∞ Infinite Mixtures of Infinite Factor Analysers ∞

Keefe Murphy[1, 2] Isobel Claire Gormley[1, 2] Cinzia Viroli[3]

[1]School of Mathematics and Statistics, UCD
[2]Insight Centre for Data Analytics, UCD
[3]Department of Statistical Sciences, University of Bologna

keefe.murphy@ucd.ie

Motivation

- Model-based approaches to clustering are well-established methods that uncover sub-groups of observations:
  - **reproducibility** due to statistical modelling basis
  - **objectivity** through access to principled model selection tools
  - **interpretability** through provision of parameter estimates and associated uncertainties
- However, they lose **tractability** when p, the dimension of the feature vectors, is comparable to or even greater than N, the number of observations
- Typical issues include handling large covariance matrices, optimisation issues, run times, selecting the number of clusters, ...

Methods for clustering big p data

- Bouveyron & Brunet-Saumard (2014) provide a synoptic overview
- **Classical distance-based**: e.g. k-means scale relatively well to large p, but focus on detecting differences in mean signals
- **Dimension reduction + clustering**: caution is required[1] but typically computationally cheap[2]
- **Regularisation**: eases covariance matrix inversion[3]

[1] Chang (1983)
[2] Rahman & Johnson (2018); Taschler et al. (2019)
[3] Fraley & Raftery (2007)

Methods for clustering big p data

- **Penalisation**: lasso-like method[4]
- **Hybrids**: estimate some parameters, but 'avoid' the covariance matrix[5]
- **Parsimonious mixtures**: Gaussian mixture models in `mclust`[6]
- **Subspace clustering**: exploit the 'simple structure' phenomenon and model data in low-dimensional subspaces, e.g. `pgmm`[7]

[4] Zhou et al. (2009); Städler et al. (2017)
[5] Cai et al. (2019)
[6] Scrucca et al. (2016)
[7] Murtagh (2009); McNicholas et al. (2018)

Motivation

- We focus on factor-analytic Gaussian mixture models as a subspace clustering method
- Typically, the numbers of clusters G and subspace dimension q are specified in advance of model fitting, and remain fixed
- The pair (G, q) which optimises some model selection criterion (which one?) is usually chosen
- As the model search space can become vast, models in which $q_g \neq q_{g'}$ are rarely considered

Motivation

- We introduce the *family of infinite mixtures of infinite factor analysers models*, in particular the flagship **IMIFA** model
- Bayesian nonparametric approach to both clustering and dimension reduction
- Facilitates automatic inference on the number of clusters G and the numbers of *cluster-specific latent factors* q_g
- Advantages:
  - flexible
  - computationally efficient
  - enables uncertainty quantification
  - removes reliance on model selection criteria

Agenda

- Mixtures of Factor Analysers (**MFA**)
- Mixtures of **Infinite** Factor Analysers (**MIFA**)
- Overfitted Mixtures of **Infinite** Factor Analysers (**OMIFA**)
- **Infinite** Mixtures of **Infinite** Factor Analysers (**IMIFA**)
- Examples & Results
- Discussion

(Finite) Mixtures of (Finite) Factor Analysers (MFA)

- **MFA** is a Gaussian latent variable model, simultaneously achieving dimension reduction and clustering in high-dimensional data settings

- Supposes p-dimensional feature vector $\mathbf{x}_i \; \forall \; i = 1, \ldots, N$ arises with probability $\pi_g \; \forall \; g = 1, \ldots, G$ from a *cluster-specific* **FA** model:

$$\mathbf{x}_i - \boldsymbol{\mu}_g \;=\; \boldsymbol{\Lambda}_g \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_{ig}$$

(Finite) Mixtures of (Finite) Factor Analysers (MFA)

- **MFA** is a Gaussian latent variable model, simultaneously achieving dimension reduction and clustering in high-dimensional data settings

- Supposes p-dimensional feature vector $\mathbf{x}_i \; \forall \; i = 1, \ldots, N$ arises with probability $\pi_g \; \forall \; g = 1, \ldots, G$ from a *cluster-specific* **FA** model:

$$\mathbf{x}_i - \boldsymbol{\mu}_g \;=\; \boldsymbol{\Lambda}_g \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_{ig}$$

where:

$\boldsymbol{\mu}_g =$ mean vector for cluster g

$\boldsymbol{\eta}_i = q$-vector of unobserved latent factors where $0 \leq q \ll p$

$\boldsymbol{\Lambda}_g = p \times q$ loadings matrix for cluster g

$\boldsymbol{\epsilon}_{ig} =$ error vector $\sim \mathrm{N}_p\left(\mathbf{0}, \boldsymbol{\Psi}_g\right)$

$\boldsymbol{\Psi}_g =$ diagonal matrix of non-zero uniquenesses for cluster g

$\boldsymbol{\pi} =$ cluster mixing proportions

(Finite) Mixtures of (Finite) Factor Analysers (MFA)

- A latent indicator $\mathbf{z}_i$ is introduced, s.t.

$$z_{ig} = \begin{cases} 1 & \text{if } i \in \text{cluster } g \\ 0 & \text{otherwise} \end{cases}$$

- Provides parsimonious covariance structure:

$$f\left(\mathbf{x}_i \,|\, \boldsymbol{\theta}\right) = \sum_{g=1}^{G} \pi_g \mathrm{N}_p \left(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g^{\top} + \boldsymbol{\Psi}_g\right)$$

- Flat conditionally conjugate priors assumed:
  facilitates MCMC sampling via Gibbs updates

- Isotropic constraint: $\boldsymbol{\Psi}_g = \psi_g \boldsymbol{\mathcal{I}}_p$ provides link to **MPPCA**[8]

[8] Tipping & Bishop (1999)

Prior Specification

- Assume:

$$
\begin{aligned}
(\lambda_{j1g}, \ldots, \lambda_{jqg}) &\sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\mathcal{I}}_q) \quad \text{(for now)} \\
\boldsymbol{\eta}_i &\sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\mathcal{I}}_q) \\
(\pi_1, \ldots, \pi_G) &\sim \mathrm{Dir}(\boldsymbol{\alpha} = \alpha, \ldots, \alpha) \quad (\alpha = 1 \text{ for now}) \\
\mathbf{z}_i &\sim \mathrm{Mult}(1, \boldsymbol{\pi}) \\
\boldsymbol{\mu}_g &\sim \mathrm{N}_p(\tilde{\boldsymbol{\mu}}, \varphi^{-1}\boldsymbol{\mathcal{I}}_p) \\
\psi_{jg}^{-1} &\sim \mathrm{IG}(\alpha_0, \beta_j)
\end{aligned}
$$

- $\tilde{\boldsymbol{\mu}}$ is the overall sample mean & the scalar φ controls the level of diffusion

- The variable-specific scales β_j are derived from the (estimated) sample precision matrix[9]

[9] Früwirth-Schnatter & Lopes (2010)

(Finite) Mixtures of **Infinite** Factor Analysers (MIFA)

- Allows each Λ_g matrix to theoretically have infinitely many factors, using a multiplicative gamma process shrinkage prior[10] (MGP)

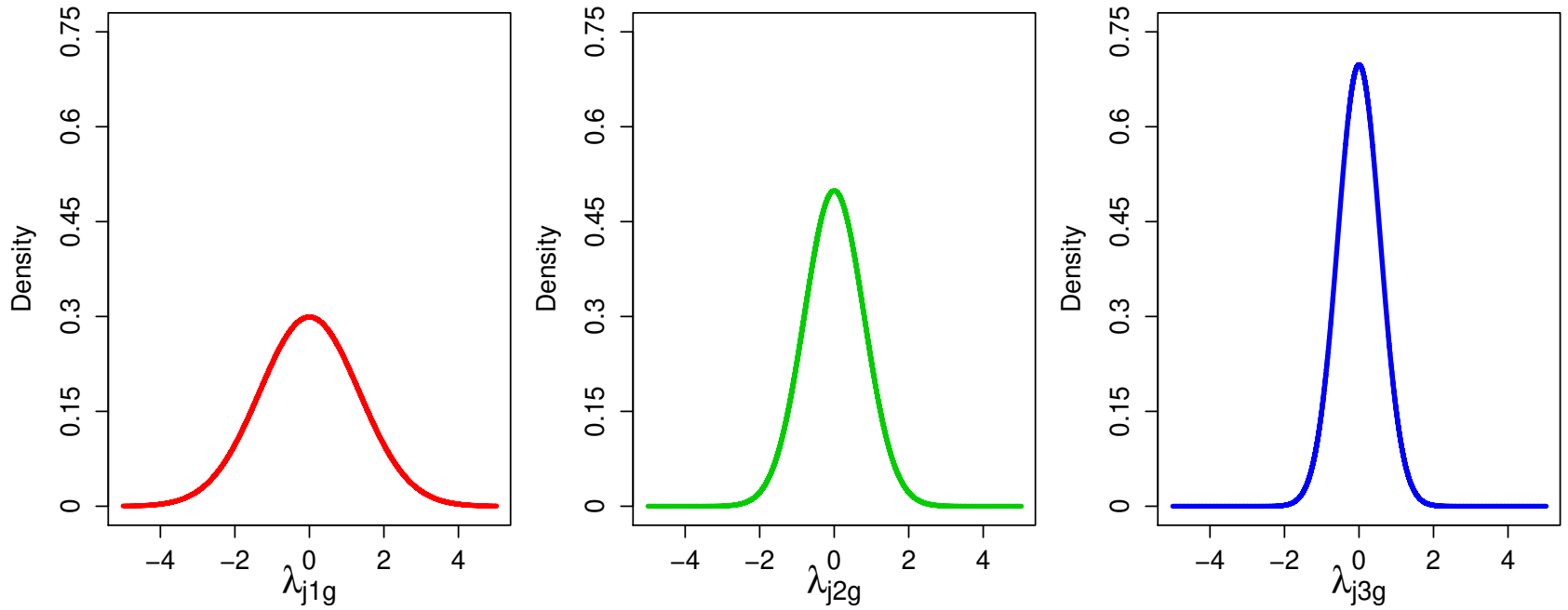
$$\begin{aligned}
\textbf{Loadings}: \quad & \lambda_{jkg} \sim \mathrm{N}\left(0, \phi_{jkg}^{-1}\tau_{kg}^{-1}\sigma_g^{-1}\right) \\
\textbf{Local Shrinkage}: \quad & \phi_{jkg} \sim \mathrm{Ga}\left(\nu_1, \nu_2\right) \\
\textbf{Column Shrinkage}: \quad & \tau_{kg} = \prod_{h=1}^{k} \delta_{hg} \quad \delta_{1g} \sim \mathrm{Ga}\left(\alpha_1, 1\right) \quad \delta_{hg} \sim \mathrm{Ga}\left(\alpha_2, 1\right) \ \forall\ h \geq 2 \\
\textbf{Cluster Shrinkage}: \quad & \sigma_g \sim \mathrm{Ga}\left(\varrho_1, \varrho_2\right)
\end{aligned}$$

