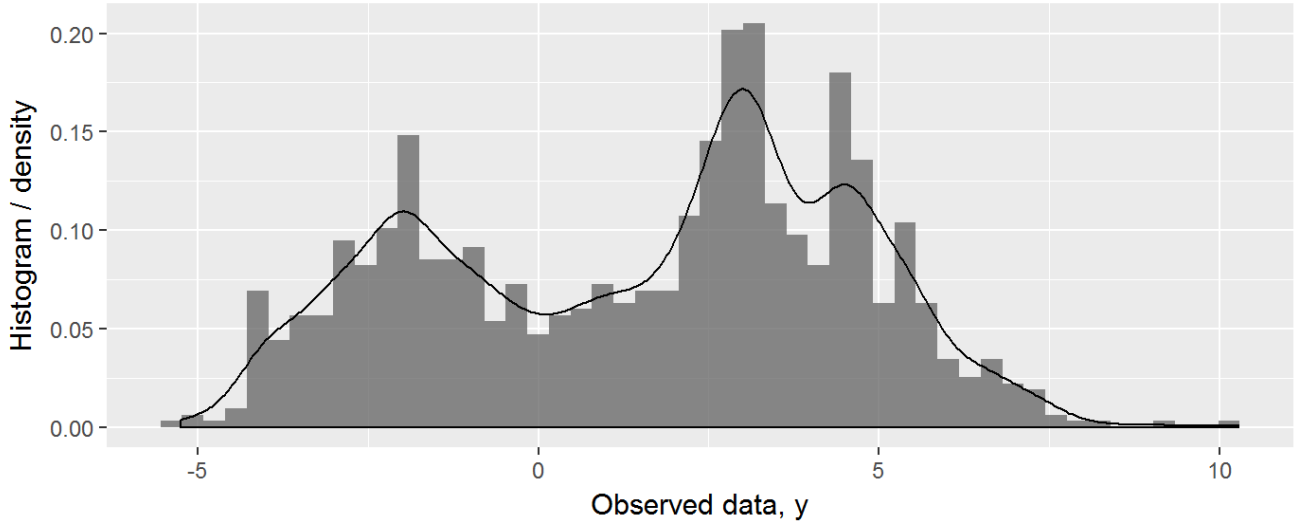
A to Z: 20 years of progress on the label switching problem

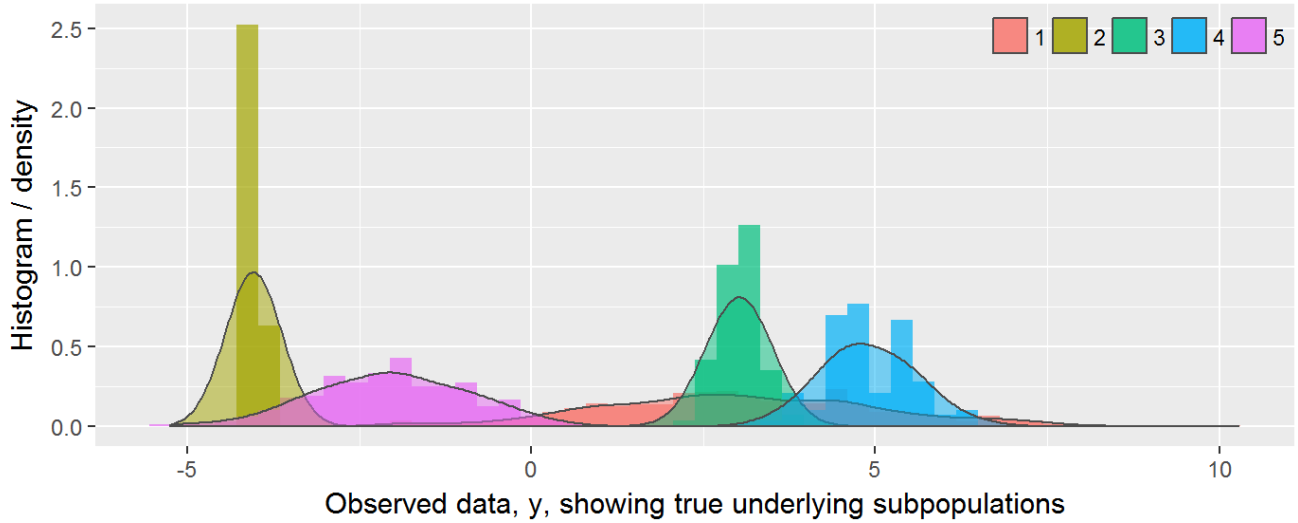
Earl Duncan

Bayes on the Beach Conference

15 Nov 2017

- Consider the following data, $\boldsymbol{y} = (y_1, \ldots, y_{1000})$:





- The K-component mixture model is expressed as

$$\boldsymbol{Y} \sim p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\phi}, \boldsymbol{\lambda}) = \prod_{i=1}^{N} \sum_{k=1}^{K} w_k f_k(y_i|\boldsymbol{\phi}_k, \boldsymbol{\lambda})$$

where $\boldsymbol{y} = (y_1, \ldots, y_N)$ is the observed data, $\boldsymbol{\phi}_k$ and $\boldsymbol{\lambda}$ denote unknown component-specific and common parameter(s) respectively, and $f_k(\cdot)$ is the k^{th} component density with corresponding mixture weight w_k subject to:

$$\sum_{k=1}^{K} w_k = 1 \text{ and } w_k \geq 0 \text{ for } k = 1, \ldots, K.$$

- Note: $\boldsymbol{\theta} = \left\{\boldsymbol{\phi}_{1,(1,\ldots,K)}, \ldots, \boldsymbol{\phi}_{R-1,(1,\ldots,K)}, \boldsymbol{w}\right\}$.

Marin, J-M., K. Mengersen, and C. P. Robert. 2005. "Bayesian modelling and inference on mixtures of distributions" In *Handbook of Statistics* edited C. Rao and D. Dey. New York: Springer-Verlag.

