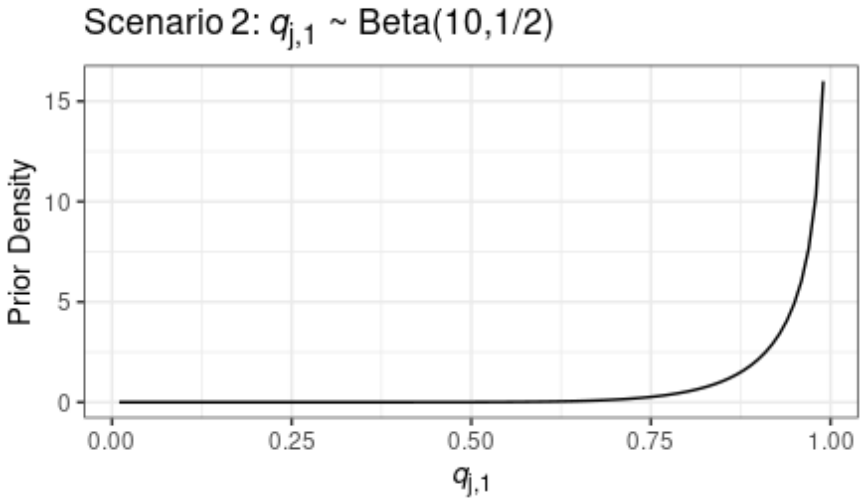
$$y_t | X_t = j \sim \text{Normal}(\mu_j, 1)$$

Priors:

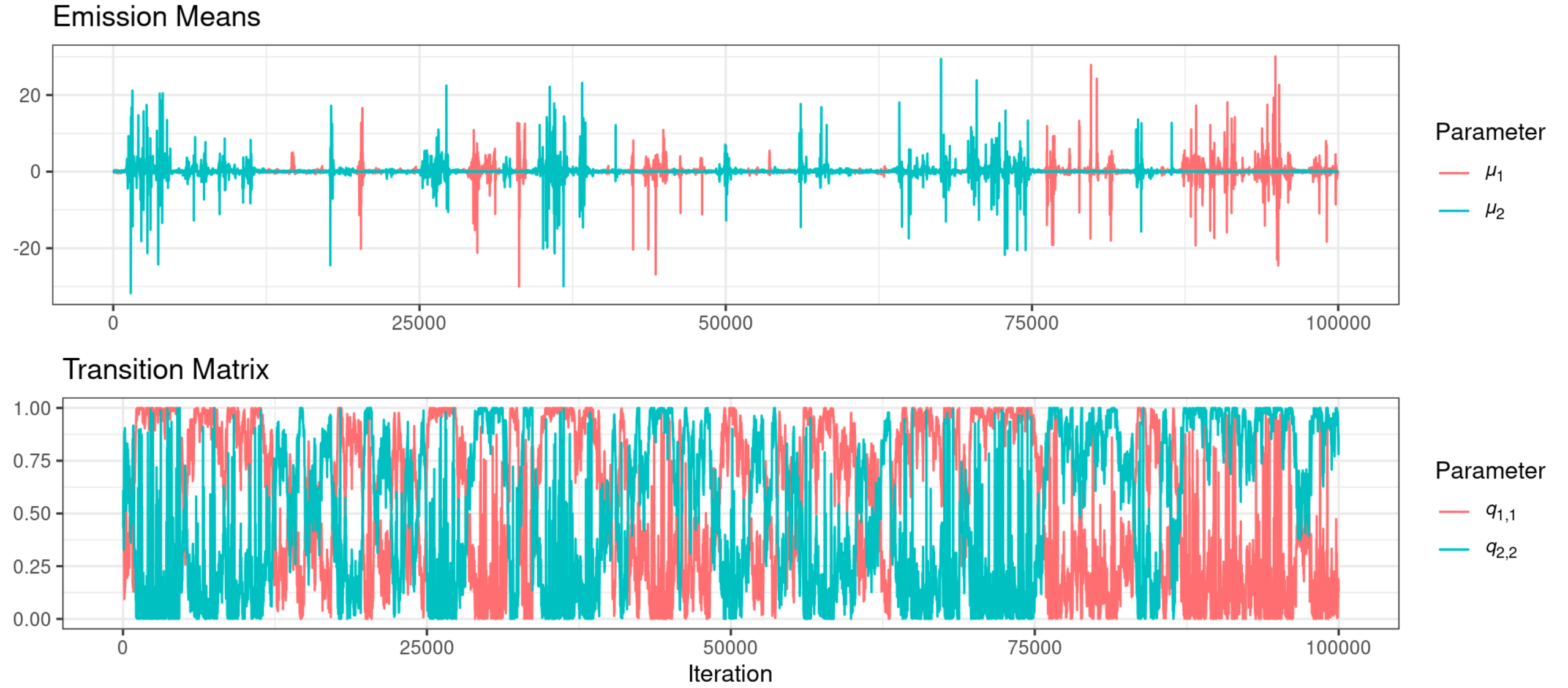
$$\mu_1, \mu_2 \sim \text{Normal}(\bar{y}, 10^2)$$

