Finding structures in observations: consistent(?) clustering analysis.

Clara Grazian

School of Mathematics and Statistics, University of Sydney

clara.grazian@sydney.edu.au

OBayes Conference

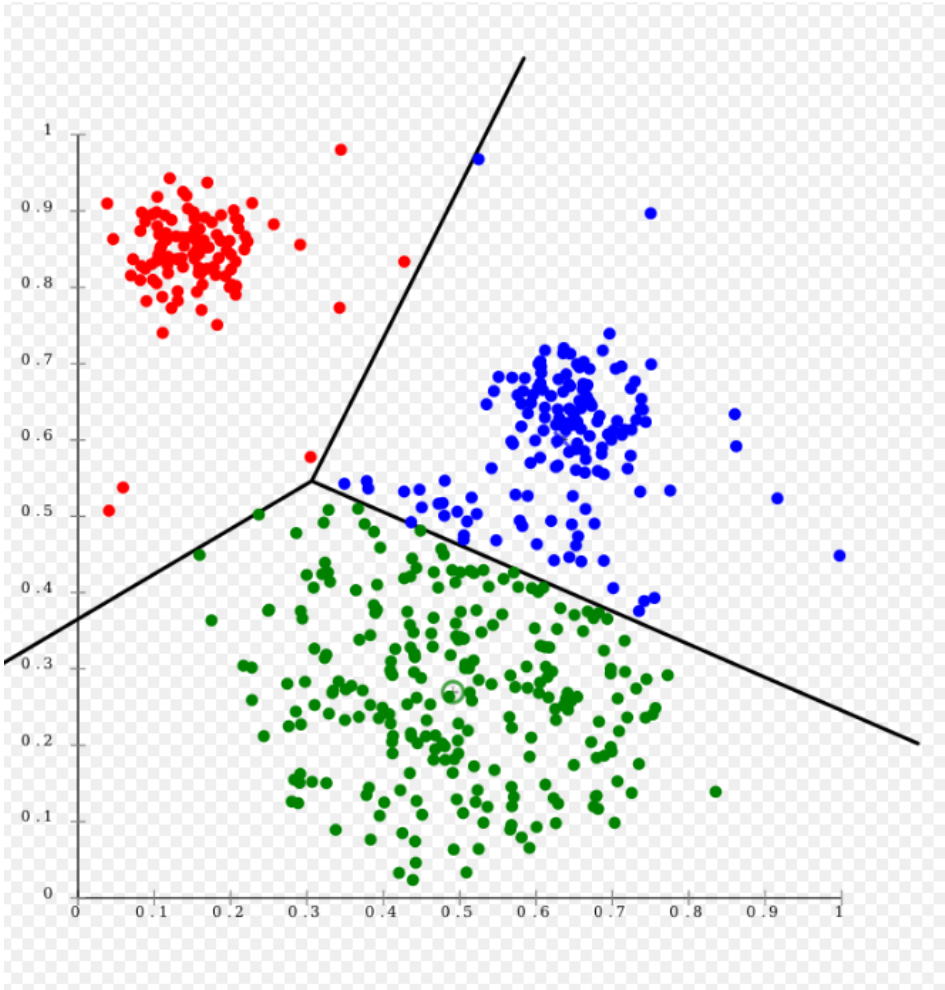
(Santa Cruz)

10 September 2022

Clustering

Clustering:

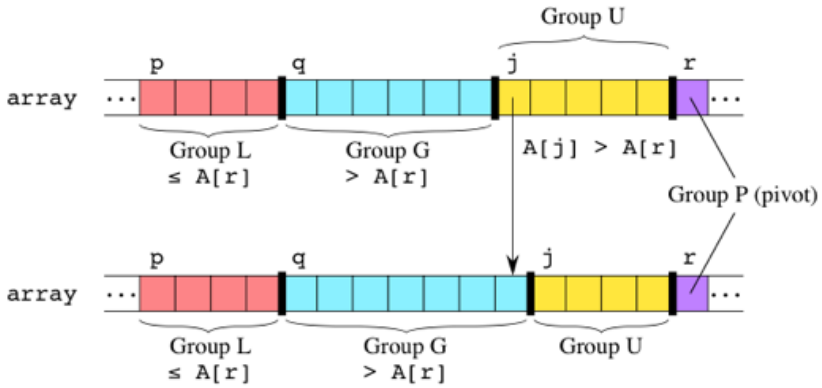
- "unsupervised learning"
- requires data, but no labels
- **detect patterns**, e.g.
  - online search results
  - customer shoppint patterns
  - effect of pollution
  - animal behaviours
  - cells, tissues, etc
  - regions of images
- common initial analysis: useful when you have no idea
- how to interpret results?



Clustering algorithms

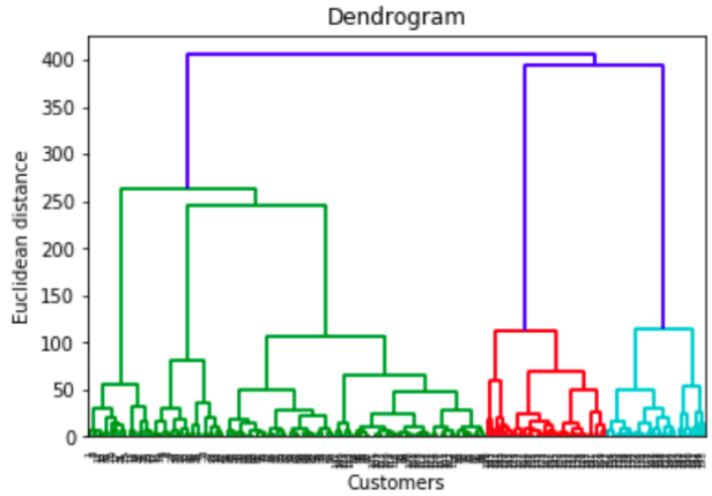
Partitioning algorithms:

- k-means
- mixture models
- spectral custering



Hierarchical algorithms:

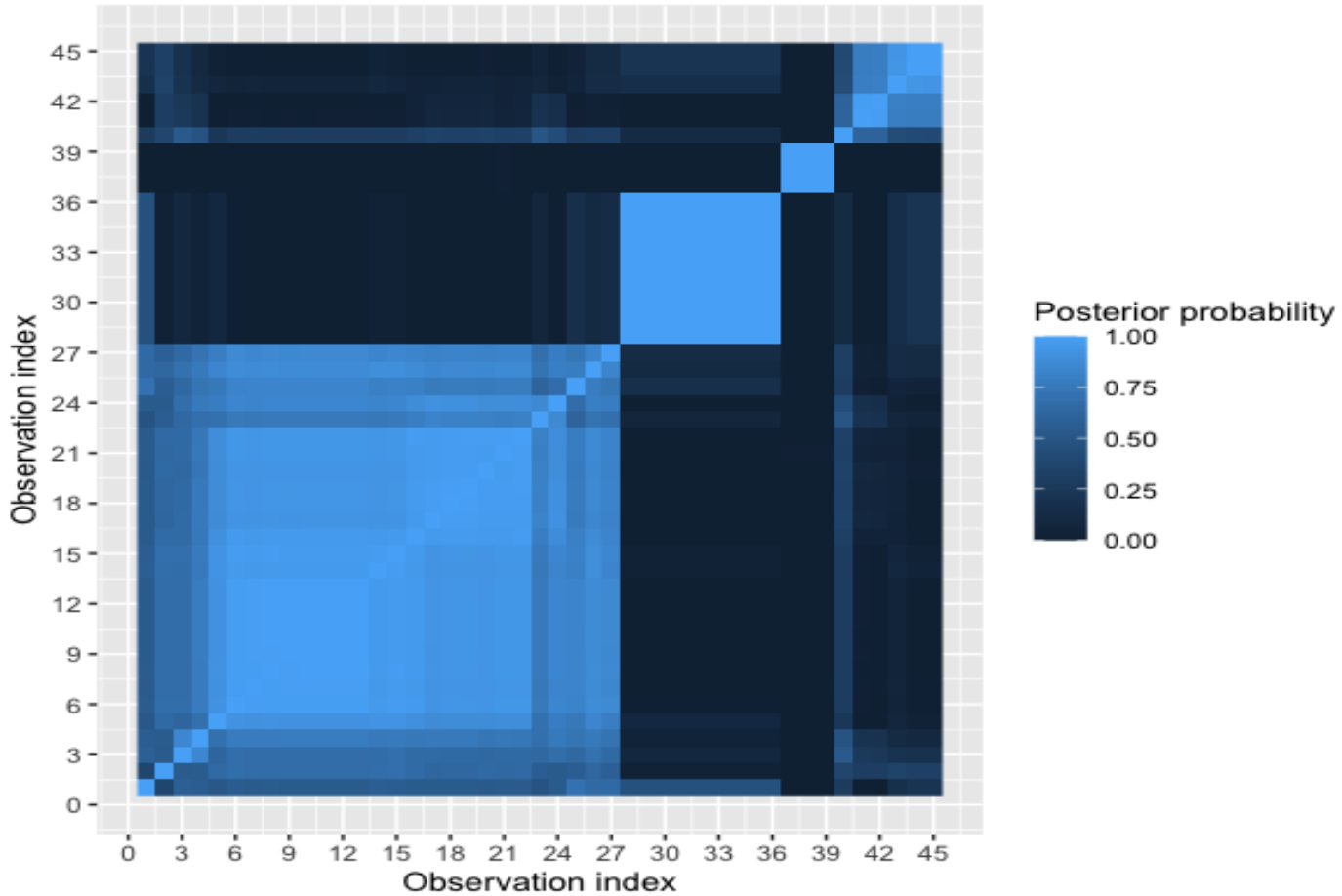
- bottom-up, agglomerative
- top-down, divisive