- A latent allocation variable Z_i is used to identify which component Y_i belongs to.

$$Y_i|z_i, \boldsymbol{\phi}, \boldsymbol{\lambda} \sim f_{z_i}\left(y_i|\boldsymbol{\phi}_{z_i}, \boldsymbol{\lambda}\right)$$
$$Z_i|\boldsymbol{w} \sim \text{Cat}(w_1, \dots, w_K)$$

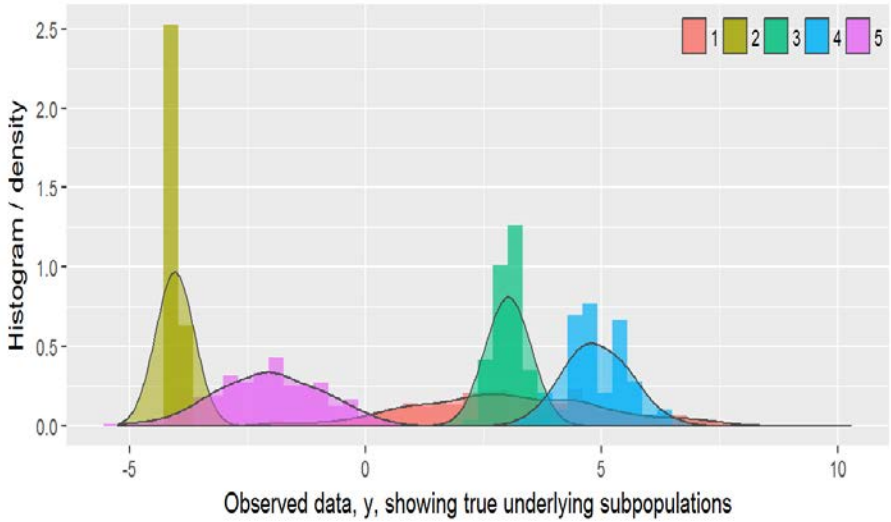
- What happens if we swap the labels? E.g.

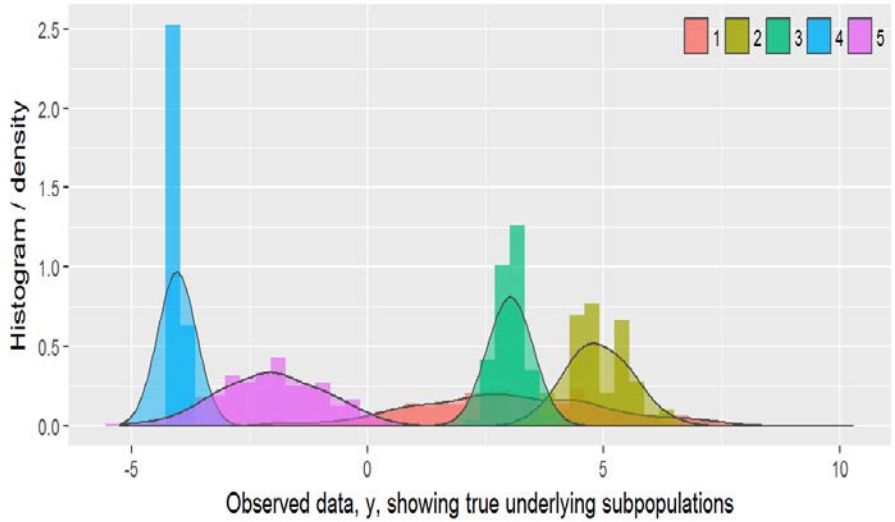
$$z_1 \coloneqq z_2 \quad \Rightarrow \quad f_1 \coloneqq f_2$$
$$z_2 \coloneqq z_1 \quad \Rightarrow \quad f_2 \coloneqq f_1$$

- The likelihood is *exchangeable* meaning that it is invariant to permutations of the labels identifying the mixture components

$$p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\lambda}) = p(\boldsymbol{y}|\tau(\boldsymbol{\theta}),\boldsymbol{\lambda})$$

for any permutation τ.

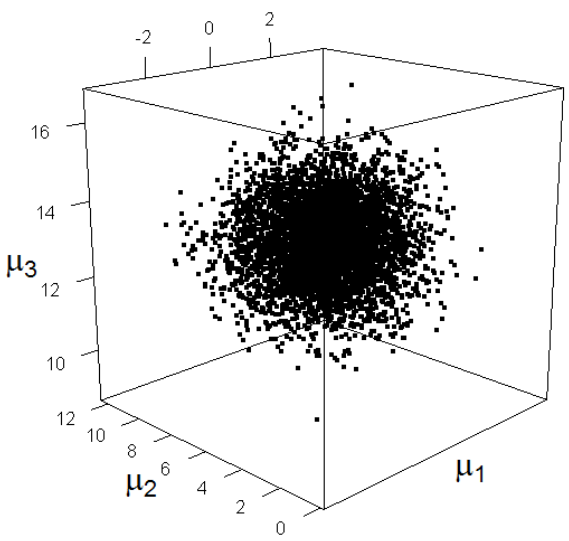




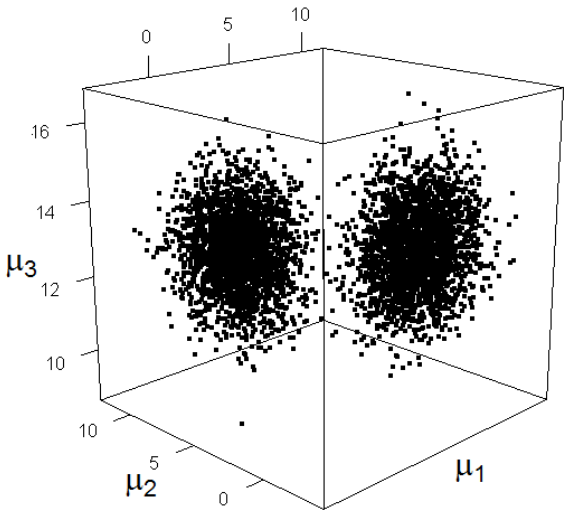
- If the posterior distribution is invariant to permutations of the labels, this is known as ***label switching*** (LS).

- LS will occur if:
  - the prior is (at least partly) exchangeable; and
  - the sampler is efficient at exploring the posterior hypersurface.
- The posterior will have (up to) $K!$ symmetric modes.