$$q_{1,1}, q_{2,1} \sim \text{Beta}(\alpha_1, \alpha_2)$$





Scenario 1


Emission Means
20
0
-20
0
25000
50000
75000
100000
Parameter
μ_1
μ_2
Transition Matrix
1.00
0.75
0.50
0.25
0.00
0
25000
50000
75000
100000
Iteration
Parameter
$q_{1,1}$
$q_{2,2}$

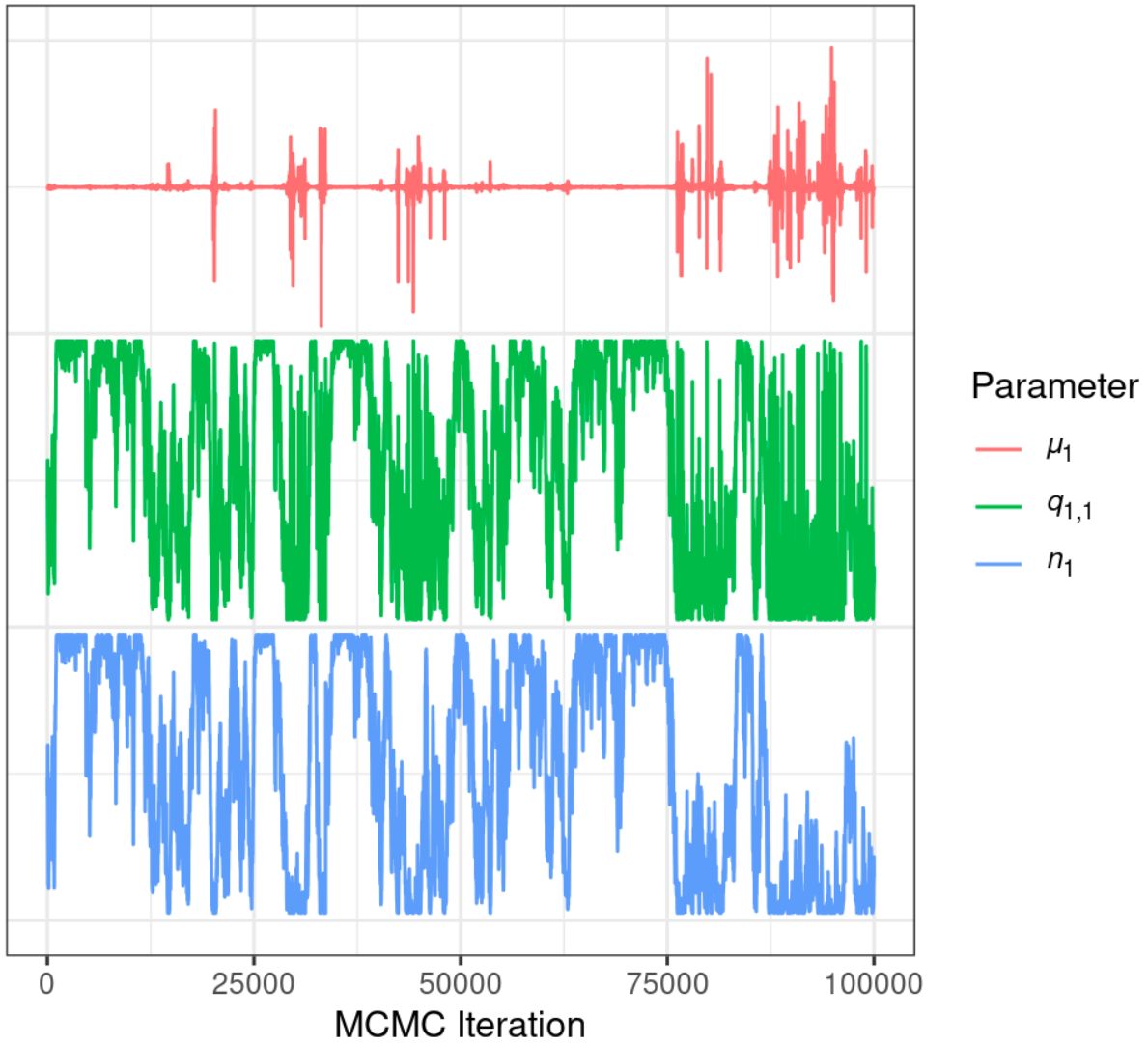
Scenario 1

n_1: Number of observations in state 1.

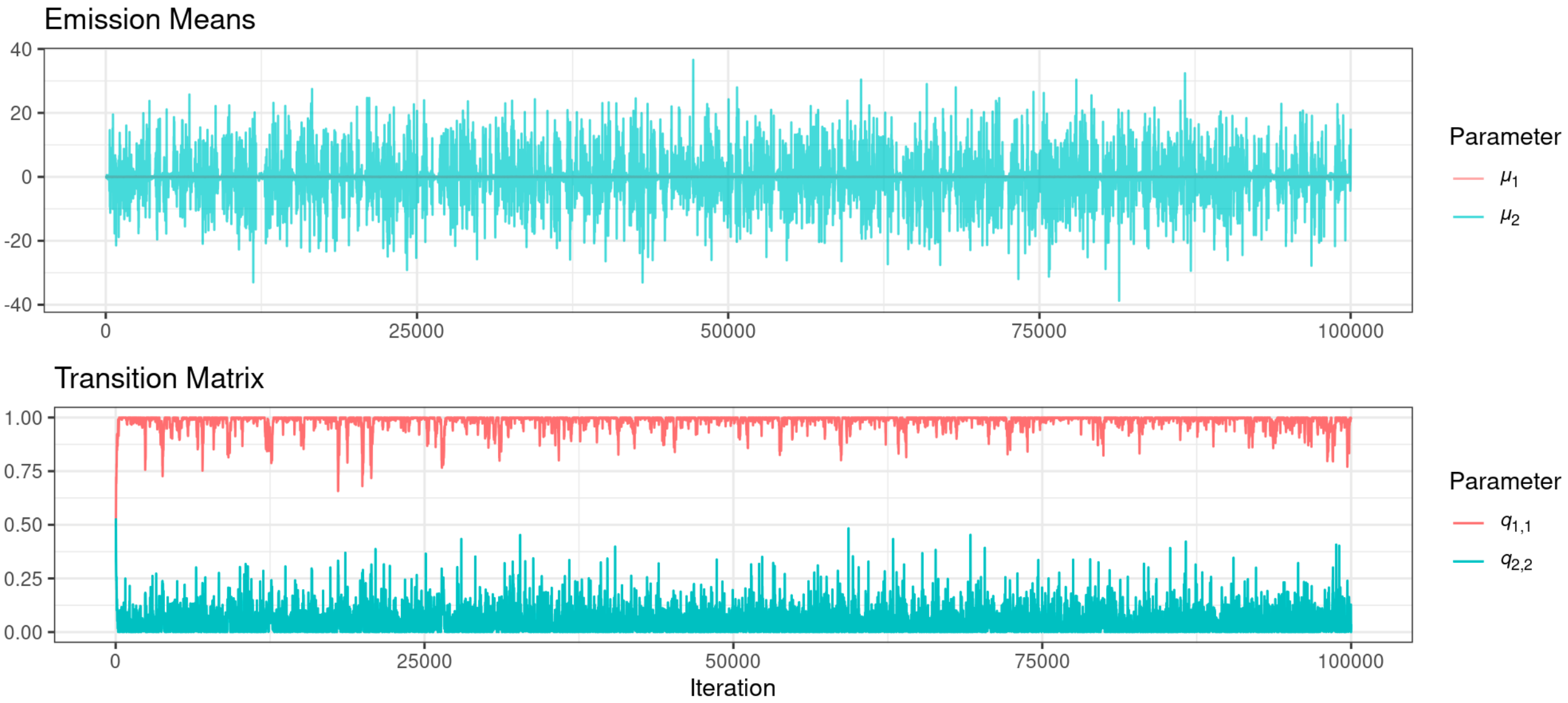
MCMC chain shows very slow mixing in transition matrix parameters.