- Increasingly shrinks loadings toward zero as column index $k \to \infty$ under certain hyperparameter settings[11]

- Conditional conjugacy facilitates block updates of the loadings matrices

[10] Bhattacharya & Dunson (2011)
[11] Durante (2017)

- Prior distribution of loadings in *1st*, *2nd* & *3rd* columns of typical $\mathbf{\Lambda}_g$ matrix



- **MIFA** allows different clusters be modelled by different numbers of factors

- **MIFA** significantly reduces model search to one for G only, as q_g is estimated automatically during model fitting

Adaptive Gibbs Sampler

- Used to automatically truncate infinite loadings matrices and estimate q_g
- Don't want to 'lose' important factors;
  don't want the computational burden of redundant factors
- Check which loadings columns have some proportion of elements within some small neighbourhood of 0
- If none, simulate new columns from the $\mathbf{MGP}$ prior and new scores $\boldsymbol{\eta}$; otherwise, discard redundant columns
- Decrease adaptation frequency exponentially fast after burn-in.
- Use the modal number of effective factors in each cluster as the $\widehat{q}_g$ estimate

Other Practical Considerations...

- **Identifiability**: rotational invariance addressed via Procrustean methods[12]; each sampled $\mathbf{\Lambda}_g$ mapped to a template matrix at burn-in to ensure sensible posterior means
- **Label switching**: addressed offline by also mapping $\mathbf{z}_i$ samples to a template, using the square-assignment algorithm[13]
- **Model selection**: optimal G chosen via $\mathbf{BICM}$[14], which is particularly useful for nonparametric models where the number of 'free' parameters is hard to quantify

[12] Ghosh & Dunson (2008)
[13] Carpaneto & Toth (1980)
[14] Raftery et al. (2007)

Overfitted Mixtures of **Infinite** Factor Analysers (**OMIFA**)

- Overfitted mixtures[15] obviate the need to choose the optimal G, as a simple alternative to transdimensional MCMC methods
- Papastamoulis (2018) simultaneously proposed an overfitted mixture of finite factor analysers (**OMFA**), which is a member of the **IMIFA** family
- Initially overfit the number of clusters expected to be present and estimate G by the number of non-empty clusters visited most often
- Prior on the mixing proportions plays an important role: small values of the Dirichlet hyperparameter α encourage emptying excess components
- Following Früwirth-Schnatter and Malsiner-Walli (2019), assume a 'sparse' Gamma hyperprior for α, allowing it to be learned
- Employing the (**MGP**) prior on the infinite loadings matrices and modifying the adaptation to account for empty components gives rise to **OMIFA**

[15] Rousseau & Mengersen (2011); van Havre et al. (2015)

Infinite Mixtures of **Infinite** Factor Analysers (**IMIFA**)

- Infinite mixture models are another approach to automating estimation of G
- **IMIFA** employs a nonparametric Pitman-Yor process prior **PYP** and is thus a **PYP-MGP** mixture model:

$$
\begin{aligned}
(\mathbf{x}_i \,|\, i \in g, \boldsymbol{\theta}_g) &\sim f(\mathbf{x}_i; \boldsymbol{\theta}_g) \\
\boldsymbol{\theta}_g &\sim H \\
H &\sim \mathbf{PYP}(\alpha, d, H_0)
\end{aligned}
$$

where $\boldsymbol{\theta}_g =$ cluster-specific parameters $\{\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g\}$

$\alpha =$ concentration parameter

$d =$ discount parameter $\in [0, 1)$ s.t. $\alpha > -d$

$H_0 =$ base distribution

- When $d = 0$, the **PYP** reduces to the Dirichlet process (**DP**)

- **Stick-Breaking**[16] Representation:

$$\upsilon_g \sim \text{Beta}(1-d, \alpha+gd), \quad \boldsymbol{\theta}_g \sim H_0, \; f\left(\mathbf{x}_i \,|\, \boldsymbol{\theta}\right) = \sum_{g=1}^{\infty} \pi_g \; \text{N}_p\left(\mathbf{x}_i; \boldsymbol{\theta}_g\right)$$

$$\pi_g = \upsilon_g \prod_{l=1}^{g-1}(1-\upsilon_l), \qquad H = \sum_{g=1}^{\infty} \pi_g \; \delta_{\boldsymbol{\theta}_g} \quad \sim \textbf{PYP}\left(\alpha, d, H_0\right)$$

[16] Pitman (1996)
[17] Kalli et al. (2011)
[18] Papaspiliopoulos & Roberts (2008)

Sampling Infinite Mixture Models

- **Stick-Breaking**[16] Representation:

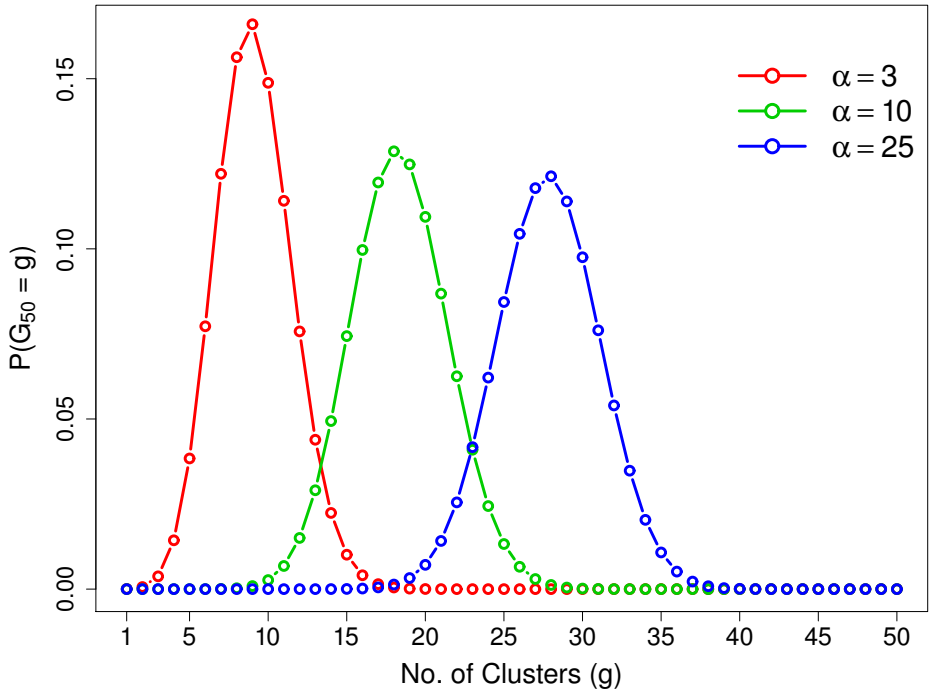
$$\upsilon_g \sim \text{Beta}(1-d, \alpha+gd), \quad \boldsymbol{\theta}_g \sim H_0, \; f\left(\mathbf{x}_i \,|\, \boldsymbol{\theta}\right) = \sum_{g=1}^{\infty} \pi_g \, \text{N}_p\left(\mathbf{x}_i; \boldsymbol{\theta}_g\right)$$

$$\pi_g = \upsilon_g \prod_{l=1}^{g-1}(1-\upsilon_l), \qquad H = \sum_{g=1}^{\infty} \pi_g \, \delta_{\boldsymbol{\theta}_g} \quad \sim \textbf{PYP}\left(\alpha, d, H_0\right)$$