Examples of clustering

Clustering gene expression data:

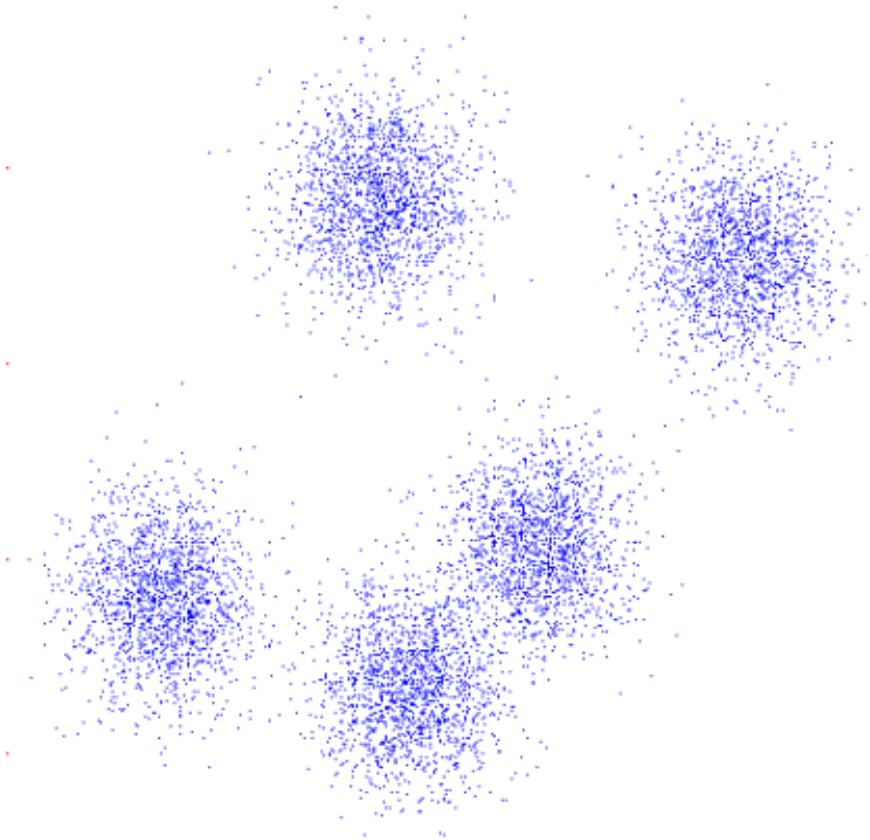
Find clusters of cells with similar biological expression



K-means

An iterative algorithm:

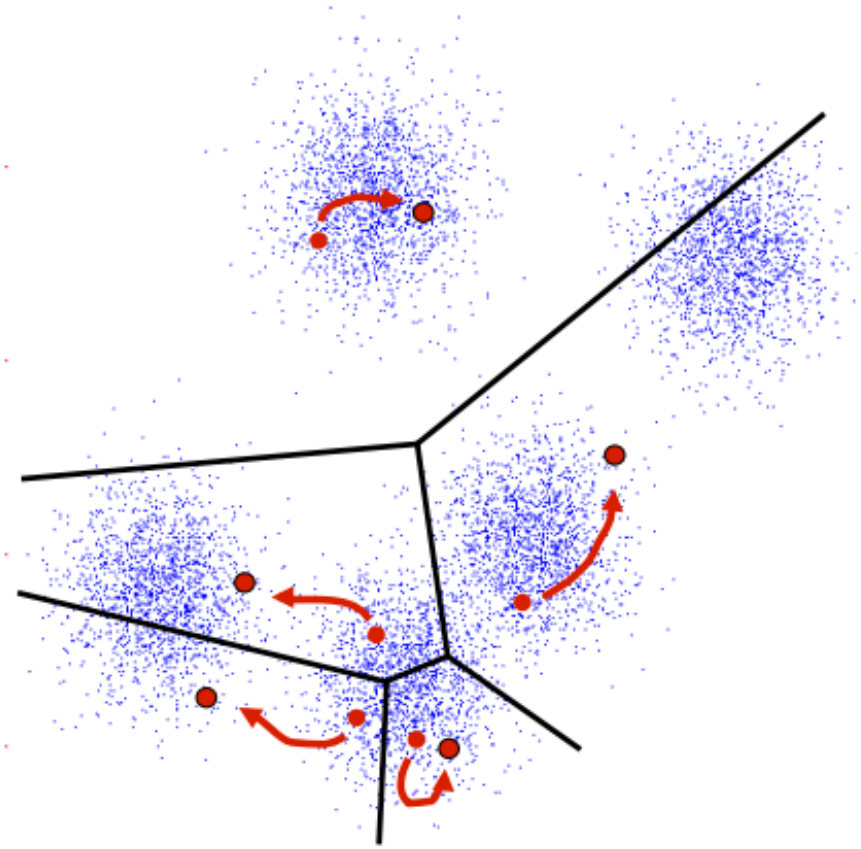
- **Initialise:** pich K random points as cluster centers
- **Alternate:**
  - assign data points to closest cluster center
  - change the cluster center to the average of its assigned points
- **Stop:** when there is no change in assignments