“Sparks” in emission mean parameters occur when the state empties. i.e. a draw from the prior.

Similar behaviour seen in $q_{1,1}$ and $q_{2,2}$.



Scenario 2


Emission Means
40
20
0
-20
-40
0
25000
50000
75000
100000
Parameter
μ_1
μ_2
Transition Matrix
1.00
0.75
0.50
0.25
0.00
0
25000
50000
75000
100000
Iteration
Parameter
$q_{1,1}$
$q_{2,2}$

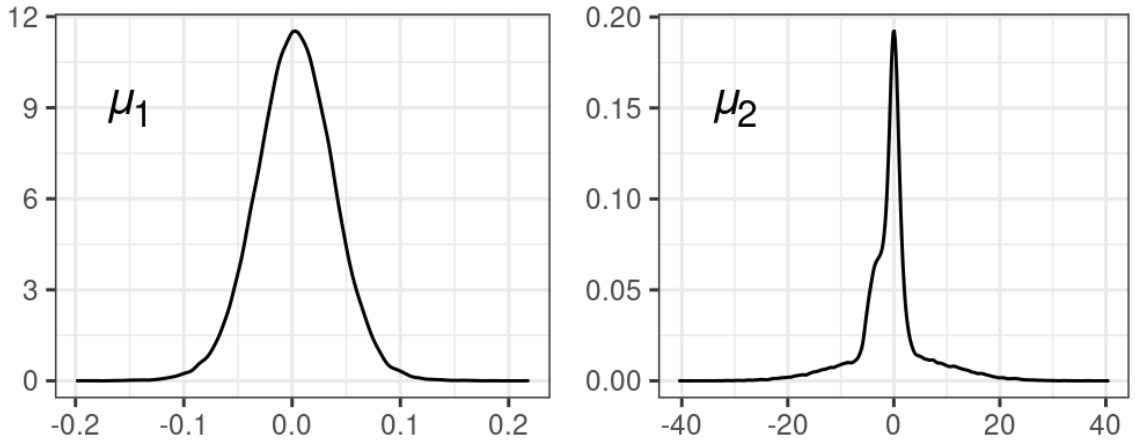
Scenario 2

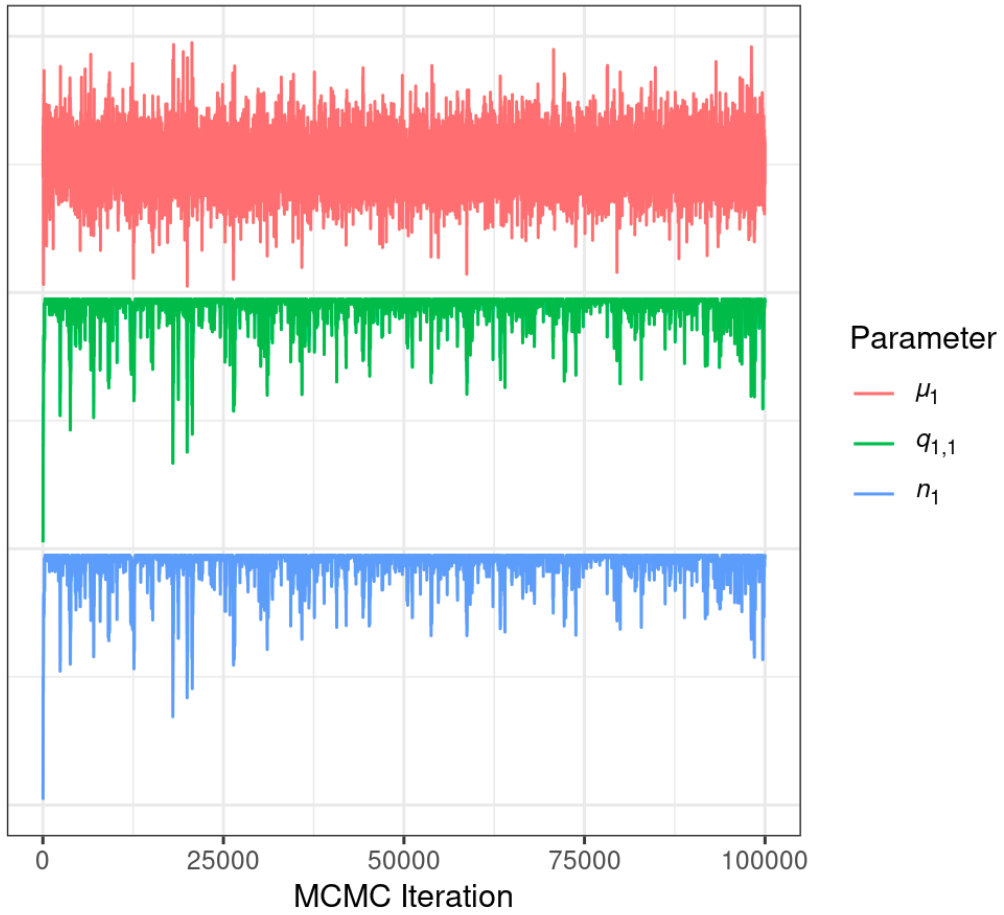
n_1: Number of observations in state 1.