- 'Independent' **Slice-Efficient Sampler**[17] introduces an auxiliary variable s.t. $\mathbf{x}_i \,|\, u_i$ can be written as a finite mixture model
- Facilitates adaptively truncating sufficient number of **'active' components** needed to be sampled at each iteration
- **Label-switching moves**[18] incorporated in order to improve mixing due to highly multimodal state spaces

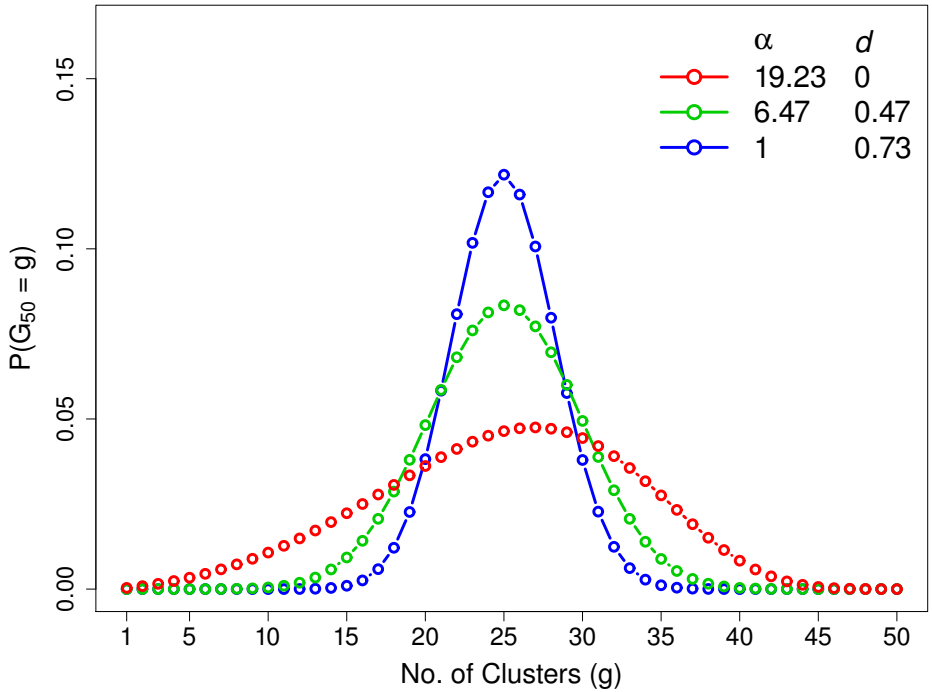
[16] Pitman (1996)
[17] Kalli et al. (2011)
[18] Papaspiliopoulos & Roberts (2008)

Dirichlet Processes (DP)



- Under the **DP**, $v_g \sim \text{Beta}(1, \alpha)$
- Mass shifts to the right with increasing dispersion as α increases
- Reliable prior information on the number of clusters is required; the high-peakedness of the distributions prevents the wrong prior information from being overruled

Pitman-Yor Processes (PYP)



- Under the **PYP**, $v_g \sim \text{Beta}(1-d, \alpha+gd)$: $d > 0$ obtains heavy-tailed less informative prior with no tractability sacrifices

- A joint hyperprior of the form $\text{p}(\alpha, d) = \text{p}(d)\,\text{p}(\alpha \mid d)$ is assumed:
  - $(\alpha \mid d) \sim \text{Ga}(\alpha + d \mid a, b), \qquad \alpha + d \in (-d, \infty)$
  - Spike-and-slab prior $d \sim \kappa\delta_0 + (1-\kappa)\,\text{Beta}(d \mid a', b')$ used to assess whether data arose from **DP** or **PYP** at little extra computational cost

- $f(\mathbf{X}, \boldsymbol{\eta}, \mathbf{Z}, \mathbf{u}, \boldsymbol{\Upsilon}, \boldsymbol{\theta}) \propto f(\mathbf{X} \,|\, \boldsymbol{\eta}, \mathbf{Z}, \mathbf{u}, \boldsymbol{\Upsilon}, \boldsymbol{\theta}) f(\boldsymbol{\eta}) f(\mathbf{Z}, \mathbf{u} \,|\, \boldsymbol{\Upsilon}, \boldsymbol{\pi}) f(\boldsymbol{\Upsilon} \,|\, \alpha, d) f(\boldsymbol{\theta})$

$$= \left\{ \prod_{i=1}^{N} \prod_{g \in \mathcal{A}_\xi(u_i)} \mathrm{N}_p(\mathbf{x}_i; \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \boldsymbol{\eta}_i, \boldsymbol{\Psi}_g)^{z_{ig}} \right\} \left\{ \prod_{i=1}^{N} \mathrm{N}_q(\boldsymbol{\eta}_i; \mathbf{0}, \boldsymbol{\mathcal{I}}_q) \right\}$$

$$\left\{ \prod_{i=1}^{N} \prod_{g=1}^{\infty} \left(\frac{\pi_g}{\xi_g} \mathbb{1}\left(u_i < \xi_g\right) \right)^{z_{ig}} \right\} \left\{ \prod_{g=1}^{\infty} \frac{(1-\upsilon_g)^{\alpha+gd-1}}{\upsilon_g^d \, \mathrm{B}(1-d, \alpha+gd)} \right\} f(\boldsymbol{\theta})$$

  where $f(\boldsymbol{\theta})$ is the product of the conjugate priors

- Only the 'active' components need to be sampled from at each iteration

- Allocations sampled efficiently, accounting for tiny z_{ig} samples, using the Gumbel-Max trick[19]

- True G estimated by number of non-empty clusters visited most often; cluster-specific inference conducted only on those visits

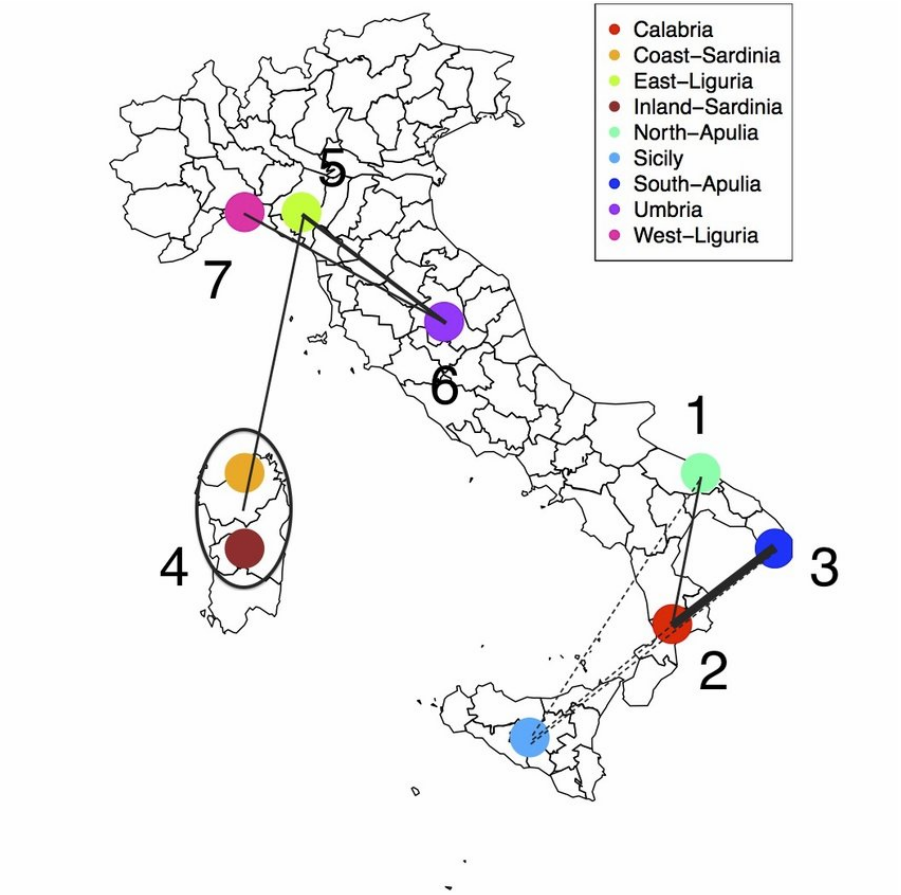
[19] Yellott (1977)

The **IMIFA** family of models

	$\mathbf{Q} = \infty$	$\mathbf{Q} < \infty$
$G = \infty$	IMIFA	IMFA
$G < \infty$	OMIFA	OMFA
$G < \infty$	MIFA	MFA
$G = 1$	IFA	FA