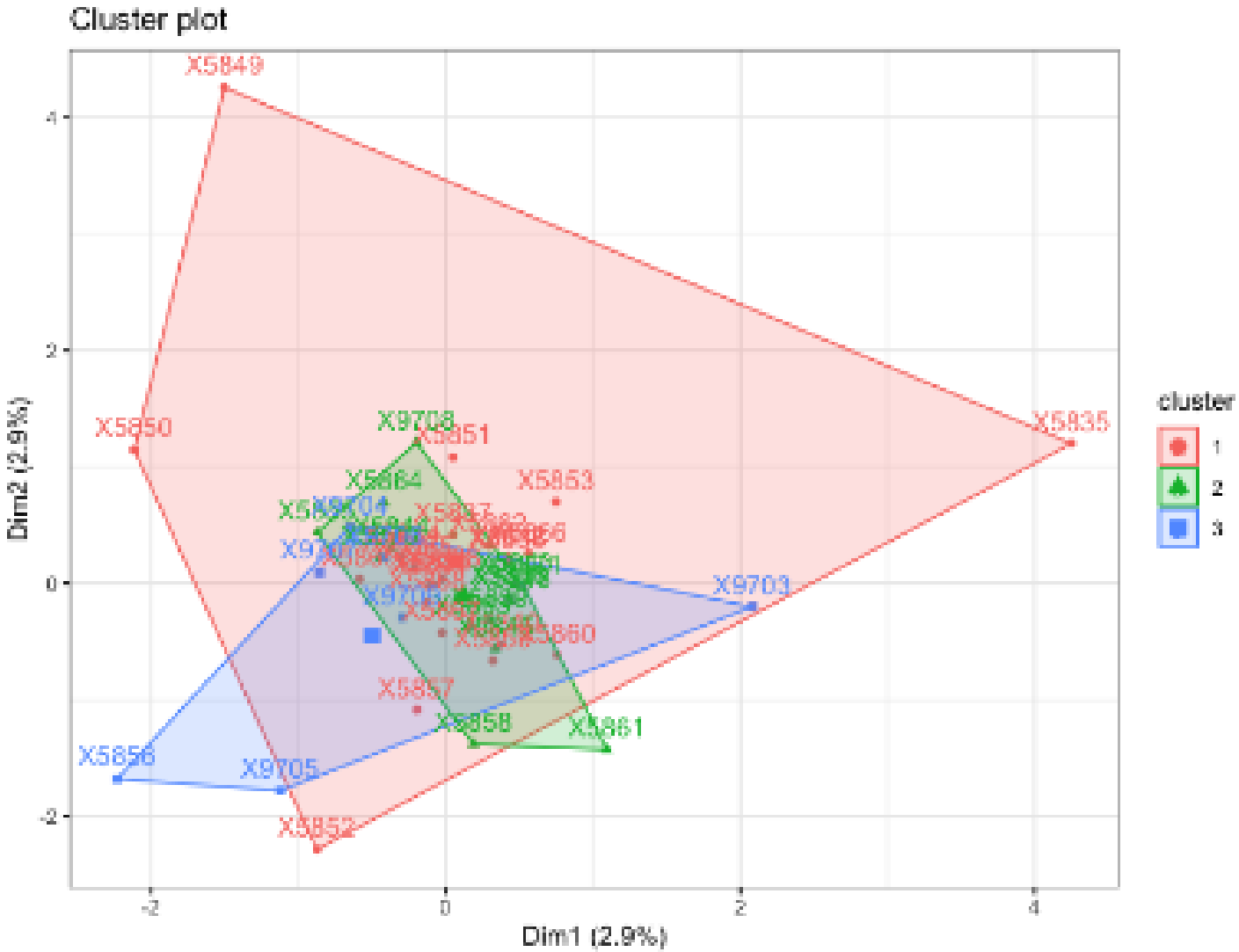
K-means

An iterative algorithm:

- **Initialise:** pich K random points as cluster centers
- **Alternate:**
  - assign data points to closest cluster center
  - change the cluster center to the average of its assigned points
- **Stop:** when there is no change in assignments



K-means


Cluster plot
X5849
X5850
X9708
X5851
X5864
X5853
X5835
X9704
X9703
X5860
X5857
X5858
X5861
X5856
X9705
X5852
cluster
1
2
3
Dim2 (2.9%)
Dim1 (2.9%)
4
2
0
-2
-2
0
2
4

Mixture Models

Consider the following mixture model

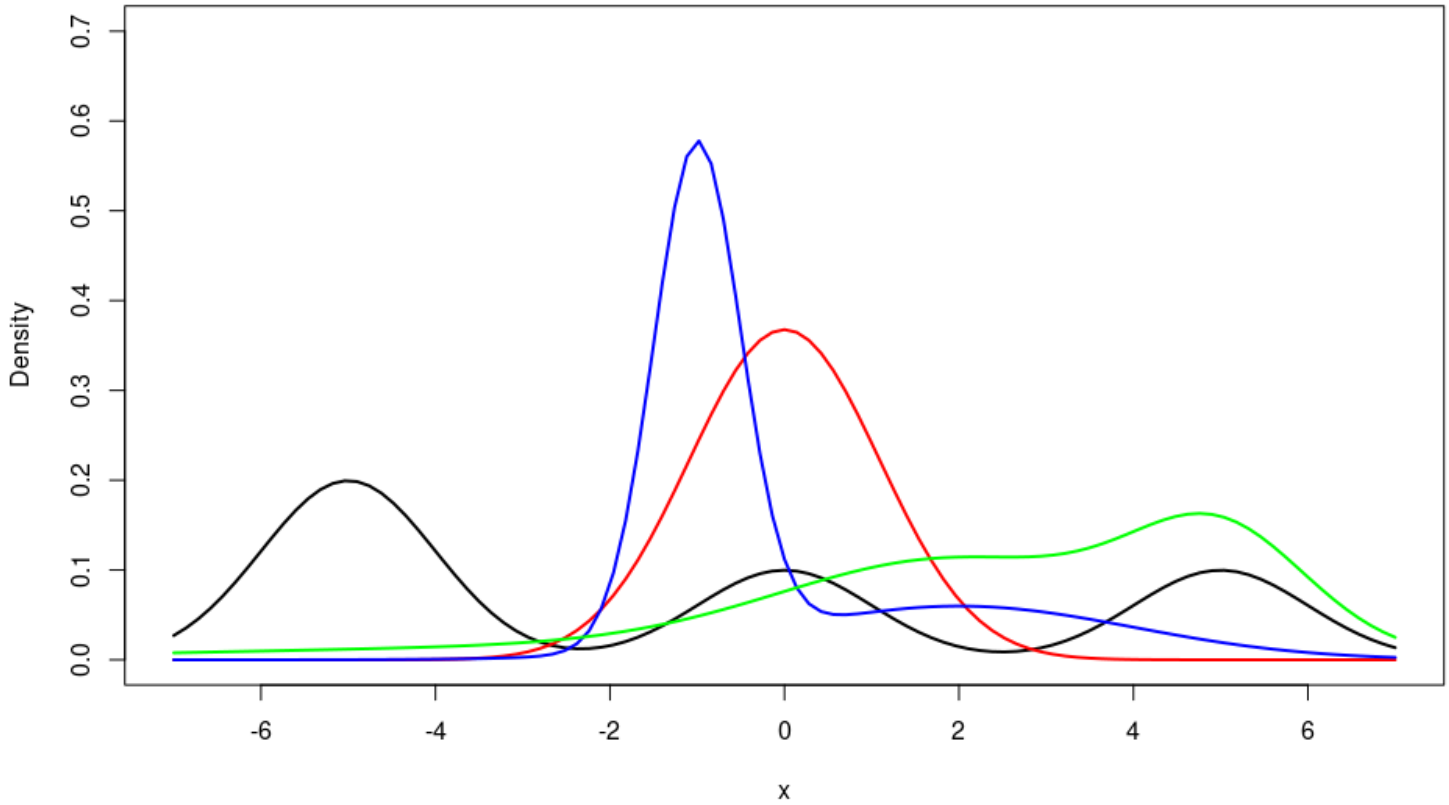
$$g(y;\psi) = \sum_{j=1}^{K} p_j f_j(y;\theta_j)$$

where

- $\psi = (\theta_1, \cdots, \theta_K, p_1, \cdots, p_K)$
- $p_j \geq 0$ for $j = 1, \cdots, K$
- $\sum_j p_j = 1$.
- $f_j(\cdot)$ is any probability distribution

These models provide a flexible tool for statistical inference (even in a nonparametric setting, see Lindsay (1995), Roeder (1992) and Roeder and Wasserman (1997)).

Mixture models



We need a prior distribution!

Often in Bayesian inference, we want to reduce the effect of the prior on the posterior distribution, in case we do not have strong prior information.

Sometimes so-called "improper" priors are used

$$\int_{\Theta} \pi(\theta) d\theta = \infty.$$

This is not a pdf (or a pmf), therefore the Bayes' theorem cannot be applied.

However they are used in practice as limit of proper prior distributions, when the assure a proper posterior distribution.

Default priors for mixture models

It is delicate to produce a noninformative prior for the parameters of a mixture model, since they are often *improper*.

Why can't we use improper priors?