MCMC chain shows better mixing than Scenario 1.

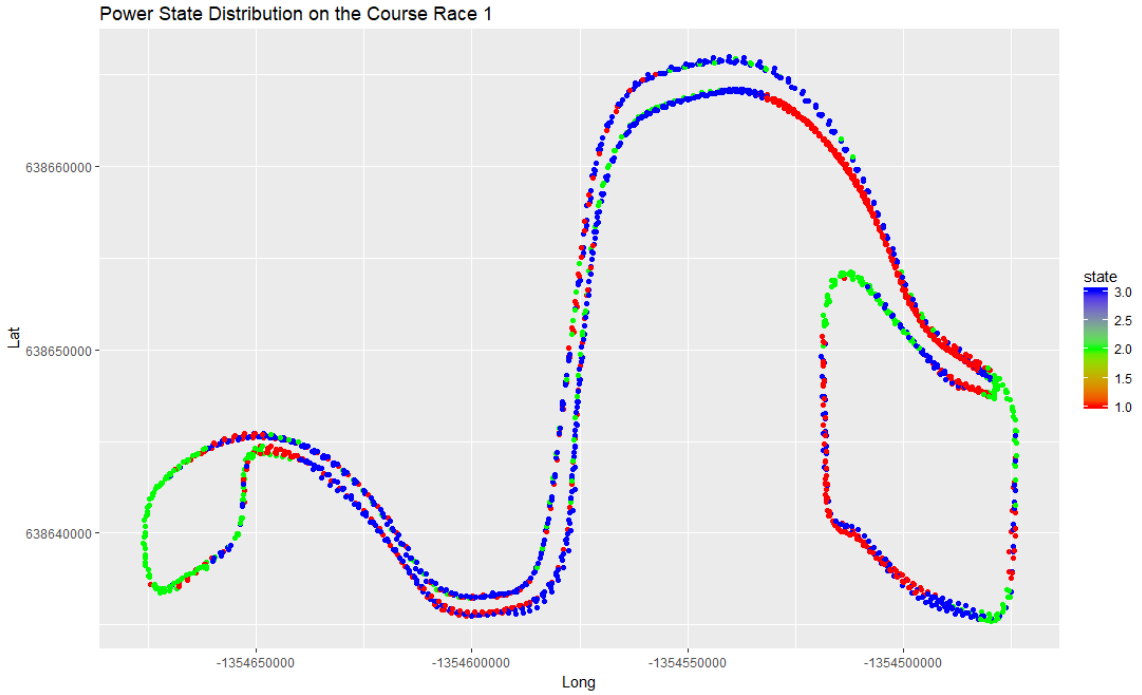
Data are mostly emptied from state 2.