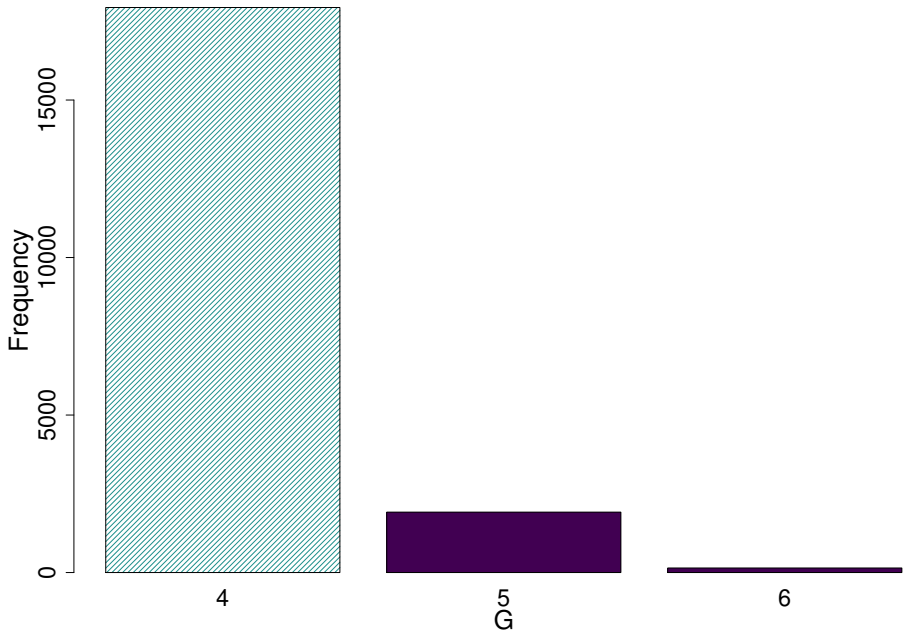
Italian olive oils[20]

- Data on %-composition of 8 fatty acids found in 572 Italian olive oils
- Oils are from 3 *areas*, Sardinia, Southern Italy and Northern Italy, composed of 9 *regions*:



[20] Forina et al. (1983)

- Fit **IMIFA** with $50,000$ iterations, $10,000$ burn-in, and every 2^{nd} iteration thinned



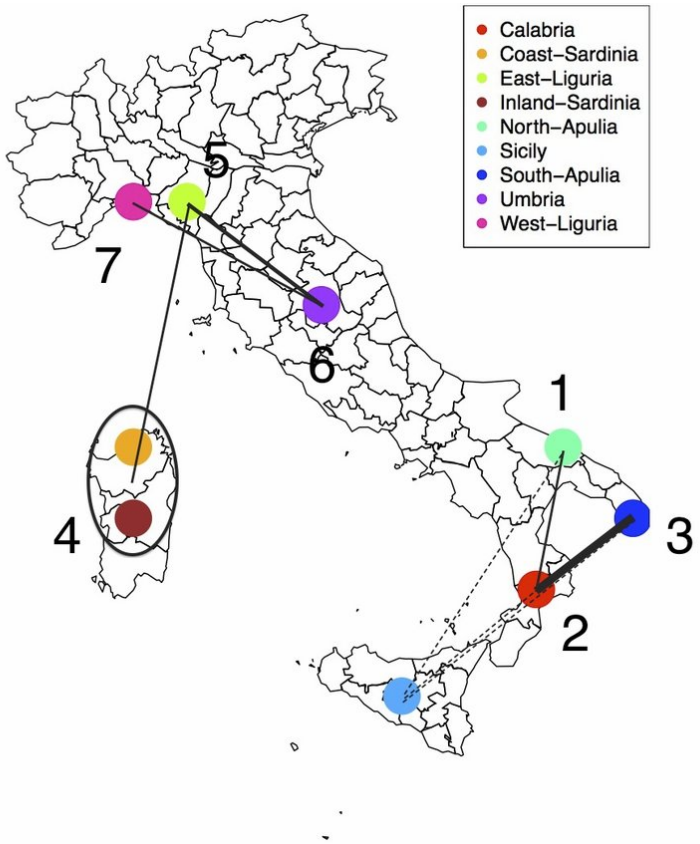
Adj. Rand = 0.9371

	1	2	3	4
South	323	0	0	0
Sardinia	0	98	0	0
North	0	0	103	48

Italian olive oils and the IMIFA model

Adj. Rand = 0.9371

	1	2	3	4
South	323	0	0	0
Sardinia	0	98	0	0
North	0	0	103	48



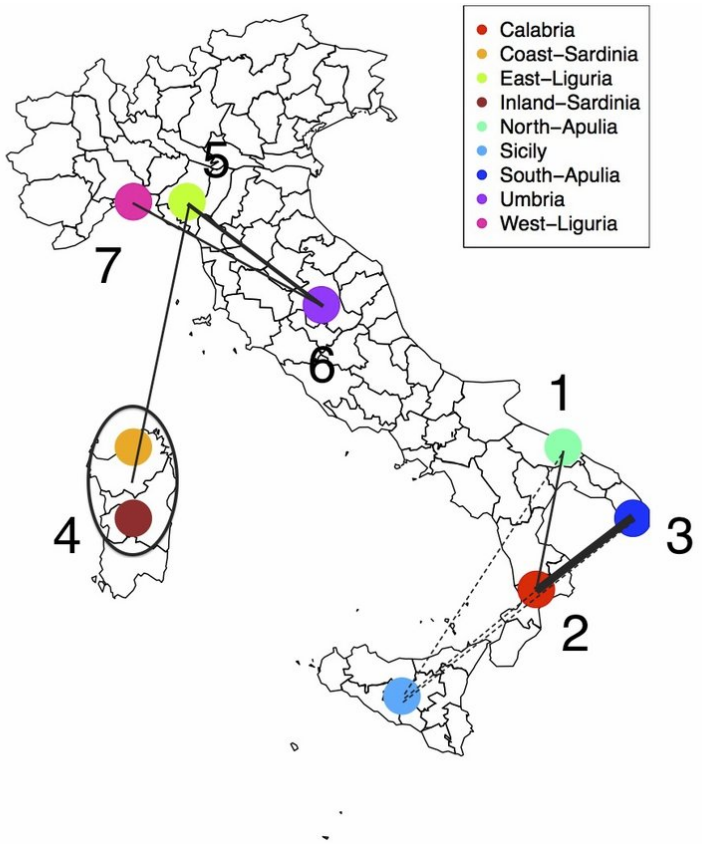
Adj. Rand = 0.9943

	1	2	3	4
South	323	0	0	0
Sardinia	0	98	0	0
Liguria	0	0	100	0
Umbria	0	0	3	48

Italian olive oils and the IMIFA model

Adj. Rand = 0.9371

	1	2	3	4
South	323	0	0	0
Sardinia	0	98	0	0
North	0	0	103	48

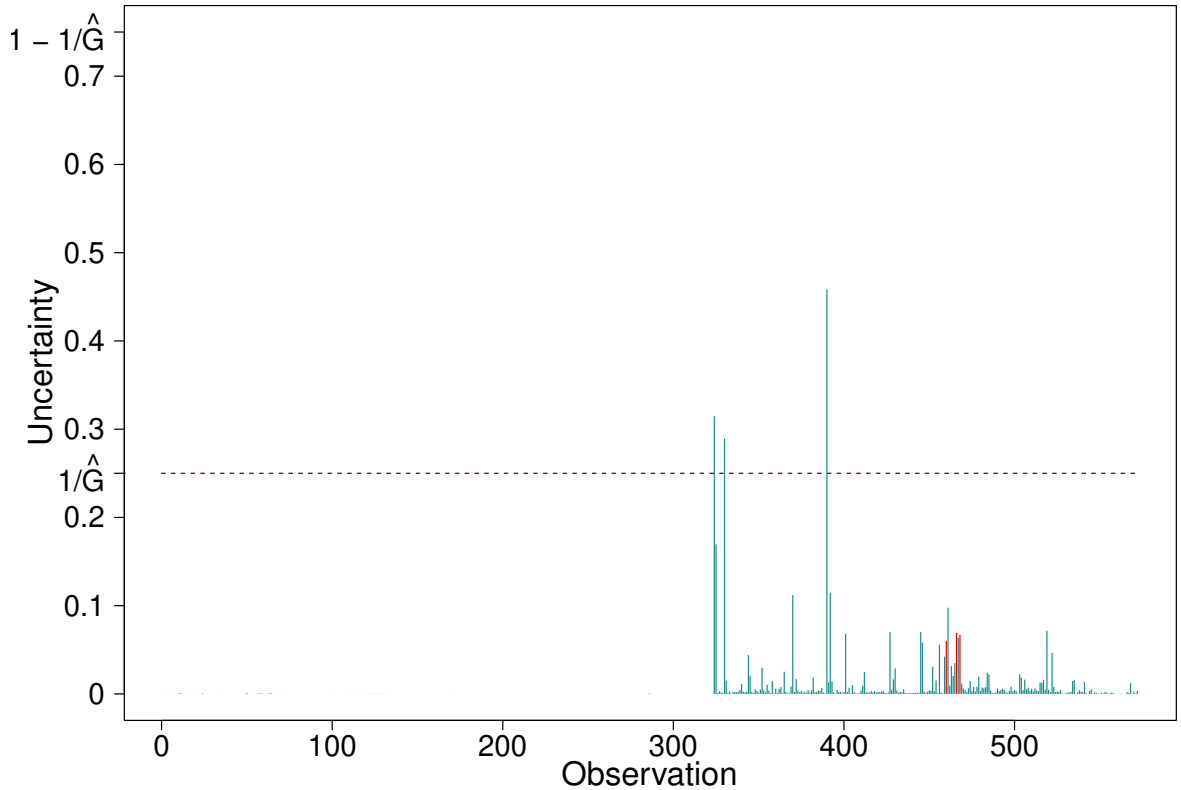


Adj. Rand = 0.9943

	1	2	3	4
South	323	0	0	0
Sardinia	0	98	0	0
Liguria	0	0	100	0
Umbria	0	0	3	48

- $\hat{q}_1 = 6\ [5, 6],\ \hat{q}_2 = 3\ [1, 6],$ $\hat{q}_3 = 6\ [3, 6],\ \hat{q}_4 = 2\ [1, 4].$
- Large Southern Italy cluster requires large number of factors.
- Flexibility to model other clusters with less factors.