Example:

Consider independent improper priors

$$\pi(\theta_1, \cdots, \theta_K) \propto \prod_{j=1}^{K} \pi(\theta_j)$$

such that $\int_\Theta \pi(\theta_j) d\theta_j = \infty$

Classical noninformative solutions - contd

The mixture model is a classical example of latent variable model, then it can be rewritten as

$$g(y;\psi) = \sum_{S=\mathscr{S}_k} \prod_{j=1}^{K} f(y;S,\theta_j)\pi(\theta_j)\pi(S \mid p)\pi(p)$$

where the summation runs over all k^N possible classifications S.
Then the complete-data likelihood is non-informative if there is an empty component (let's say the j-th)

$$\int \prod_{i:S_i=j} f(y_i;\theta_j)\pi(\theta_j)d\theta_j \propto \int \pi(\theta_j)d\theta_j = \infty$$

Random partition models

- A random partition model is a probability distribution over $\mathscr{P}_n$

$$\{p(\rho_n = (S_1, \ldots, S_K)) : \rho_n \in \mathscr{P}_n\}$$

- One main approach is to to define $p(\rho_n)$ through discrete random probability measures

$$f(y_1, \ldots, y_n | \theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} f(y_i | \theta_i)$$

$$\theta_1, \ldots, \theta_n | G \overset{i.i.d.}{\sim} G$$

$$G \sim \text{discrete RPM}$$

For example, if $G(\cdot) = \sum_{h=1}^{K} p_h \delta_{\psi_h}$ with $P(\sum_{h=1}^{K} p_h = 1) = 1$, then

$$g(y_i | \{\psi_h\}) = \sum_{h=1}^{K} p_h f(y_i | \psi_h)$$

Random partition models

- Discreteness of G implies existence of ties among $\theta_1, \ldots, \theta_n$
- If $\psi_1, \ldots, \psi_K$ denote the corresponding unique values then we can define ρ_n through indicators given as

$$c_i = j \Leftrightarrow \theta_i = \psi_j \text{ or equivalently } \theta_i = \psi_{c_i}$$

  and so $S_j = \{i \in [n] : \theta_i = \psi_j\}$, where $[n] = \{1, \ldots, n\}$ is the set of n indices.
- This is an induced random partition model.

Is it easy to define a prior over the partition space?

- $\rho_n = (S_1, \ldots, S_K)$ a partition of $[n]$ into $K = |\rho_n| \geq 1$ nonempty (and mutually exclusive) subsets;
- $\mathscr{P}_n$: set of all partitions of $[n]$;
- the size of $\mathscr{P}_n$ increases as the Bell number; e.g. $B_{10} = 115,975$

Finite mixture models

- Consider $K \sim p_K(k)$

$$y_i|K=k,p_1,\ldots,p_K,\theta_1,\ldots,\theta_K \overset{i.i.d.}{\sim} \sum_{h=1}^{K} p_h f(y_i|\theta_h)$$

$$\theta_1,\ldots,\theta_k|K=k \overset{i.i.d.}{\sim} p_0(\theta)$$

$$(p_1,\ldots,p_k)|K=k \sim Dir(\gamma,\ldots,\gamma)$$

- The induced partition model is then

$$p(\rho_n=(S_1,\ldots,S_K)) = \left(\sum_{h=1}^{\infty} \frac{h_{(k)}}{(\gamma h)^{(n)}} p_K(h)\right)\left(\prod_{s\in(S_1,\ldots,S_K)} \gamma^{|s|}\right)$$

where $x^{(m)} = x(x+1)\ldots(x+m-1)$ and $x_{(m)} = x(x-1)\ldots(x-m+1)$.

CASE 1: Overfitted mixtures

- One approach consists in fixing K to a large value and use inference to estimate some of the weights as equal to zero, in order to identify the correct $k < K$ number of clusters.
- Playing on the prior for $(p_1, \ldots, p_K)$
- [Rousseau and Mengersen (2011)] show the asymptotic behaviour of the posterior distribution in a mixture model for overfitted mixtures: the posterior distribution concentrates on a sparse representation of the true density; this is exhibited by a subset of componenets that adequately describe the density remaining and any superfluous components becoming empty.
- IMPORTANT: need for a prior on the weigths that favour small weights (Dirichlet with parameters 1/2).

[Grazian & Robert, 2018] Jeffreys' prior for mixture models

We recall that the Jeffreys prior was introduced by Jeffreys (1939) as a default prior based on the Fisher information matrix

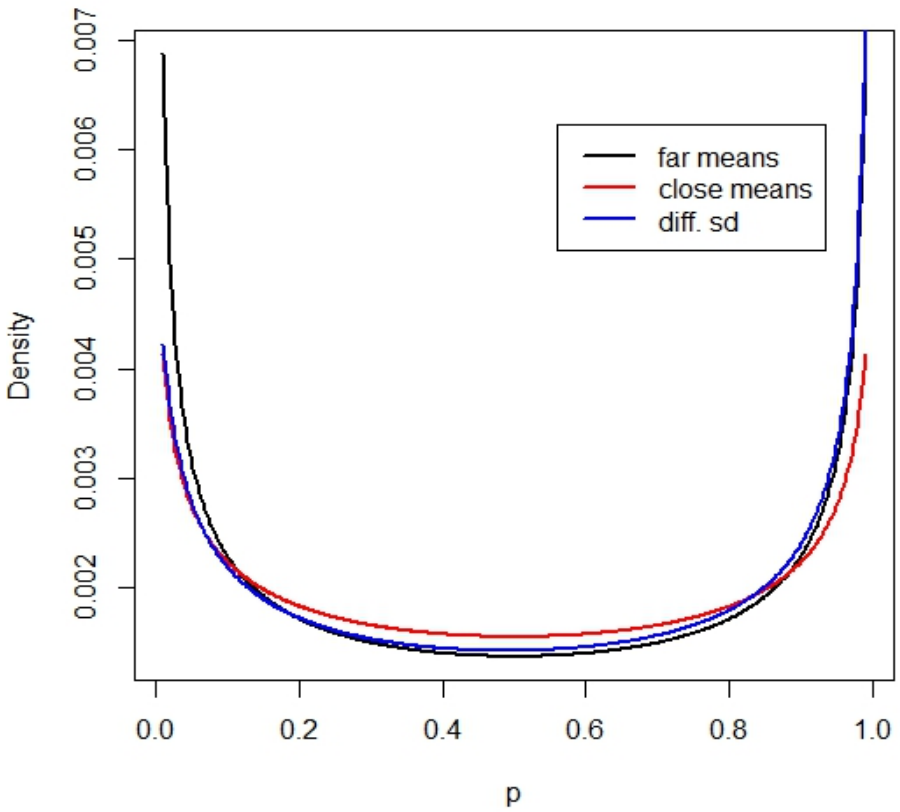
$$\pi^{\mathrm{J}}(\theta, \mathbf{p}) \propto |I(\theta, \mathbf{p})|^{1/2} = \det\left(\mathbb{E}_g\left[-\frac{d^2}{d\psi^2}\log g(y; \mathbf{p}, \theta)\right]\right)^{1/2}$$

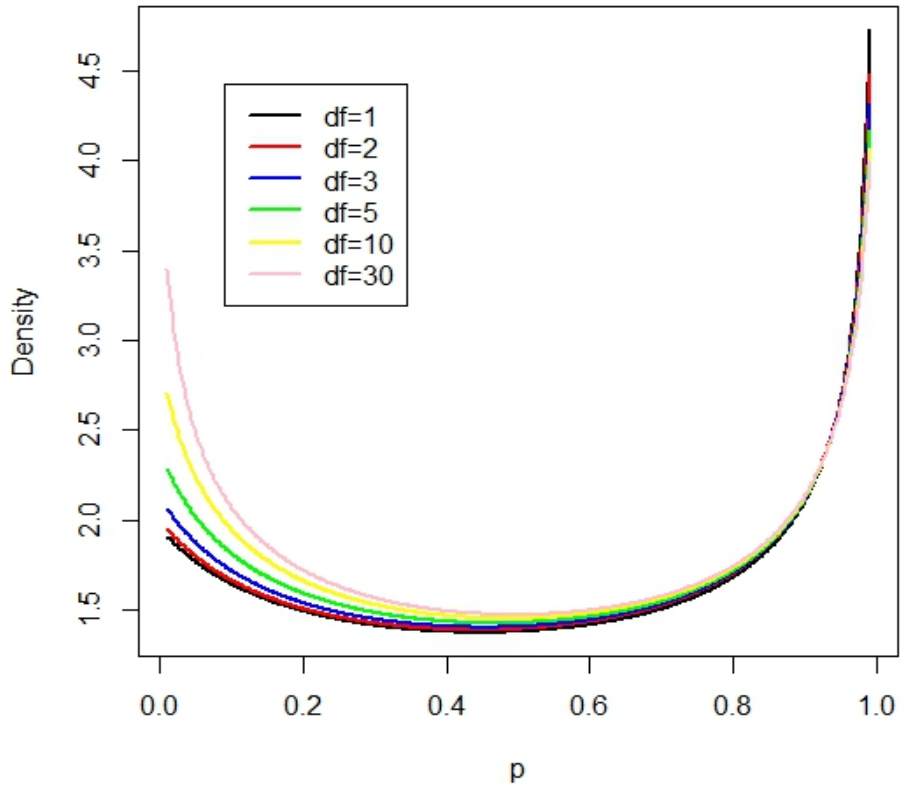
- when using the Jeffreys' prior for all the parameters of the model, the posterior is improper (OH NO!)
- but when the Jeffreys' prior is used only for the weights, it can be shown that it leads to the same results as Rousseau & Mengersen (2011)!