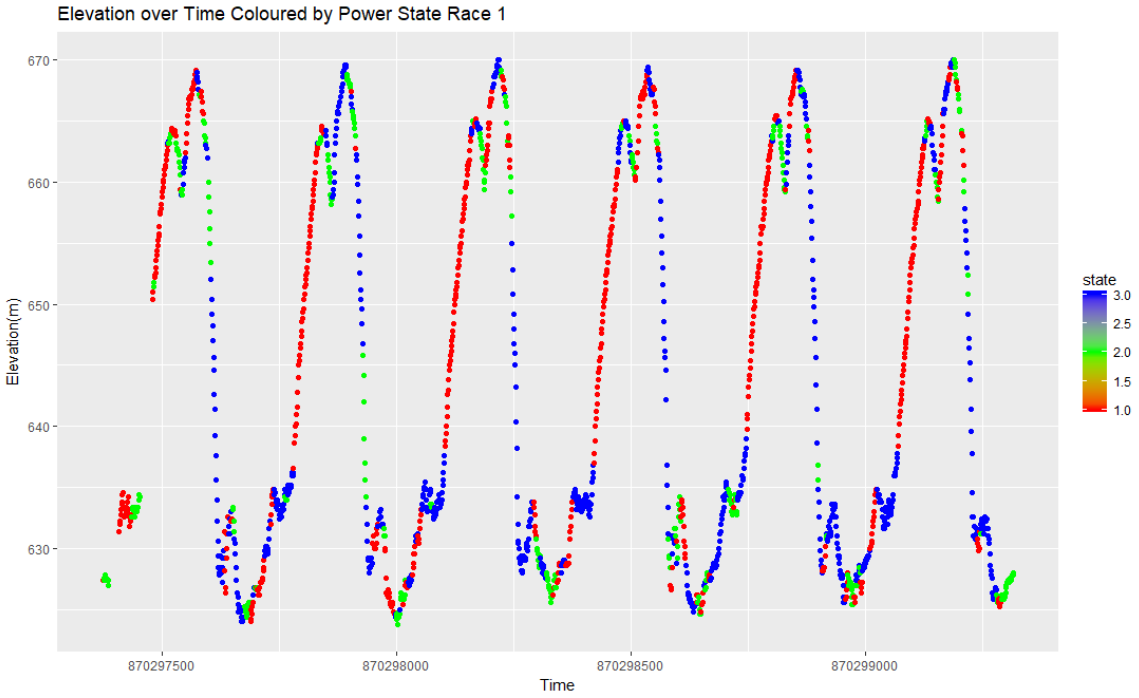
Posterior densities show approximate normality for μ_1 but not μ_2.





Current application





Overall conclusions

1. Overfitting HMMs in such a way as to empty out the stationary distribution of extra states is theoretically and practically possible.
2. While the number of occupied states was not proven to be a consistent estimator of the true number, careful choice of hyperparameters can encourage extra states to be rarely allocated observations in practice.
3. We suggest the choice of $\underline{a} = \frac{a_0}{n}$ for some a_0: then possibly $\widehat{K}_A$ becomes consistent; simulations pointed to this, but it is still only conjecture.
4. In practice, check the posterior samples and distribution of the number of components for evidence that $\underline{a}$ is sufficiently small (i.e. concentrated distribution for K_A).
5. The value of $\overline{a}$ dictated by the asymptotic constraints is concerning in practice due to the relationship between this parameter and sample size. Use PPT or similar to allow mixing while emptying out extra states, especially for small *n*.
6. The column prior is theoretically justified and leads to better MCMC behaviour. Second choice is a mixture prior. The diagonal prior allowed escape from merged configuration but was observed to be inconsistent.

References

CL Alston, KL Mengersen, JM Thompson, PJ Littlefield, D Perry, AJ Ball (2005) Extending the Bayesian mixture model to incorporate spatial information in analysing sheep CAT scan images. Australian Journal of Agricultural Research, 56(4):373–388.

CL Alston, KL Mengersen, CP Robert, JM Thompson, PJ Littlefield, D Perry, AJ Ball (2007) Bayesian mixture models in a longitudinal setting for analysing sheep CAT scan images. Computational Statistics & Data Analysis, 51(9):4282–4296.