- Clustering uncertainties: $\widehat{U}_i = \min_{g \in \{1,\ldots,\widehat{G}\}} \{1 - \widehat{z}_{ig}\}$



Italian olive oils and the **IMIFA** model family

- Consider other members of the **IMIFA** family, with $G = 1, \ldots, 9$, $q = 0, \ldots, 6$ and **BICM** used for model selection where necessary.

Italian olive oils and the **IMIFA** model family

- Consider other members of the **IMIFA** family, with $G = 1, \ldots, 9$, $q = 0, \ldots, 6$ and **BICM** used for model selection where necessary.

Model	# Models	Rel. Time	α	d	G	$\mathbf{Q}$	Adj. Rand
IMIFA	1	1.00	0.48	0.01	4	6, 3, 6, 2	0.94
IMFA	7	4.14	0.62	0.01	5	6, 6, 6, 6, 6	0.91

Italian olive oils and the **IMIFA** model family

- Consider other members of the **IMIFA** family, with $G = 1, \ldots, 9$, $q = 0, \ldots, 6$ and **BICM** used for model selection where necessary.

Model	# Models	Rel. Time	α	d	G	$\mathbf{Q}$	Adj. Rand
IMIFA	1	1.00	0.48	0.01	4	6, 3, 6, 2	0.94
IMFA	7	4.14	0.62	0.01	5	6, 6, 6, 6, 6	0.91
OMIFA	1	1.19	0.02	–	4	6, 3, 6, 4	0.93
OMFA	7	5.11	0.02	–	5	6, 6, 6, 6, 6	0.85

- Consider other members of the **IMIFA** family, with $G = 1, \ldots, 9$, $q = 0, \ldots, 6$ and **BICM** used for model selection where necessary.

Model	# Models	Rel. Time	α	d	G	$\mathbf{Q}$	Adj. Rand
IMIFA	1	1.00	0.48	0.01	4	6, 3, 6, 2	0.94
IMFA	7	4.14	0.62	0.01	5	6, 6, 6, 6, 6	0.91
OMIFA	1	1.19	0.02	–	4	6, 3, 6, 4	0.93
OMFA	7	5.11	0.02	–	5	6, 6, 6, 6, 6	0.85
MIFA	9	3.41	1	–	5	6, 3, 6, 6, 4	0.92
MFA	63	13.86	1	–	2	5, 5	0.82

Italian olive oils and the **IMIFA** model family

- Consider other members of the **IMIFA** family, with $G = 1, \ldots, 9$, $q = 0, \ldots, 6$ and **BICM** used for model selection where necessary.

Model	# Models	Rel. Time	α	d	G	$\mathbf{Q}$	Adj. Rand
IMIFA	1	1.00	0.48	0.01	4	6, 3, 6, 2	0.94
IMFA	7	4.14	0.62	0.01	5	6, 6, 6, 6, 6	0.91
OMIFA	1	1.19	0.02	–	4	6, 3, 6, 4	0.93
OMFA	7	5.11	0.02	–	5	6, 6, 6, 6, 6	0.85
MIFA	9	3.41	1	–	5	6, 3, 6, 6, 4	0.92
MFA	63	13.86	1	–	2	5, 5	0.82

- Optimal models chosen by **BICM** were not all optimal in a clustering sense: candidate $G = 4$ **MIFA** model yields Adj. Rand $= 0.94$.

- Fully automatic **IMIFA** requires one quick run, gives optimal clustering performance, and does not rely on model selection criteria.

- Comparison to other state-of-the-art methods:
  - `mclust`: mixture of Gaussians (Scrucca et al (2017))
  - MFMA: mixture of factor mixture analysers (Viroli (2010))
  - `pgmm`: parsimonious Gaussian mixture models (McNicholas et al. (2018))

- Comparison to other state-of-the-art methods:
  - `mclust`: mixture of Gaussians (Scrucca et al (2017))
  - MFMA: mixture of factor mixture analysers (Viroli (2010))
  - `pgmm`: parsimonious Gaussian mixture models (McNicholas et al. (2018))

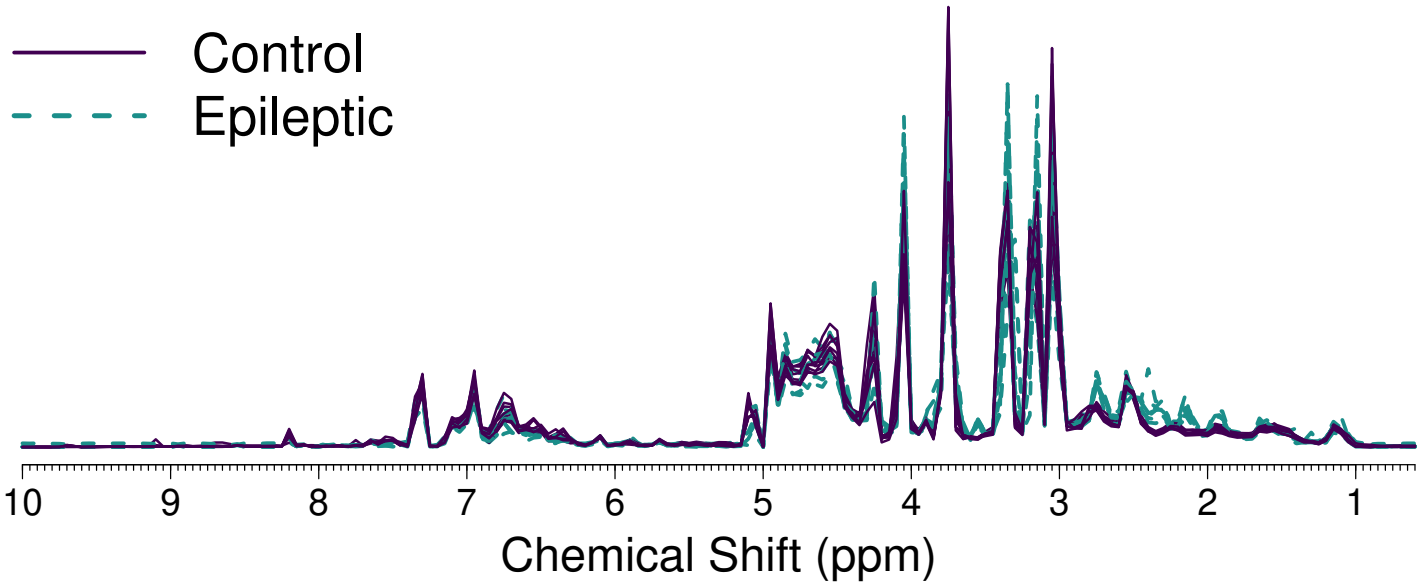
Model	# Models	Rel. Time	α	d	G	$\mathbf{Q}$	Adj. Rand
IMIFA	1	1.00	0.48	0.01	4	6, 3, 6, 2	0.94

- Comparison to other state-of-the-art methods:
  - `mclust`: mixture of Gaussians (Scrucca et al (2017))
  - MFMA: mixture of factor mixture analysers (Viroli (2010))
  - `pgmm`: parsimonious Gaussian mixture models (McNicholas et al. (2018))

Model	# Models	Rel. Time	α	d	G	$\mathbf{Q}$	Adj. Rand
IMIFA	1	1.00	0.48	0.01	4	6, 3, 6, 2	0.94
`mclust`	115	0.01	–	–	6	–	0.56
MFMA	1350	4.68	–	–	4	5, 5, 5, 5	0.68
`pgmm`	588	4.46	–	–	5	6, 6, 6, 6, 6	0.53

Spectral Metabolomic Data[21]