Jeffreys prior for the weights

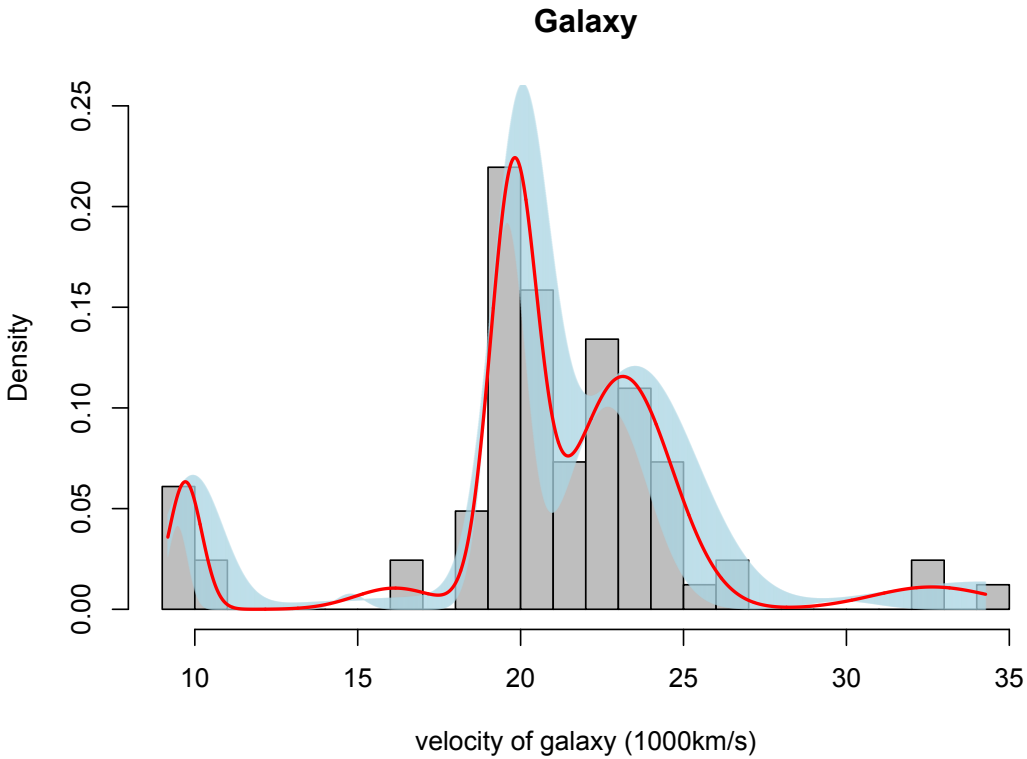
Instead, we fix the Jeffreys prior only for the weights conditionally on the other parameters

$$\pi^J(p_1, \ldots, p_K | \theta_1, \ldots, \theta_K) \propto |I(p_1, \ldots, p_K)|^{1/2}$$





The galaxy dataset



Dataset:	**galaxy**
p_1	0.437
	(23.139, 1.507)
p_2	0.390
	(19.790, 0.715)
p_3	0.080
	(9.709, 0.503)
p_4	0.056
	(32.630, 1.842)
p_5	0.037
	(16.138,1.226)
$\sum_{\ell=6}^{10} p_\ell$	0.000

CASE 2: A prior on the number of the # of components

The induced partition model is then

$$p(\rho_n = (S_1, \ldots, S_K)) = \left(\sum_{h=1}^{\infty} \frac{h_{(k)}}{(\gamma h)^{(n)}} p_K(h) \right) \left(\prod_{s \in (S_1, \ldots, S_K)} \gamma^{|s|} \right)$$

where $x^{(m)} = x(x+1)\ldots(x+m-1)$ and $x_{(m)} = x(x-1)\ldots(x-m+1)$.

A prior on the number of components

For model

$$g(y;\psi) = \sum_{j=1}^{K} p_j f_j(y;\theta_j)$$

the prior can be specified as

$$\pi(k, \mathbf{p}, \theta) = p_K(k)\pi(\mathbf{p} \mid k)\pi(\theta \mid k).$$

The posterior for k is then given by

$$p_K(k \mid y) \propto \int f(y \mid k, \mathbf{p}, \theta) \times p_K(k)\pi(\mathbf{p} \mid k)\pi(\theta \mid k)\, d\mathbf{p}\, d\theta.$$