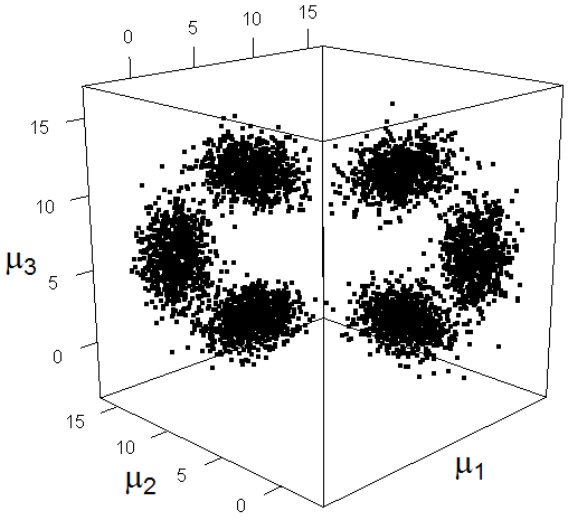
No label switching



LS between groups 1 and 2



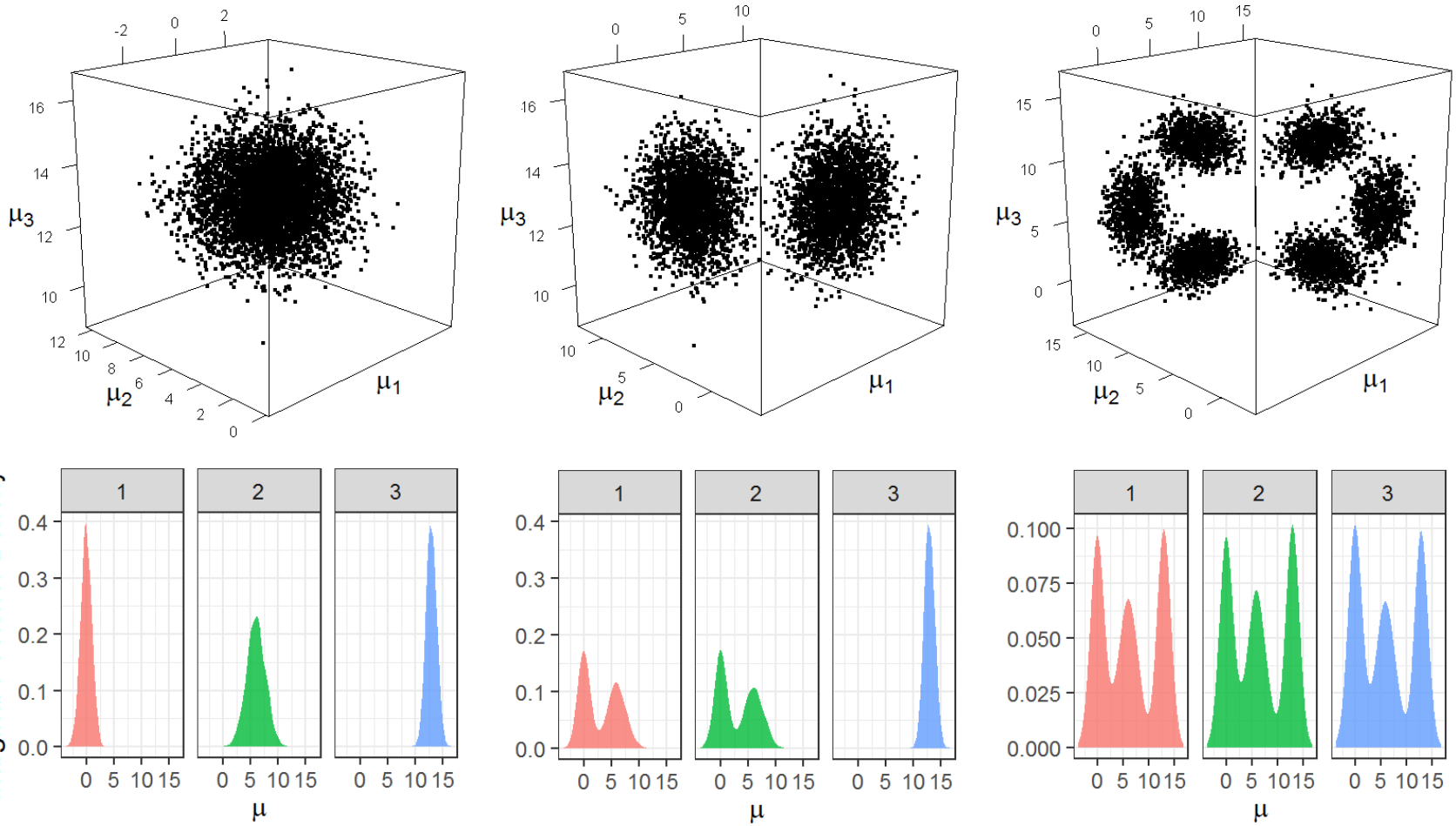
LS between all 3 groups



- Why is LS a problem?

- …Because the marginal posterior distributions are identical for each component. **So how can we make inferences???**

No label switching | **LS between groups 1 and 2** | **LS between all 3 groups**



- One of the earliest solutions to LS:
  - Use an **A**rtificial identifiability constraint (AIC) on some parameters, e.g.

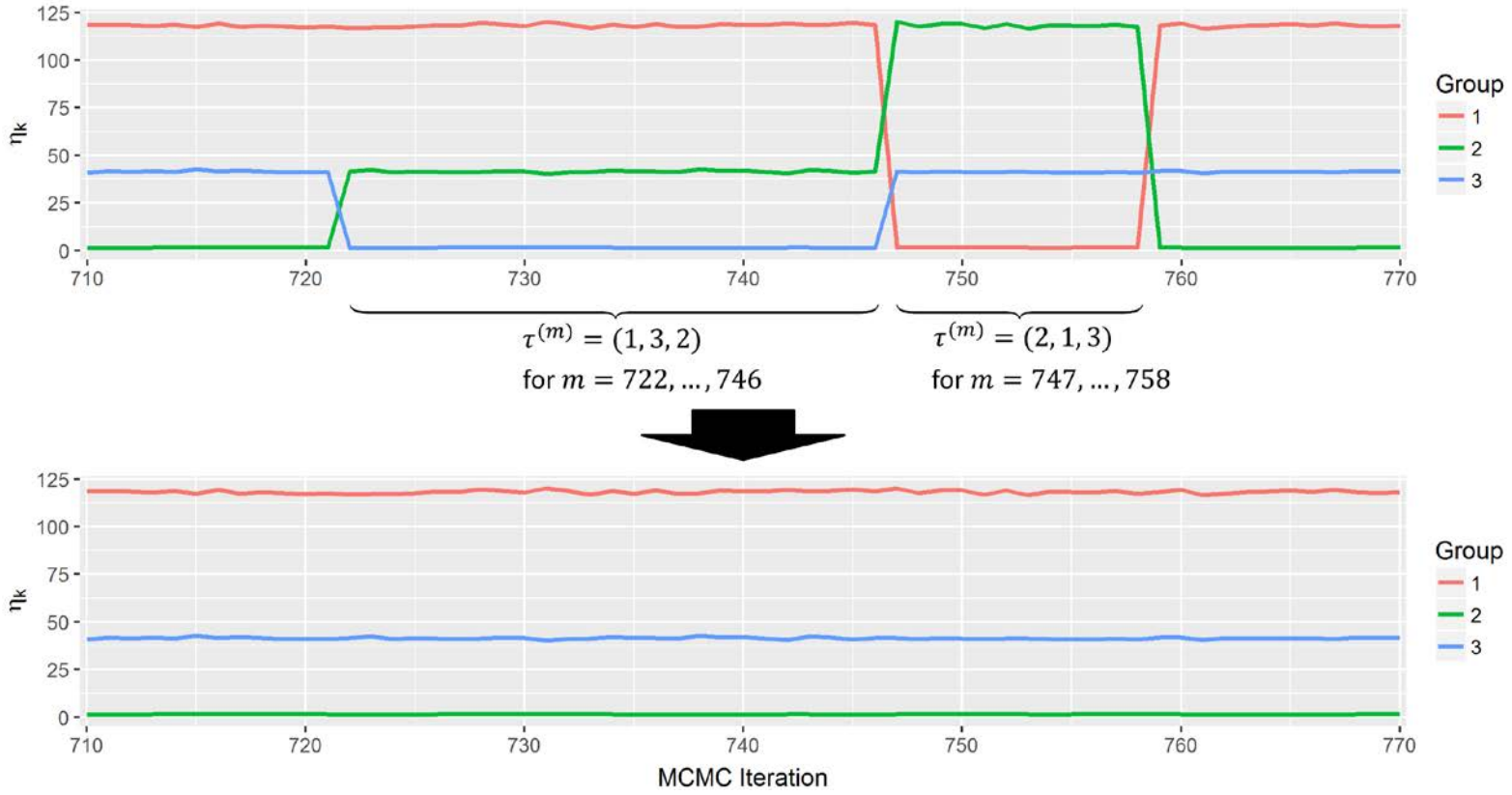
$$f_k(y_i|\boldsymbol{\phi}_k) = \mathcal{N}(y_i; \mu_k, \sigma_k^2)$$
$$p(\boldsymbol{\theta})\mathbb{I}(\mu_1 < \cdots < \mu_K)$$