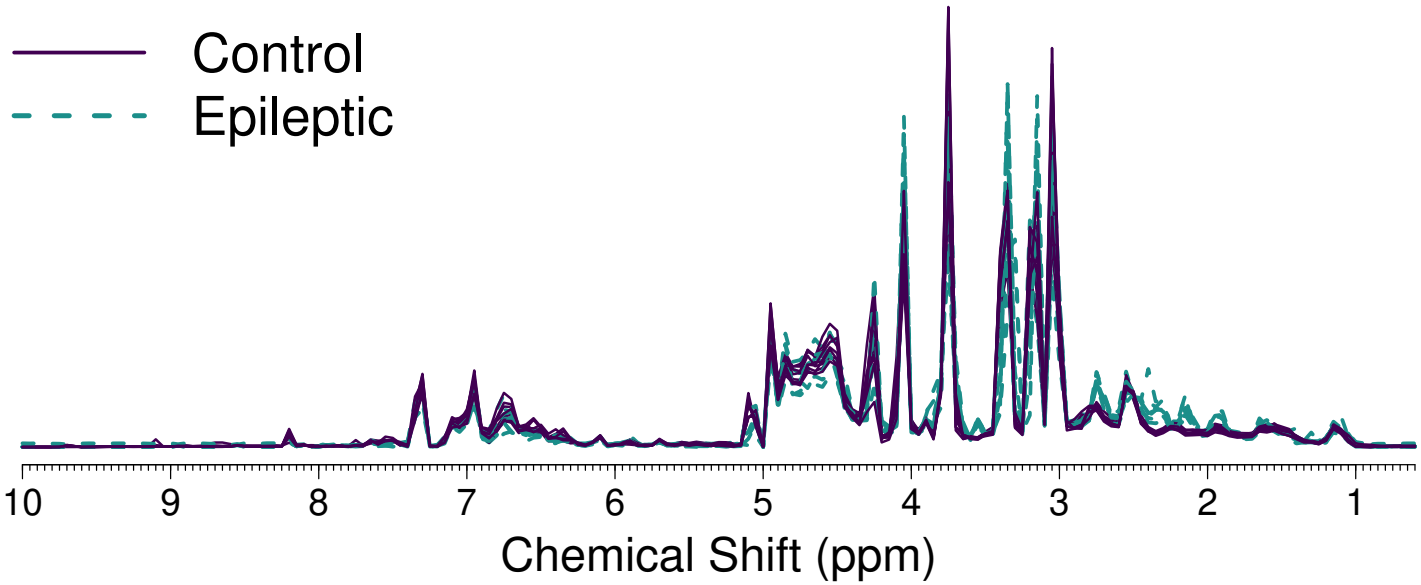
- Urine samples of $N = 18$ subjects; half have epilepsy, half are controls
- NMR spectra with $p = 189$ peaks ($N \ll p$).



[21] Carmody et al. (2010)

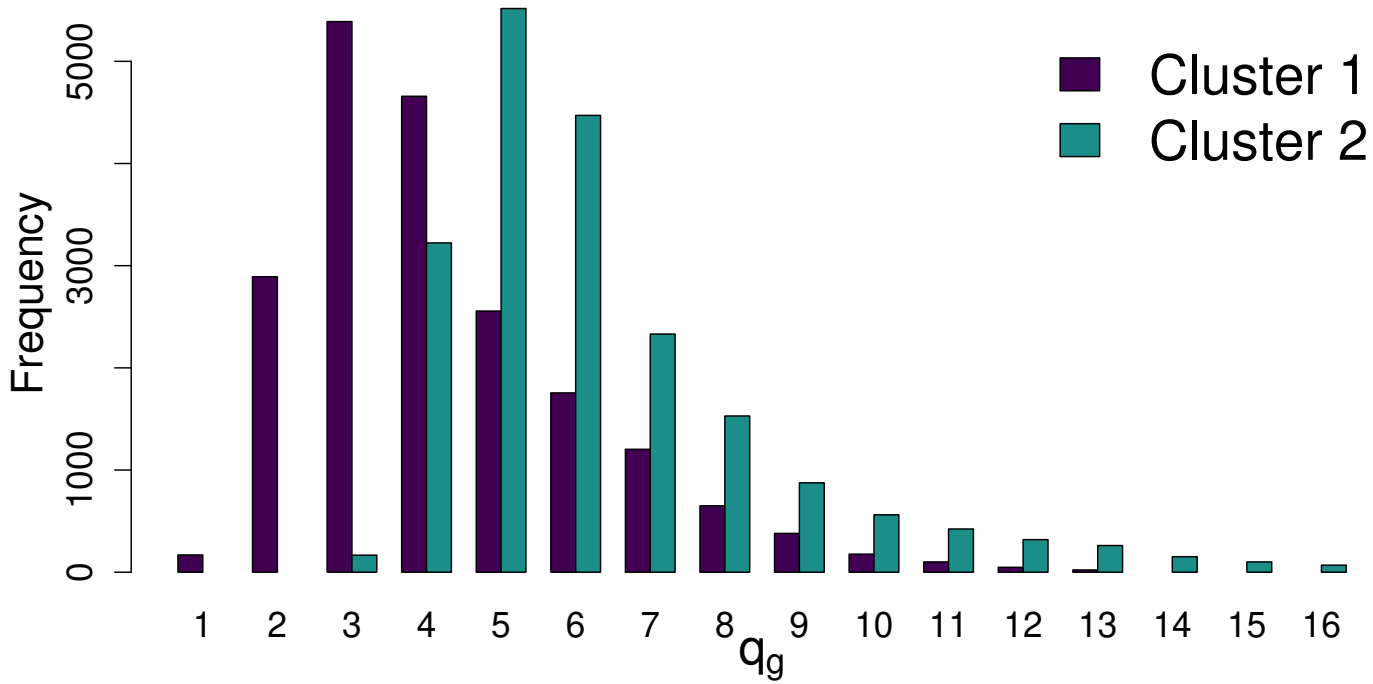
Spectral Metabolomic Data[21]

- Urine samples of $N = 18$ subjects; half have epilepsy, half are controls
- NMR spectra with $p = 189$ peaks ($N \ll p$).

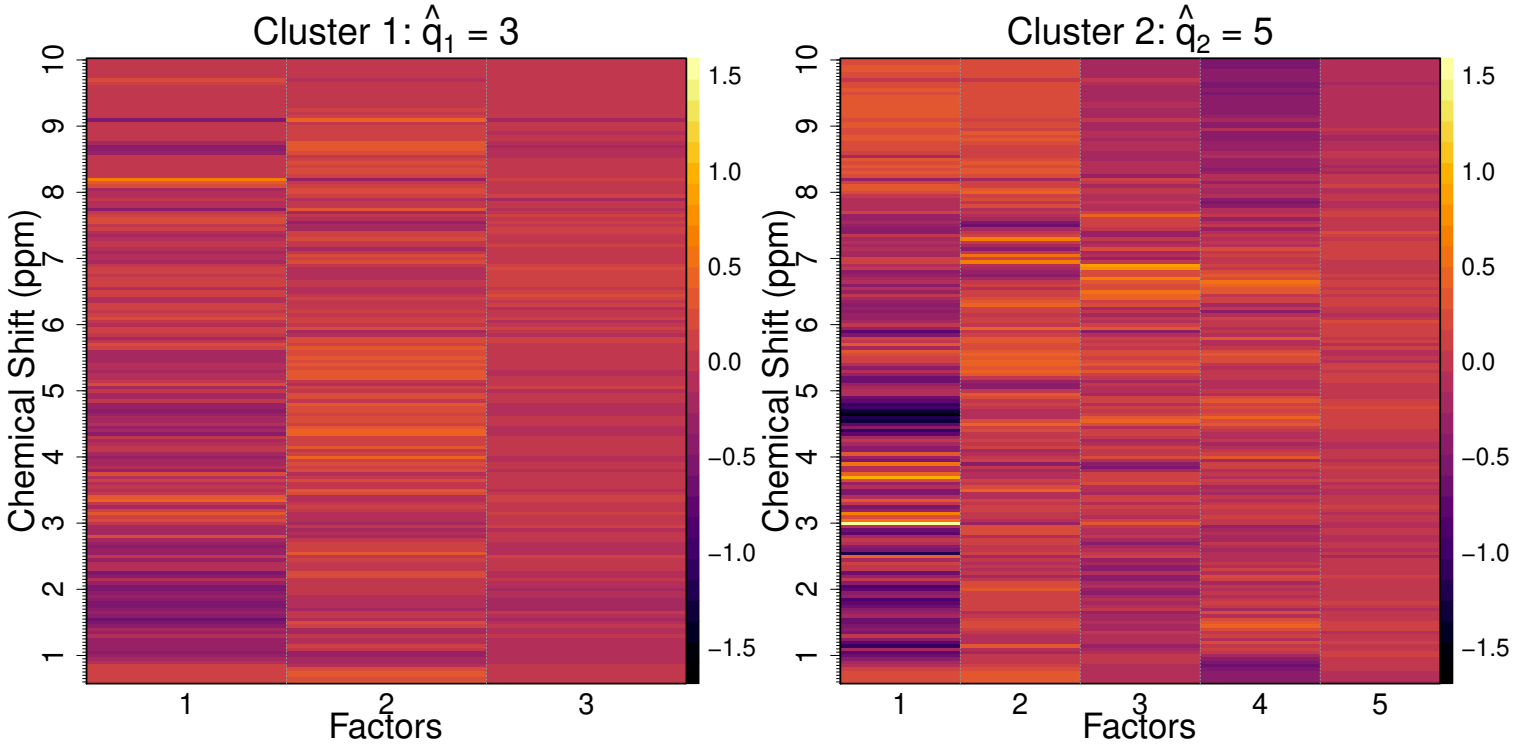


- **MFA** for $G = 2$ and $q = 0, \ldots, 10$:
  $\widehat{G} = 2$ and $\widehat{q} = 3$, with 4 subjects clustered 'incorrectly'
- **MIFA** for $G \in \{1, \ldots, 5\}$: $\widehat{G} = 2$ is optimal, 1 subject clustered 'incorrectly'