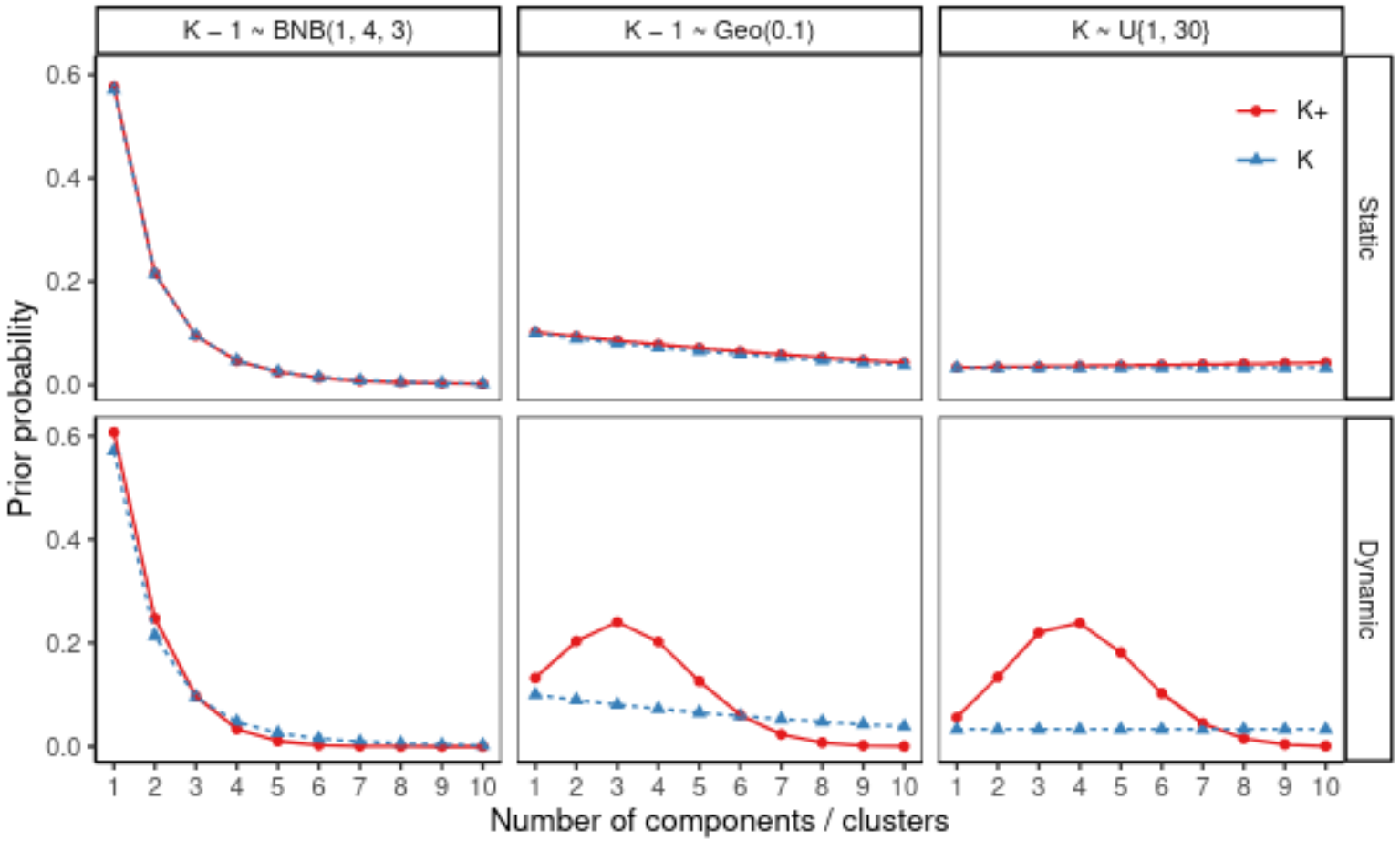
Some difficulties

- Although for practical purposes the range of values K can take is finite, it may be appropriate to define a prior over $\mathbb{N}$.
- In fact, by truncating the support of K there may be possible distortions of the posterior around the boundary, affecting the inferential results.
- **BUT** the prior on K must be proper, as proved by Nobile (2005).
- Remember the **inconsistency problems** in the nonparametric setting - see Miller and Harrison (2014).

Possible prior distributions

- $K \sim Unif(0, 30)$ (Richardson and Green, 1997)
- $K \sim Pois(1)$ (Nobile and Fearnside, 2007)
- $K \sim BNB(1, a, b)$
  (Grazian et al. 2020, Früwirth-Schnatter et al., 2021)
  - Früwirth-Schnatter et al. (2021) propose to combine a prior on K with an adaptive prior on the weights $(p_1, \ldots, p_K) \sim Dir(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$ — "dynamic" version of the mixtures

Possible prior distributions



Früwirth-Schnatter et al., 2021

A loss-based prior [Grazian et al., 2020]

Idea: inconsistency problems can be prevented by penalising larger values $\rightarrow$ we can define the prior on K with a loss-based approach.

To obtain the **loss-based prior** on K, we define the prior on K by assigning a prior on the space of models determined by the mixtures with $k = 1, 2, \ldots$ components.

- we can assign a *worth* to each mixture
- we include a component of loss due to the complexity of the model

$$\text{Loss}(k) = \text{Loss}_I(k) + \text{Loss}_C(k)$$

A loss-based prior: information loss [Grazian et al., 2020]

The quantification of the loss comes from Berk (1966): *if the model is misspecified, the posterior distribution asymptotically tends to accumulate at the most similar model so to minimise the loss in information, in terms of Kullback-Leibler divergence*
If we consider a mixture $M_s = \{g_s(x|\psi_s), \pi_s(\psi_s)\}$ (where $\psi_s = (p_s, \theta_s)$

$$\text{Loss}_I(k) = \mathbb{E}_{\pi_s}\left\{\inf_{\psi_m, m\neq j} D_{KL}\Big(g_s(x|\psi_s)\|g_m(x|\psi_m)\Big)\right\},$$

The above loss is linked to the prior mass by means of the *self-information* loss function which associate a loss to a probability statement. As such,

$$p_K(k) \propto \exp\{\text{Loss}_I(k)\}.$$

The loss attains its minimum at zero: Consider mixture $g_k = \sum_{j=1}^{k} p_j f_j(x|\theta_j)$ and $g_{k+1} = \sum_{j=1}^{k} \breve{p}_j f_j(x|\breve{\theta}_j) + \breve{p}_{k+1} f_{k+1}(x|\breve{\theta}_{k+1})$; the minimum is obtained when $\breve{p}_j = p_j$ and $\breve{\theta}_j = \theta_j$ and $\breve{p}_{k+1} = 0$.

A loss-based prior: complexity loss [Grazian et al., 2020]

To fully describe the *worth* of a mixture model it is also necessary to take into consideration its complexity.

If we keep the mixture model with k components, the loss would be related to the number of parameters that have to be estimated, and therefore the number of components.

$$\text{Loss}_C(k) = \text{U}(\text{keep } k) = -c \cdot k.$$

Therefore,

$$p_K(k) \propto \exp\{-c \cdot k\},$$

where $c > 0$ is included as loss functions are defined up to a constant.

A reparametrization

Theorem

Consider the prior distribution for the number of components of a finite mixture model, where we set $p = \exp\{-c\}$ and $k = 1, 2, \ldots$. If we choose $p \sim \text{Beta}(\alpha, \beta)$, with $\alpha, \beta > 0$, then

$$p_K(k|p) = p^{k-1}(1-p),$$

which is a geometric distribution with parameter $1-p$, and

$$p_K(k) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(k+\beta-1)\Gamma(\alpha+1)}{\Gamma(k+\alpha+\beta)},$$

which is a beta-negative-binomial distribution where the number of failures before the experiment is stopped is equal to 1, and shape parameters α and β.

The linear complexity loss is a choices; other choices are also possible.

A loss-based prior

The prior $p_K(k)$ just defined

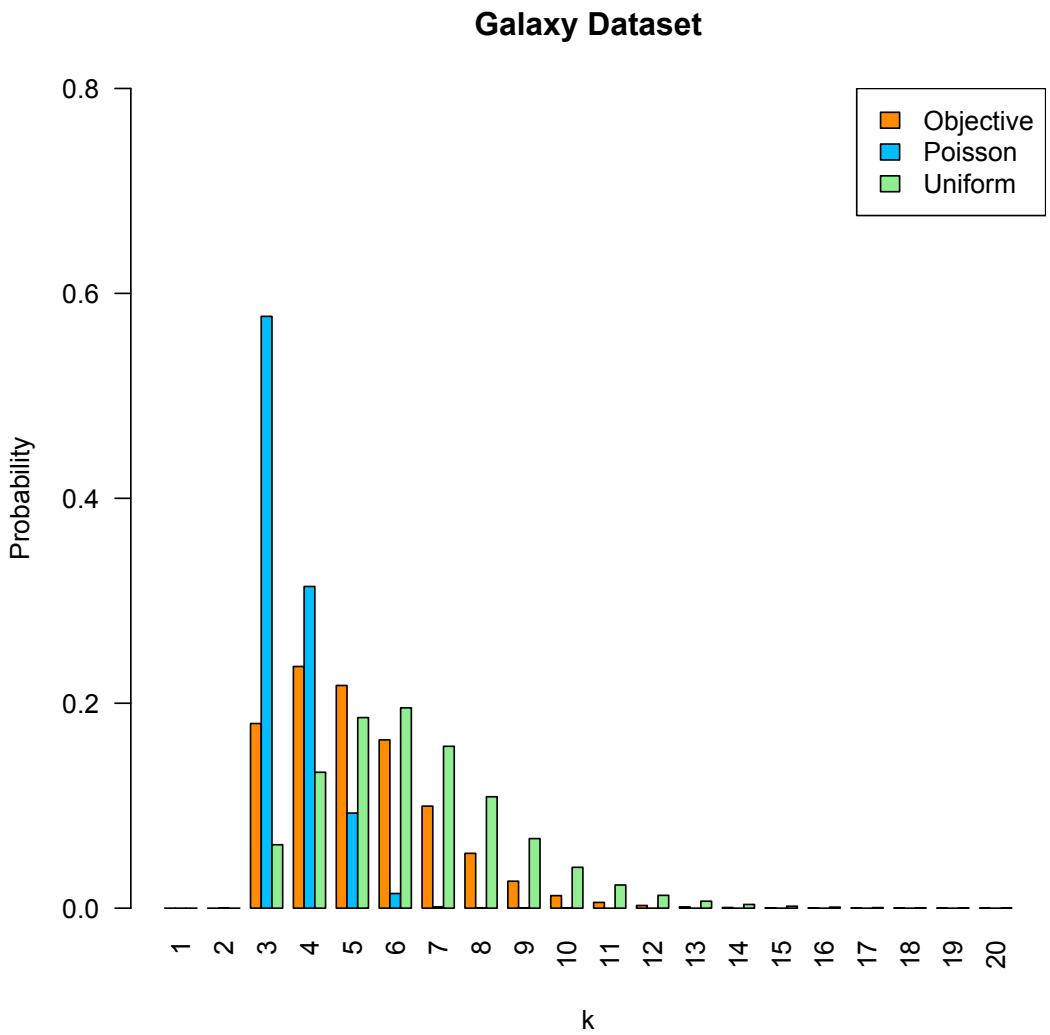
- is defined on the whole support of K, $\mathbb{N}$
- is proper
- has moments

$$\mathbb{E}(K) = \mathbb{E}(\mathbb{E}\{K|p\}) = \mathbb{E}(p^{-1}) = \frac{\alpha+\beta-1}{\alpha-1}, \qquad \text{for } \alpha > 1,$$

$$\text{Var}(K) = \mathbb{E}(\text{Var}\{K|p\}) + \text{Var}(\mathbb{E}\{K|p\}) = \frac{\alpha\beta(\alpha+\beta-1)}{(\alpha-2)(\alpha-1)^2}, \quad \text{for } \alpha > 2.$$

where β can be used to control how many components we assume a priori and α can be used to control the variance.