- Not a good solution!
  - Choosing a suitable AIC is not straightforward.
    - Why not $\sigma_1^2 < \cdots < \sigma_K^2$?.
    - What about multivariate mixtures?
    - What if components are poorly separated?
  - Destroys the non-informativeness of the exchangeable prior.
    - Why not use an informative (non-exchangeable) prior instead?
  - Can have a large influence on the shape of the posterior.
  - Does not guarantee removal of symmetry in the posterior.

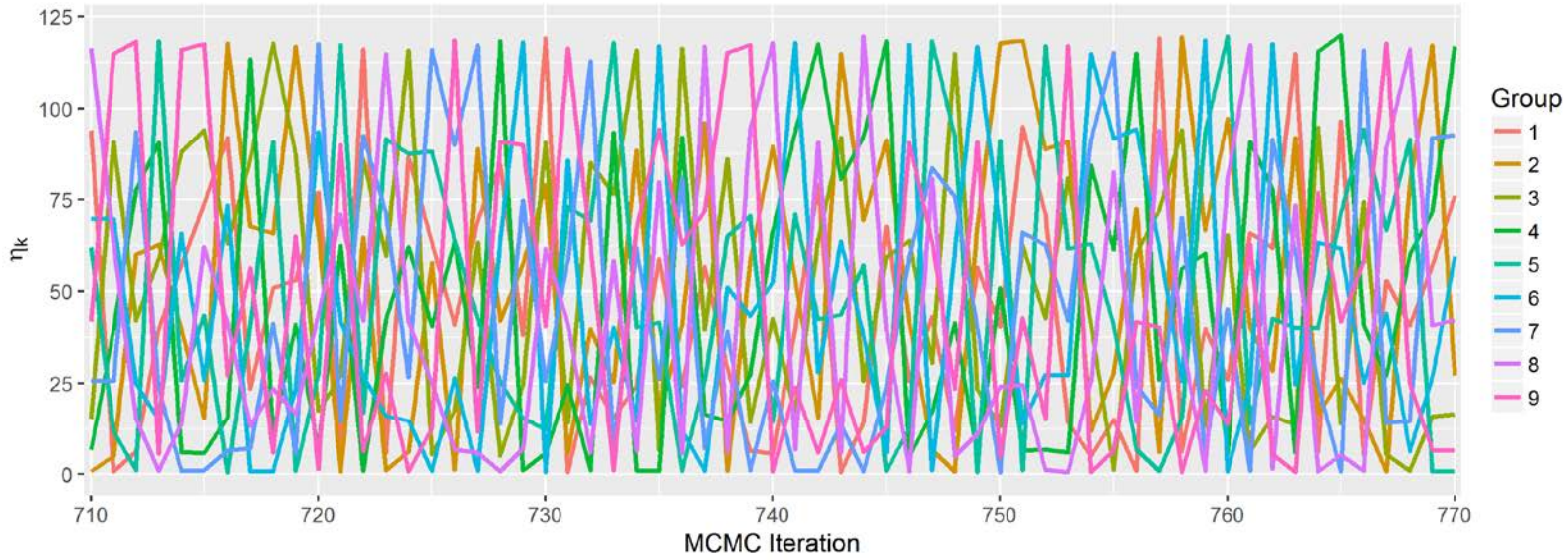
- More decision theoretic solutions have been proposed:

	AIC	Artificial identifiability constraint
2000	KL	Kullack-Leibler divergence algorithm
2005	PRA	Pivotal reordering algorithm
2009	BM	Bernoulli mixture algorithm
2009	BMP	Bernoulli mixture permutation algorithm
2010, 2014	ECR	Equivalence classes representatives algorithms • Non-iterative (ECR) and iterative versions (ECR 1 and ECR 2)
2010	SJW	(Named after authors: Sperrin, Jaki, Wit)
2014	DB	Data-based relabelling algorithm • Non-iterative (DB) and iterative (DB it.)
2015	PU	Pivotal unit relabelling algorithm
2015	ZS	Zswitch relabelling algorithm
2017	ZS 2	Zswitch 2 relabelling algorithm

- Aside from AIC, these approaches aim to *reverse* the effect of label switching by determining the correct permutations $\tau^{(m)}$ for $m = 1, \ldots, M$ (number of MCMC iterations).
- For simple models, this could be done manually:



- Not feasible for large K.



  - Need to use relabelling algorithms!
- Algorithm efficiency is a concern.
  - Searching all $K!$ permutations for the correct one can be very slow.

- Aside: how many permutations in a Rubik's cube?
  - $8! \times 3^7 \times (12!/2) \times 2^{11} \approx 4.325 \times 10^{19} \approx 21!$
  - And yet humans can solve it fast!

World record set 2nd September 2017

https://www.youtube.com/watch?v=np2G0yr5xI0

- The KL, PRA, BMP, all ECR, both DB, ZS, and ZS 2 algorithms find τ by minimising the posterior expectation of some loss function, $\mathrm{E}[\mathcal{L}(a;\boldsymbol{\theta},\boldsymbol{z})|\boldsymbol{y}\,]$.
- Since the likelihood is invariant to permutations of the parameters, the loss function should also be permutation invariant, i.e.

$$\mathcal{L}(a;\boldsymbol{\theta},\boldsymbol{z}) = \mathcal{L}(a;\tau(\boldsymbol{\theta}),\tau^{-1}(\boldsymbol{z})).$$

- If $\mathcal{L}_0(a;\boldsymbol{\theta},\boldsymbol{z})$ denotes a loss function which is not permutation invariant, we define

$$\mathcal{L}(a;\boldsymbol{\theta},\boldsymbol{z}) = \min_{\tau} \mathcal{L}_0\big(a;\tau(\boldsymbol{\theta}),\tau^{-1}(\boldsymbol{z})\big).$$

- If the loss function $\mathcal{L}_0$ is of the form

$$\mathcal{L}_0(a; \boldsymbol{\theta}, \boldsymbol{z}) = \sum_{k=1}^{K} \mathcal{L}_0(a; \boldsymbol{\theta}_k, \boldsymbol{z}(k))$$

then minimising $\mathcal{L}_0$ is equivalent to minimising

$$\sum_{k=1}^{K} c_{\tau(k),k}$$

where $c_{j,k} = \mathcal{L}_0\big(a; \boldsymbol{\theta}_j, \boldsymbol{z}(j)\big)$ is the cost of assigning the k^{th} element of τ the value j, i.e. $\tau(k) = j$.