[21] Carmody et al. (2010)



- **IMIFA** finds $\widehat{G} = 2$, with Adj. Rand$= 1$
- 95% CI: $\widehat{q}_1 = 3$ $[2, 9]$ and $\widehat{q}_2 = 5$ $[4, 13]$
- Cluster 2 captures the epileptic group: more complex model required



- Elevated loadings in cluster 2 for the first two factors for chemical shift values between 8 and 10
- This activity is not present for other factors in either cluster

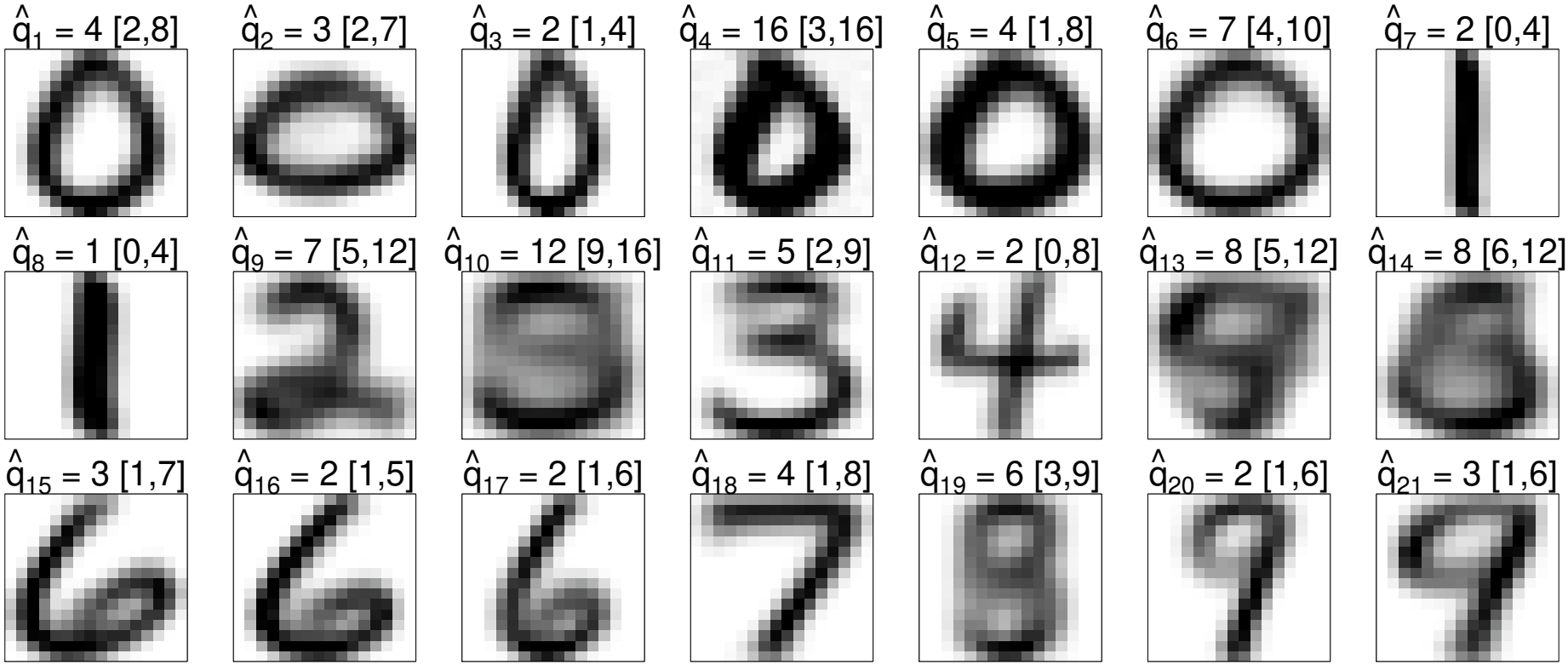
USPS Digits[22]

- $N = 7,291$ images of the digits 0,...,9, taken from handwritten zip codes from the United States Postal Service (USPS)
- Each digit is represented by a 16×16 grayscale grid concatenated into a $P = 256$-dimensional vector.
- Fitting many **MFA** or **MIFA** models is practically infeasible for this data
- **IMIFA** uncovers $\widehat{G} = 21$ clusters and assigns images of the same digit, albeit written differently, to different clusters with different $\widehat{q}_g$ values
- Allowing cluster-specific numbers of factors helps characterise digits with different geometric features and complexities

[22] Hastie et al. (2001)

USPS Digits

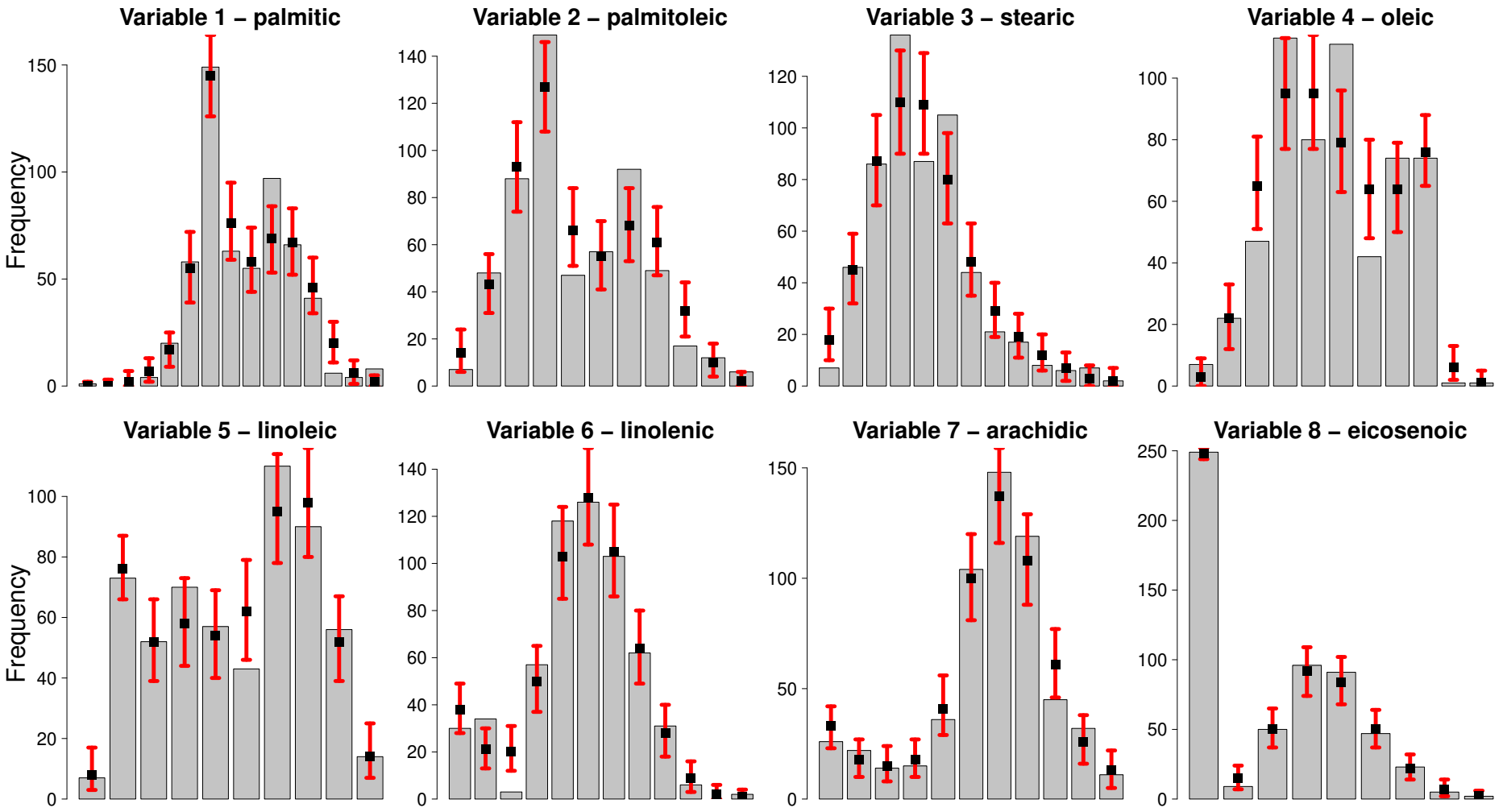
- Cluster posterior means, with cluster-specific numbers of factors and uncertainty:



Model Checking

- Aspect of (Bayesian) clustering methods that is often ignored
- Posterior predictive checking:
  how to do it in multimodal and multidimensional settings?
- Small p: examine histograms comparing bin counts of modelled versus replicate data for each variable