The galaxy dataset



CASE 3: Introducing covariates

Suppose that the observations depend on a covariate, e.g. they are time-dependent. $\rightarrow$ we can use hidden Markov models!
Let

- $\{t_1, t_2, \ldots, t_T\} \equiv \mathscr{T}$: set of observed time points
- $\mathbf{y} = \{\mathbf{y}_t\}_{t \in \mathscr{T}}$: the data
- $\mathbf{c} = \{c_t\}_{t \in \mathscr{T}}$: a latent variable indicating the cluster each observation belongs to, with $c_t \in \{1, 2, \ldots, K\} \equiv \mathscr{K}$ and K

We assume that the data come from a mixture-type model

$$g(\mathbf{y}|\mathbf{c}, \{\theta_k\}_{k \in \mathscr{K}}) = \prod_{t \in \mathscr{T}} \prod_{k \in \mathscr{K}} f(y_t|\theta_k)^{\mathbb{I}_k(c_t)}$$

i.e. given the latent variables $\mathbf{c}$, the observations y_t are independent.

Logit-normal process

One possible solution

$$g(\mathbf{y}|\mathbf{c}, \{\theta_k\}_{k\in\mathscr{K}}) = \prod_{t\in\mathscr{T}}\prod_{k\in\mathscr{K}} f(y_t|\theta_k)^{\mathbb{I}_k(c_t)},$$
$$c_t \sim \text{Discrete}(\mathbf{p}_t),$$
$$\mathbf{p}(t) \sim \text{LogitGP}(\mathbf{A}, \mu(t), \boldsymbol{C}(h))$$

So that the probabilities $\mathbf{p}_t = \{\mathbf{p}_{t,k}\}_{k\in\mathscr{K}}$ are discrete-time observations of an underlying and non-observed continuous-time process $\mathbf{p}(t)$.

We have that

- $\mathbf{A}$ is a co-regionalization matrix
- $\mu(t)$ is a mean function
- $\mathbf{C}(h)$ is a correlation function with h being a temporal distance

LogitN distribution and LogitGP

[Aitchison, 1986] proposed the LogitN distribution to model compositional data as an alternative to the Dirichlet distribution.

The vector $\mathbf{p}_t$ is defined as

$$p_{t,k} = \frac{e^{\omega_{t,k}}}{\sum_{j=1}^{K} e^{\omega_{t,j}}}, \quad k \in 1, \ldots, K$$

where $\omega_{t,k}$ are real valued variables.

Remark: adding a constant to each $\omega_{t,k}$ produces the same vector of probabilities, and an identifiability constraint is therefore needed; the $K-$th element is set to zero ($\omega_{t,K} = 0$) treated as the *reference element*.

ω_t can be the realisation of a $K-1$ dimensional GP $\omega(t)$.

The covariance of the $\mathbf{p}_t$

Attention must be paid! The covariance among each element and the sum of all the element is

$$\text{Cov}(p_{t,k}, p_{t,1} + \cdots + p_{t,k} + \cdots + p_{t,K}) = 0$$

where $p_{t,1} + \cdots + p_{t,k} + \cdots + p_{t,K} = 1$. Therefore we have

$$-\text{Var}(p_{t,k}) = \sum_{\substack{h=1 \\ k \neq h}}^{K} \text{Cov}(p_{t,k}, p_{t,h}).$$

Aitchison (1986) pointed out that a more consistent measure of dependence between compositional elements can be measure as

$$\tau_{ij,kl}(t, t') = \text{Cov}\left(\log \frac{p_{t,i}}{p_{t,k}}, \log \frac{p_{t',j}}{p_{t',l}}\right), \; i, j, k, l \in 1, \ldots, K,$$

The covariance of the $\mathbf{p}_t$

Let's keep things simple and suppose that $\mathbf{p}_t \sim LogitN(\mu_t, \mathbf{\Sigma}_t)$, where μ_t is $K-1$ dimensional vector and $\mathbf{\Sigma}_t$ is a $(K-1)\times(K-1)$ square matrix.

- It can be proved that a *LogitN* process has independent components (in term of log-ratio), i.e. $\tau_{ij,kl}(t,t') = 0$ for arbitrary i, j, k and l, at time lag $|t-t'|$ only if the variance of the Gaussian variable is

$$\mathbf{\Sigma}_{t,t'} = \begin{pmatrix} a_1(t,t') + a_K(t,t') & a_K(t,t') & \dots & a_K(t,t') \\ a_K(t,t') & a_2(t,t') + a_K(t,t') & \dots & a_K(t,t') \\ \dots & \dots & \dots & \dots \\ a_K(t,t') & a_K(t,t') & \dots & a_{K-1}(t,t') + a_K(t,t') \end{pmatrix}$$

  where the element $[\mathbf{\Sigma}_{t,t'}]_{i,j}$ is $\tau_{ij,KK}(t,t')$.

- The elements of $\mathbf{p}$ are iid if $\mu_t = \mathbf{0}$ and

$$\mathbf{\Sigma}_{t,t'} = \begin{pmatrix} 2a & a & \dots & a \\ a & 2a & \dots & a \\ \dots & \dots & \dots & \dots \\ a & a & \dots & 2a \end{pmatrix}$$

A new parametrization [Mastrantonio et al., 2019]

We introduce an auxiliary K-dimensional GP $\gamma(t)$, From $\gamma(t)$, we construct $\omega(t)$ as

$$\begin{aligned}\omega_k(t) &= \gamma_k(t) - \gamma_K(t),\\ \gamma(t) &= \mu(t) + \mathbf{A}\gamma^*(t),\\ \gamma_k^*(t) &\sim \text{GP}(0, C_k(h)).\end{aligned}$$

where

- $\mathbf{A}$ is a coregionalization matrix, which we require to be non-negative definite and symmetric
- $\mu(t)$ is a mean function
- $\boldsymbol{C}(h)$ is a vector of correlation functions

Covariance structure

Matrix $\mathbf{A}$ introduces dependence between the elements of $\gamma(t)$, and

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{A}'$$

is the covariance of $\gamma(t)$.

The explicit relation between $\mathbf{A}$ and $\mathbf{\Sigma}$ is

$$\mathbf{A} = \mathbf{\Delta}\mathbf{\Xi}^{\frac{1}{2}}\mathbf{\Delta}',$$

where

- $\mathbf{\Delta}$ is the matrix of the eigenvectors of $\mathbf{\Sigma}$
- $\mathbf{\Xi}$ is the diagonal matrix of the eigenvalues of $\mathbf{\Sigma}$

Then

$$\mathbf{p}(t) \sim \text{LogitGP}(\mathbf{A}, \mu(t), \boldsymbol{C}(h))$$

Identifiability issues

It is important to highlight that $\gamma(t)$ is **not identifiable** and any inference about $\mathbf{p}_t$ is in fact made by looking at ω_t through equation.

$$p_{t,k} = \frac{e^{\gamma_{t,k}-\gamma_{t,K}}}{\sum_{j=1}^{K} e^{\gamma_{t,j}-\gamma_{t,K}}} = \frac{e^{\gamma_{t,k}}}{\sum_{j=1}^{K} e^{\gamma_{t,j}}}, \quad k \in 1,\ldots,K.$$

With respect to the case where $\omega_{t,K}$ must be set to zero, this equation has a more symmetric form, since all the components of $\mathbf{p}_t$ are written in terms of exponentials of $\gamma_k(t)$ and there is no reference element.