- That is, the minimisation problem

$$\min_{\tau^{(m)} \in S} \mathcal{L}_0 \left(a; \tau^{(m)}\left(\boldsymbol{\theta}^{(m)}\right), (\tau^{-1})^{(m)}\left(\boldsymbol{z}^{(m)}\right) \right)$$

is equivalent to the linear sum assignment problem (LSAP):

$$\min_{\tau^{(m)} \in S} \sum_{k=1}^{K} c_{\tau_k^{(m)},k} = \min_{b} \sum_{j=1}^{K} \sum_{k=1}^{K} b_{j,k} c_{j,k}^{(m)}$$

subject to

$$\sum_{j=1}^{K} b_{j,k} = \sum_{k=1}^{K} b_{j,k} = 1 \qquad \text{and} \qquad b_{j,k} \in \{0,1\}.$$

- E.g. 4-component mixture:

Set of all permutations:

$$S = \left\{ \begin{matrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \\ 1 & 3 & 2 & 4 \\ 1 & 3 & 4 & 2 \\ 1 & 4 & 3 & 2 \\ 1 & 4 & 2 & 3 \\ 2 & 1 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 2 & 3 & 1 & 4 \\ 2 & 3 & 4 & 1 \\ 2 & 4 & 1 & 3 \\ 2 & 4 & 3 & 1 \\ 3 & 1 & 2 & 4 \\ 3 & 1 & 4 & 2 \\ 3 & 2 & 1 & 4 \\ 3 & 2 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 3 & 4 & 2 & 1 \\ 4 & 1 & 2 & 3 \\ 4 & 1 & 3 & 2 \\ 4 & 2 & 1 & 3 \\ 4 & 2 & 3 & 1 \\ 4 & 3 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{matrix} \right\}$$

Constraint matrix:

$$b = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

- Kullback-Leibler (KL) divergence algorithm (Stephens 2000):

1) Initialise the $M \times K$ matrix of permutations $\boldsymbol{\mathcal{T}} = \{\tau^{(1)}, \ldots, \tau^{(M)}\}$. This is usually initialised so that $\tau^{(m)} = \{1, \ldots, K\}$ for all m.

2) For $i = 1, \ldots, N$ and $k = 1, \ldots, K$, calculate

$$\hat{p}_{i,k} = \frac{1}{M} \sum_{m=1}^{M} p^{(m)}_{i,\tau^{(m)}(k)} \quad \text{where} \quad p_{ik} = \frac{w_k f_k(y_i|\phi_k, \lambda)}{\sum_{j=1}^{K} w_j f_j(y_i|\phi_j, \lambda)}$$

3) For $m = 1, \ldots, M$, determine $\tau^{(m)}$ by solving the LSAP using costs

$$c^{(m)}_{j,k} = \sum_{i=1}^{N} p^{(m)}_{i,j} \log\left(\frac{p^{(m)}_{i,j}}{\hat{p}_{i,k}}\right).$$

4) If an improvement in $\sum_{m=1}^{M} \hat{\mathcal{L}}_0^{(m)}$ has been achieved, return to step 2) and repeat, otherwise stop.

Stephens, M. 2000b. Dealing with label Switching in mixture models. *Journal of the Royal Statistical Society Series B* **62** (4): 795-809. doi: 10.1111/1467-9868.00265

- Pivotal Reordering Algorithm (PRA) (Marin et al. 2005):

  1) Define the pivot $\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(m^*)}$ where m^* is the iteration which corresponds to the Monte Carlo approximation of the *maximum a posteriori* (MAP) estimate of $\boldsymbol{\theta} = \{\boldsymbol{\phi}_k, \mathbf{w}\}$.

  2) For $m = 1, \ldots, M$, determine $\tau^{(m)}$ by maximising the scalar product

$$\tau^{(m)} = \underset{\tau \in S}{\operatorname{argmax}} \sum_{r=1}^{R} \sum_{k=1}^{K} \theta_{r,\tau_k}^{(m)} \theta_{r,k}^*$$

  (This is equivalent to minimising the Euclidean distance between $\tau\left(\boldsymbol{\theta}^{(m)}\right)$ and $\boldsymbol{\theta}^*$.)

  Note that this problem could be formulated as a LSAP using costs

$$c_{j,k}^{(m)} = -\sum_{r=1}^{R} \theta_{r,j}^{(m)} \theta_{r,k}^* .$$

Marin, J-M., K. Mengersen, and C. P. Robert. 2005. "Bayesian modelling and inference on mixtures of distributions" In *Handbook of Statistics* edited C. Rao and D. Dey. New York: Springer-Verlag.

- Zswitch (ZS) (van Havre et al. 2015):

1) Choose one iteration m^* to be the reference, with corresponding allocation vector $\boldsymbol{z}^* = (z_1, \ldots, z_N)^{(m^*)}$ and parameter values $\boldsymbol{\theta}^*$.

2) For $m = 1, \ldots, M$:

Phase 1: Allocation-based relabelling

a) Construct a $K \times K$ matrix $\mathbf{M}$ with elements

$$\mathbf{M}_{j,k} = \sum_{i=1}^{N} \mathbb{I}\left(z_i^{(m)} = j\right) \mathbb{I}(z_i^* = k), \qquad j, k \leq K.$$

```
> M.jk
      z.ref
z.now   1   2   3
    1   0   0 200
    2  90  10   0
    3   0 200   0
```

b) For $j = 1, \ldots, K$, define the set I_j as:

$$I_j = \left\{ k : \frac{\mathbf{M}_{j,k}}{\sum_{k'=1}^{K} \mathbf{M}_{j,k'}} > \omega \right\}.$$

```
> Set.I
[[1]]
2

[[2]]
2 3

[[3]]
1
```

van Havre, Z., N. White, J. Rousseau, and K. Mengersen. 2015. Overfitting Bayesian mixture models with an unknown number of components. *PLoS ONE* **10** (7): e0131739. doi: 10.1371/journal.pone.0131739.