Model Checking



- Large p: proposed the '**P**osterior **P**redictive **R**econstruction **E**rror' (**PPRE**):

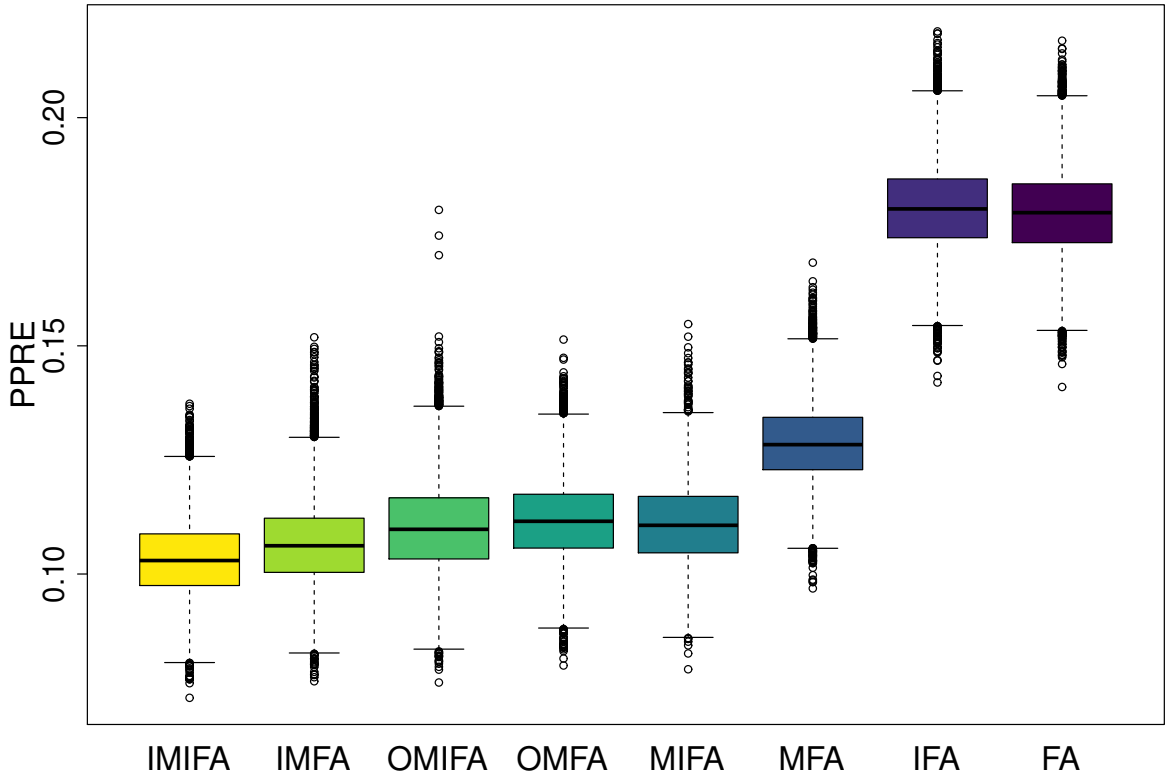
1. Transformed modelled data $\mathbf{X}$ into a $h \times p$ matrix $\mathcal{H}$
   Each column j contains bin counts of histogram for variable j
   No. of bins $h =$ max over all p variables, padded with 0s as required
2. Generate R replicate data sets $\mathbf{X}^{(r)}$ from the posterior predictive distribution, $r = 1, \ldots, R$
3. Create a similar histogram based matrix of $\mathcal{H}^{(r)}$ for each $\mathbf{X}^{(r)}$
4. Compute the Frobenius norm:

$$\left\| \boldsymbol{\mathcal{H}} - \boldsymbol{\mathcal{H}}^{(r)} \right\|_{\mathcal{F}}$$

5. Standardise to $[0, 1]$ to obtain the **PPRE** via:

$$\left| \left\| \boldsymbol{\mathcal{H}} \right\|_{\mathcal{F}} - \left\| \boldsymbol{\mathcal{H}}^{(r)} \right\|_{\mathcal{F}} \right| \leq \left\| \boldsymbol{\mathcal{H}} - \boldsymbol{\mathcal{H}}^{(r)} \right\|_{\mathcal{F}} \leq \left\| \boldsymbol{\mathcal{H}} \right\|_{\mathcal{F}} + \left\| \boldsymbol{\mathcal{H}}^{(r)} \right\|_{\mathcal{F}}$$

- **Olive oil** data:



- **Metabolomic** data: median $\mathbf{PPRE} = 0.21\ [0.18, 0.24]$

- **USPS** data: median $\mathbf{PPRE} = 0.05\ [0.04, 0.06]$

Discussion

- Proposed a family hierarchy of **IMIFA** models, from **FA** to **IMIFA**
- Potentially broader **IMIFA** family than that considered here:
  - alternative shrinkage priors, e.g. IBP, BP, spike-and-slab[23]
  - alternative mixture settings, e.g. DP-BP[24]
- PYP-MGP: continuous shrinkage ethos
  DP-BP: exact shrinkage ethos (worse performance on digit data)
- PYP-MGP is the flagship of the (current) family:
  - flexible
  - computationally efficient
  - enables uncertainty quantification
  - removes reliance on model selection criteria

[23] Legramanti (2019)
[24] Chen et al. (2010)

Discussion

- The model family includes both sparse finite (i.e. overfitted) mixtures and infinite mixtures
- Following Früwirth-Schnatter & Malsiner-Walli (2019), the priors governing their mixing proportions are 'matched', leading to 'sparsified' **PYP** mixtures
- This helps address concerns around overestimation of G under the **PYP** prior[25]
- De Blasi et al. (2015) refer to an alternative formulation of the **PYP** with $d < 0$ and $\alpha = m|d|$, for integer m
- This setting will be explored in future work, by virtue of its equivalence with a prior on m to a sparse finite mixture with a prior on G, to further unify the two model classes

[25] Miller & Harrison 2014

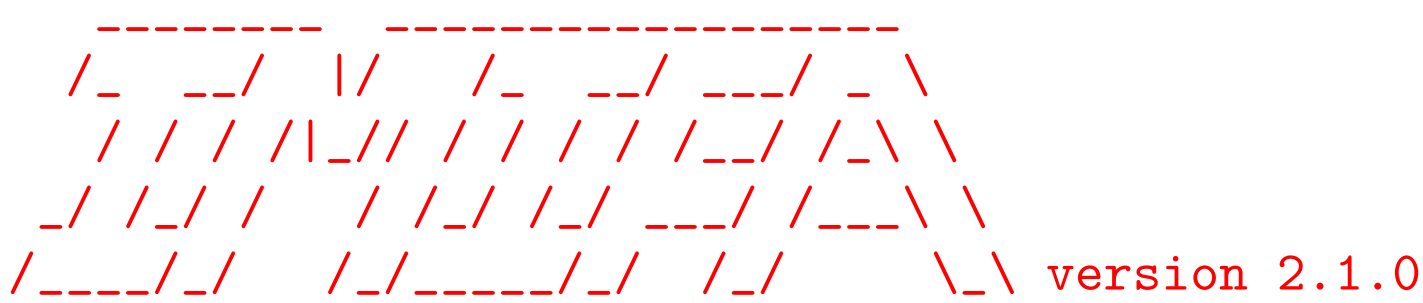
Discussion

- Wealth of potential model extensions:
  - **constrained** loadings as per `pgmm`
    (constrained uniquenesses already explored in the paper)
  - inclusion of **covariates**: mixture of experts approach
  - **semi-supervised** settings with infinite factors
  - new shrinkage priors – **computational efficiency**?
  - **robustifying IMIFA** family models with multivariate skew/t-distributions
  - power-posterior **tempering**[26]
  - **variable selection**
  - **heteroscedastic factors**: improved inference / addressing rotational invariance?

[26] Miller & Dunson (2018)

References

- `IMIFA` R package on CRAN: cran.r-project.org/web/packages/IMIFA



- Article in *Bayesian Analysis*:

Murphy, K., C. Viroli, and I. C. Gormley (2019). Infinite mixtures of infinite factor analysers. *Bayesian Analysis*, advance publication, 1–27.
URL: https://projecteuclid.org/euclid.ba/1570586978

Other work

- Murphy, K. and T. B. Murphy (2019). **Gaussian parsimonious clustering models with covariates and a noise component**. *Advances in Data Analysis and Classification*, advance publication, 1–33.
  URL: https://doi.org/10.1007/s11634-019-00373-8

- Murphy, K., T. B. Murphy, R. Piccarreta, and I. C. Gormley (2019). **Clustering longitudinal life-course sequences using mixtures of exponential-distance models**.
  arXiv pre-print: 1908.07963

- Murphy, K., C. Viroli, and I. C. Gormley (2019). **Infinite mixtures of infinite factor analysers**. *Bayesian Analysis*, advance publication, 1–27.
  URL: https://projecteuclid.org/euclid.ba/1570586978

- Respective R packages: `MoEClust`, `MEDseq`, and of course `IMIFA`