CL Alston, KL Mengersen (2010) Allowing for the effect of data binning in a Bayesian Normal mixture model. Computational Statistics & Data Analysis, 54(4):916–923.

CL Alston-Knox, KL Mengersen, R Denham, CM Strickland (2018) Modelling habitat and planning surveillance using landsat imagery: a case study using imported red fire ants. Biological Invasions 20 (5), 1349 -1367.

M Cespedes, J McGree, CC Drovandi, K Mengersen, LB Reid, JD Doecke, J Fripp (2017) A Bayesian hierarchical approach to jointly model structural biomarkers and covariance networks. Arxiv.

M Cespedes, J Fripp, JM McGree, CC Drovandi, K Mengersen, JD Doecke (2017) Comparison of neurodegeneration over time between health ageing and Alzheimer's disease cohorts via Bayesian inference. BMJ Open, 7(2).

MG Falk, CL Alston, CA McGrory, SJ Clifford, EA Heron, D Leonte, M Moores, C Walsh, TN Pettitt, KL Mengersen (2015) Recent Bayesian approaches for spatial analysis of 2-D images with application to environmental modelling, Environmental and Ecological Statistics, 22 (3), p571-600.

References

M Moores, G Nicholls, A Pettitt, K Mengersen (2018) Scalable Bayesian Inference for the Inverse Temperature of a Hidden Potts Model. Bayesian Analysis.

MT Moores, CE Hargrave, T Deegan, M Poulsen, F Harden, K Mengersen (2015) An external field prior for the hidden Potts model with application to cone-beam computed tomography. Computational Statistics and Data Analysis 86 27-41.

MT Moores, C Drovandi, K Mengersen, CP Robert (2015) Pre-processing for approximate Bayesian computation in image analysis. Statistics and Computing, 25(1), pp. 23-33.

CM Strickland, RL Burdett, R Denham, KL Mengersen (2014) PySSM : a Python module for Bayesian inference of linear Gaussian state space models. Journal of Statistical Software, 57(6), 1-37 J Holloway, KJ Helmstedt, K Mengersen, M Schmidt (2019) A decision tree approach for spatially interpolating missing land cover data and classifying satellite images. Remote Sensing, 11(15).

CM Strickland, DP Simpson, IW Turner, R Denham, KL Mengersen (2011) Fast Bayesian analysis of spatial dynamic factor models for multitemporal remotely sensed imagery. Journal of the Royal Statistical Society Series C–Applied Statistics, 60(Part 1):109–124.

CM Strickland, IW Turner, R Denham, KL Mengersen (2009) Efficient Bayesian estimation of multivariate state space models. Computational Statistics & Data Analysis, 53(12):4116–4125.

I Ullah, K Mengersen (2019) Bayesian mixture models and their Big Data implementations with application to invasive species presence-only data. Journal of Big Data 6 (1), 29.

References

Boys, R. J. and Henderson, D. a. (2004). "A Bayesian approach to DNA sequence segmentation." *Biometrics*, 60(3): 573–578. 2, 3, 15

Boys, R. J., Henderson, D. A., and Wilkinson, D. J. (2000). "Detecting homogeneous segments in DNA sequences by using hidden Markov models." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(2): 269–285. 3, 14

Frühwirth-Schnatter, S. (2008). *Finite Mixture and Markov Switching Models*. Mcmc. Springer, 1 edition. 2, 9, 33, 34

Gassiat, E. and Rousseau, J. (2012). "About the posterior distribution in hidden markov models with unknown number of states." *arXiv preprint arXiv:1207.2064*, (1997): 1–28. 3, 6, 8, 31

Nur, D., Allingham, D., Rousseau, J., Mengersen, K. L., and McVinish, R. (2009). "Bayesian hidden Markov model for DNA sequence segmentation: A prior sensitivity analysis." *Computational Statistics & Data Analysis*, 53(5): 1873–1882. 3, 14

Rousseau, J. and Mengersen, K. (2011). "Asymptotic behaviour of the posterior distribution in overfitted mixture models." *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73(5): 689–710. 2, 6, 7

van Havre, Z., White, N., Rousseau, J., and Mengersen, K. (2015). "Overfitting Bayesian Mixture Models with an Unknown Number of Components." *Plos One*, 10(7): e0131739. 1, 3, 4, 15, 16, 17, 19