- Zswitch (ZS) (van Havre et al. 2015) continued :

c) Define $\hat{S} \subseteq S$ as the set of permutations arising from the K-fold Cartesian product of each set $\{I_j\}$:

$$\hat{S} = I_1 \times \cdots \times I_K.$$

```
> S.hat
2     2     1
2     3     1
```

d) If $|\hat{S}| = 1$, set $\tau^{(m)} = \hat{S}$, otherwise set:

Phase 2: Parameter-based relabelling

$$\tau^{(m)} = \underset{\tau \in \hat{S}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{r=1}^{R} \left| \frac{\theta^*_{r,k} - \theta^{(m)}_{r,\tau(k)}}{\theta^*_{r,k}} \right|.$$

van Havre, Z., N. White, J. Rousseau, and K. Mengersen. 2015. Overfitting Bayesian mixture models with an unknown number of components. *PLoS ONE* **10** (7): e0131739. doi: 10.1371/journal.pone.0131739.

- Zswitch is very accurate and for $K < 5$, very efficient.
- However, it requires a tuning parameter, ω.
  - Smaller ω increases accuracy (more reliance on phase 2) but also computation time.
  - Larger ω decreases computation time, but it can result in set I_j being empty.
- Additionally, the storage and computation of $\hat{S}$ can become prohibitive for large K, especially when the components overlap ($\hat{S}$ approaches S)
  - E.g. for $K = 100$, this easily exceeds 1000GB of RAM for 1 iteration!

- Zswitch 2 improves Zswitch in two main ways.
  - Convert phase 2 relabelling strategy into LSAP costs:

$$c_{j,k}^{(m)} = \sum_{r=1}^{R} \left| \frac{\theta_{r,k}^{*} - \theta_{r,j}^{(m)}}{\theta_{r,k}^{*}} \right|.$$

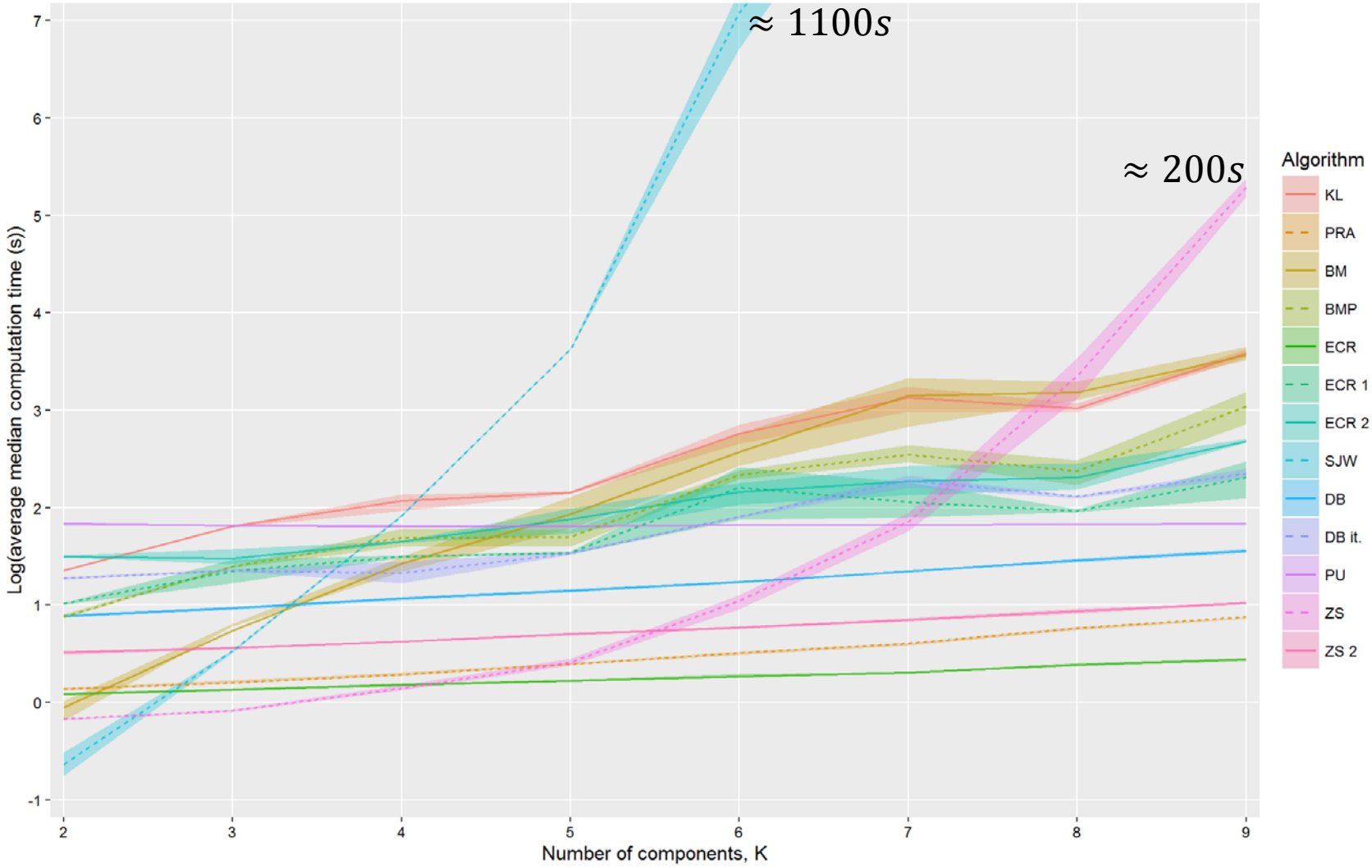
  - Combine this with the ideas of the phase 1 relabelling strategy and tuning parameter by constructing the matrix $\mathbf{M}$ exactly as before, and modifying the costs as

$$c_{j,k}^{(m)} = \begin{cases} \dfrac{1}{\mathbf{M}_{j,k}} \displaystyle\sum_{r=1}^{R} \left| \frac{\theta_{r,k}^{*} - \theta_{r,j}^{(m)}}{\theta_{r,k}^{*}} \right| & \text{if } \dfrac{\mathbf{M}_{j,k}}{\sum_{k'=1}^{K} \mathbf{M}_{j,k'}} > \omega \\ \infty & \text{otherwise} \end{cases}.$$