Advantages

It can be proved that this model assures

- invariance from the choice of the reference element;
- invariance with respect to the reordering of the labels;
- the expected structure of the covariance matrix among times, when defined on $\tau_{ij,kl}(t, t')$ elements.

CASE 4: infinite mixtures

A mixture model can be extended to consider infinite components:

$$
\begin{aligned}
y_i|\theta_i &\sim f(y_i|\theta_i) \qquad i = 1, \ldots, n \\
\theta_i|G &\sim G \\
G|\alpha, G_0 &\sim DP(\alpha, G_0),
\end{aligned}
$$

and, since G is almost surely discrete, this model can be rewritten as

$$
y_i \sim \sum_{h=1}^{\infty} \pi_h f(y_i|\psi_h) \qquad i = 1, \ldots, n
$$

where $\psi_1, \psi_2, \ldots$ are independent draws from the base distribution G_0.

Partitions for infinite mixtures

The EPPF of the DP is explicitly available; if $G \sim DP(\alpha, G_0)$, then

$$p(\rho_n = (S_1, \ldots, S_K)) = \frac{\alpha^K \prod_{h=1}^{K} (n_h - 1)!}{\prod_{i=1}^{n} (\alpha + i - 1)!}.$$

which is known as Ewens distribution.
(Generalization to other processes, like the PY process are available)
And the conditional EPPF for a DP mixture model, induced for a given number of clusters $K = k$, is

$$p_{DP}(\rho_n = (S_1, \ldots, S_k) | K = k) = \frac{1}{Const} \prod_{h=1}^{k} \frac{1}{n_h}$$

However, it can be shown that this EPPF favours unbalanced partitions with some small values of n_h (look at the inverse dependence on n_h) $\rightarrow$ this model is inconsistent for clustering!

CASE 4b: Adding covariates - Grazian (2022+)

Suppose $Y_t(s)$ come be represented as an infinite mixture model:

$$g(y_t(s)|\pi, \theta) = \sum_{k=1}^{\infty} \pi_{t,k}(s) g(y_t(s)|\theta_k)$$

where the mixing probability $\pi_{t,k}(s)$ is the probability that the location s belongs to component k at time t.

The mixing weights are built similarly to the spatial stick-breaking:

$$F_t(s) = \sum_{k=1}^{\infty} \pi_{t,k}(s) \delta_{\theta_k} \qquad s \in \mathscr{D}, t > 0 \qquad \text{where}$$

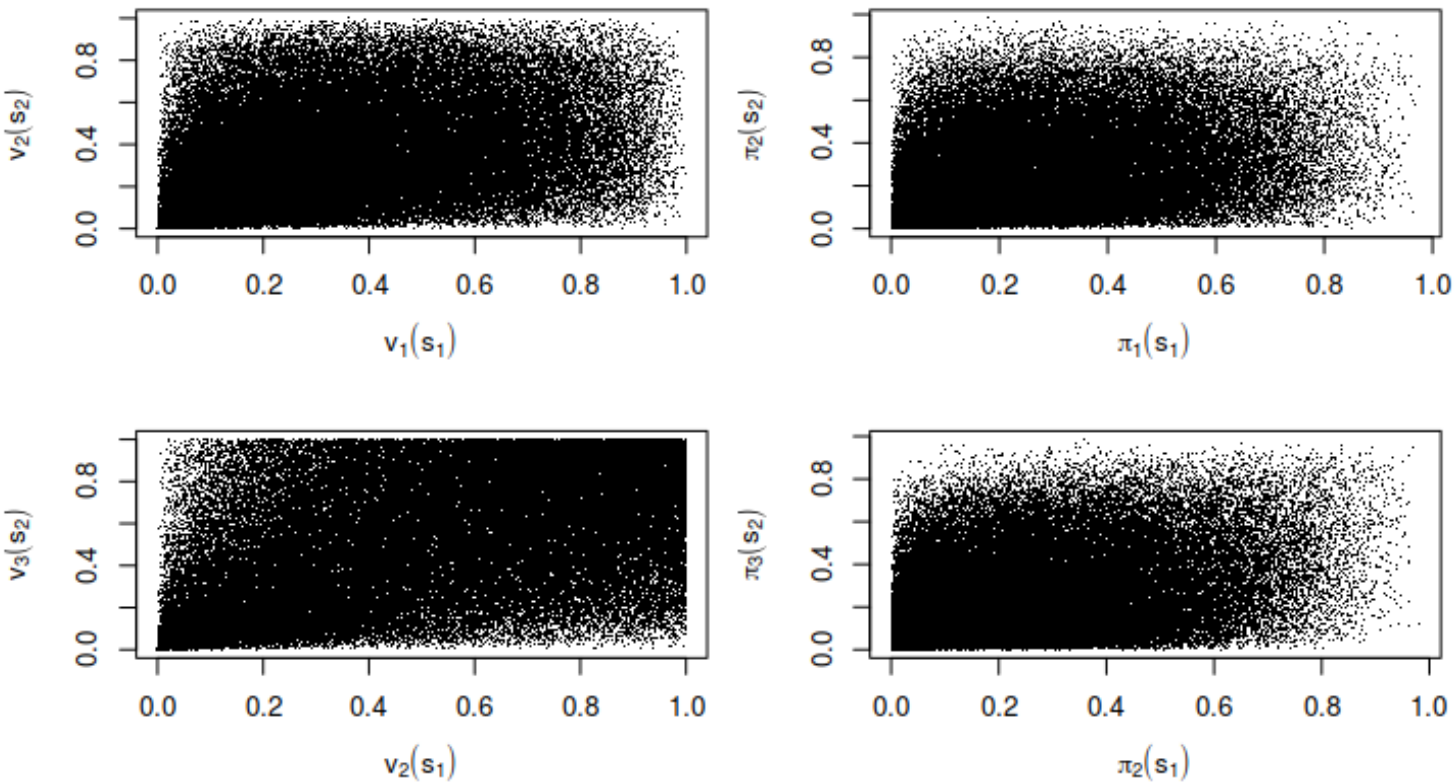
$$\pi_{t,1}(s) = V_{t,1}(s), \quad \pi_{t,k}(s) = V_{t,k}(s) \prod_{j=1}^{k-1} (1 - V_{t,j}(s)) \qquad \text{for } k = 2, \dots$$

$$V_{t,k}(s) = w_k(s, \psi, t, \zeta) V_k$$

$$V_k \sim Beta(a, b)$$

$$\theta_k \sim F_0.$$

A caveat


$v_2(s_2)$
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
$v_1(s_1)$
$\pi_2(s_2)$
$\pi_1(s_1)$
$v_3(s_2)$
$v_2(s_1)$
$\pi_3(s_2)$
$\pi_2(s_1)$

Conclusions

Consistently estimating the number of clusters in a Bayesian way is difficult.

However

- consistency can be found for overfitted mixtures
- the prior may have an important role
- advantage of reducing the number of necessary assumptions and inputs
- easy extension to multivariate setting

References

Grazian, C. and Robert, C.P. (2018) Jeffreys priors for mixture estimation: Properties and alternatives. *Computational Statistics & Data Analysis* 121, 149-163.

Mastrantonio, G., Grazian, C., Mancinelli, S., Bibbona, E. (2019) *New formulation of the logistic normal process to analyze tracking trajectories.*. *Annals of Applied Statistics*, 13(4),2483–2508.

Grazian, C., Villa, C. and Liseo, B. (2020) On a loss-based prior for the number of components in mixture models. *Statistics & Probability Letters* 158.

Grazian, C. (2022+) A review of Bayesian clustering based on mixture models.

Grazian, C. (2022+) Estimating MIC distributions and cutoffs through mixture models: an application to establish M. Tuberculosis resistance. *bioRxiv* 643429.

Grazian, C. (2022+) Spatio-temporal stick-breaking.