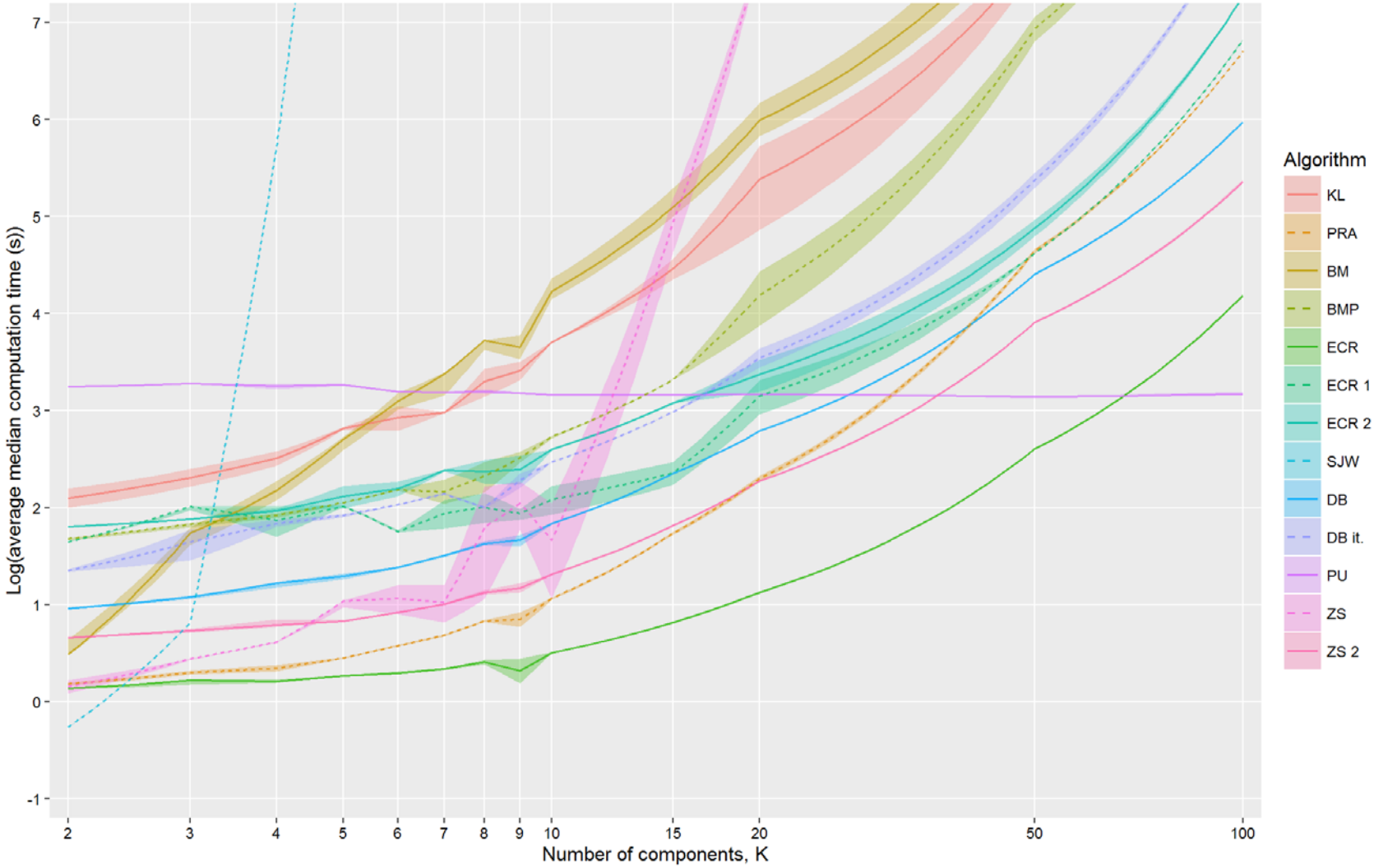
  - This circumvents problems with ω and $\hat{S}$.

- Simulation studies:
  - Poisson, Gaussian, and Gamma mixtures.
  - Test:
    - Computational efficiency (up to $K = 100$)
    - Accuracy
    - Robustness to misspecification of K

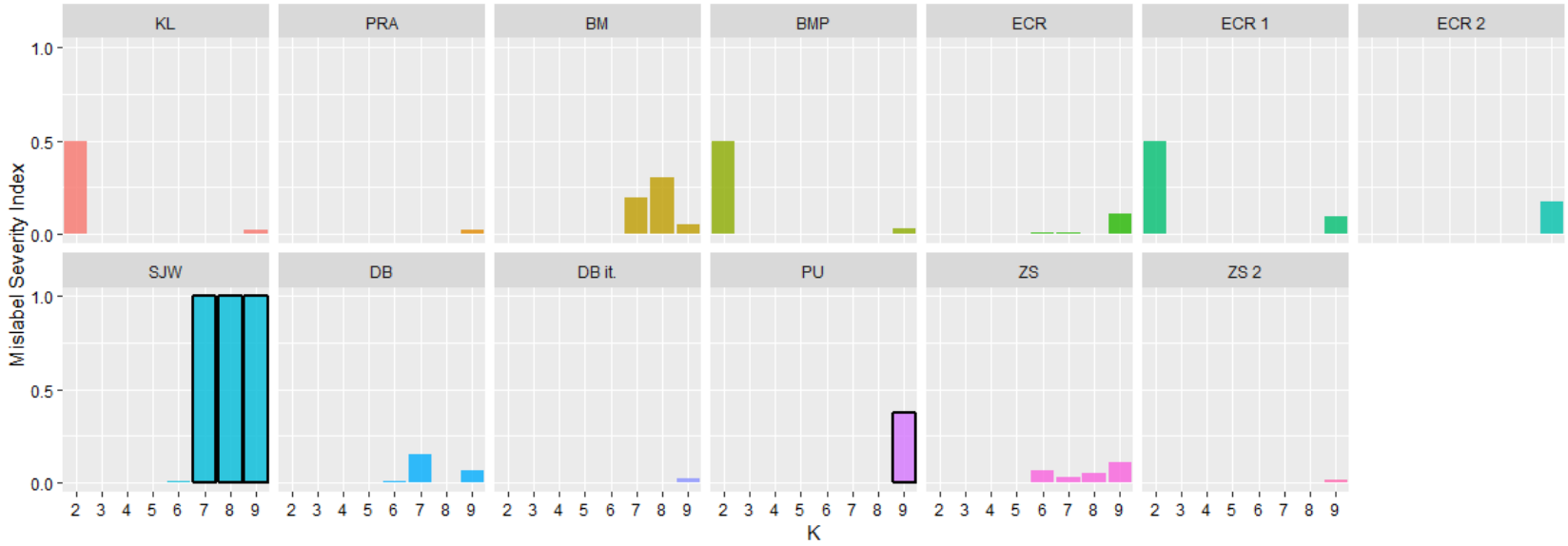
- Efficiency results (Poisson mixture):



- Efficiency results (Gaussian mixture):



- Accuracy results (Poisson mixture):

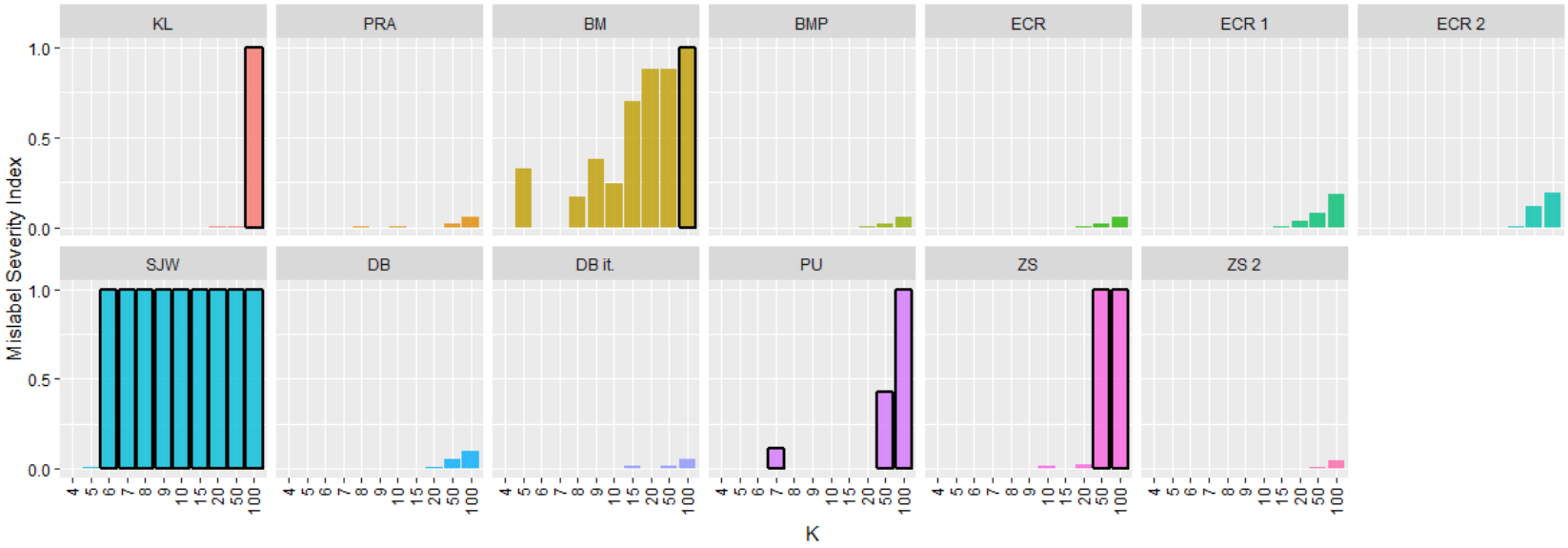


where the mislabel severity index is

$$MSI = 1 - \frac{1}{M}\sum_{m=1}^{M} A^{(m)}$$

and $A^{(m)}$ is the proportion of correct permutation indices.

- Accuracy results (Gaussian mixture):

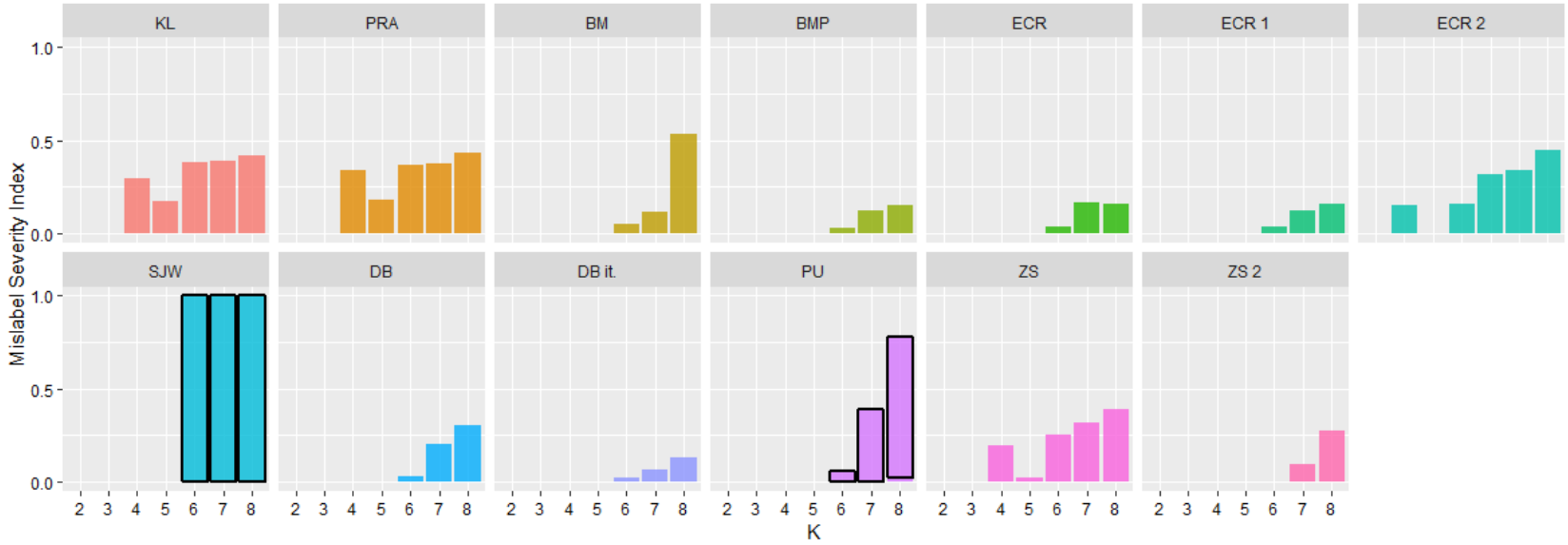


where the mislabel severity index is

$$MSI = 1 - \frac{1}{M}\sum_{m=1}^{M} A^{(m)}$$

and $A^{(m)}$ is the proportion of correct permutation indices.

- Misspecification results (Gamma mixture):



where the mislabel severity index is

$$MSI = 1 - \frac{1}{M}\sum_{m=1}^{M} A^{(m)}$$

and $A^{(m)}$ is the proportion of correct permutation indices.

- The accuracy and computational efficiency of each algorithm can vary substantially.
  - Higher computational cost ≠ higher accuracy
  - Most algorithms perform OK for small K
  - Algorithms that can be formulated as a LSAP are generally fast.
- Zswitch 2 can be viewed as an improvement on PRA and ZS.
  - Improved accuracy and computational efficiency (for large K).

- Future research:
  - Ensemble approach (e.g. PU + ZS 2)
  - Expand review of algorithms (Pan et al. 2015, Yao 2013, …)
  - Expand simulation study (e.g. larger K)

- Key references:

Marin, J-M., K. Mengersen, and C. P. Robert. 2005. "Bayesian modelling and inference on mixtures of distributions" In *Handbook of Statistics* edited C. Rao and D. Dey. New York: Springer-Verlag.

Stephens, M. 2000b. Dealing with label Switching in mixture models. *Journal of the Royal Statistical Society Series B* **62** (4): 795-809. doi: 10.1111/1467-9868.00265

van Havre, Z., N. White, J. Rousseau, and K. Mengersen. 2015. Overfitting Bayesian mixture models with an unknown number of components. *PLoS ONE* **10** (7): e0131739. doi: 10.1371/journal.pone.0131739.

The work presented in this talk is in preparation for submission to the *Journal of the Royal Statistical Society Series B*.

This work also appears as a chapter in my PhD thesis:

Duncan, E. W. 2017. *Bayesian approaches to issues arising in spatial modelling.* PhD by Publication, Queensland University of Technology. URL: https://eprints.qut.edu.au/view/person/Duncan,_Earl.html