Introduction to Bayesian Nonparametric Statistics

Vanda Inácio

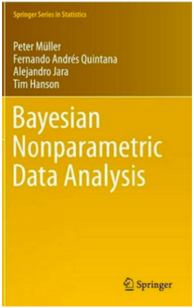
University of Edinburgh

December 2023

Outline

- $\hookrightarrow$ Introduction/motivation (why Bayesian nonparametric models?).
- $\hookrightarrow$ Finite mixture models.
- $\hookrightarrow$ Models for density estimation and regression based on Dirichlet process mixtures.

Some references (covers from `Amazon.co.uk`)

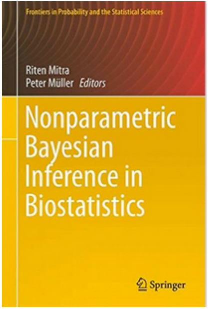


Springer Series in Statistics

Peter Müller
Fernando Andrés Quintana
Alejandro Jara
Tim Hanson

Bayesian Nonparametric Data Analysis

Springer

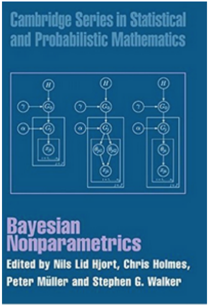


Frontiers in Probability and the Statistical Sciences

Riten Mitra
Peter Müller *Editors*

Nonparametric Bayesian Inference in Biostatistics

Springer



Cambridge Series in Statistical and Probabilistic Mathematics

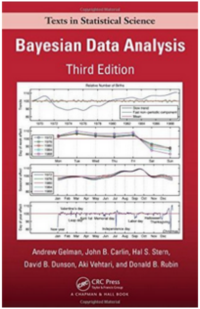
Bayesian Nonparametrics

Edited by Nils Lid Hjort, Chris Holmes, Peter Müller and Stephen G. Walker

Some references (covers from `Amazon.co.uk`)

Bayesian Non- and Semi-parametric Methods and Applications

Peter E. Rossi



Texts in Statistical Science

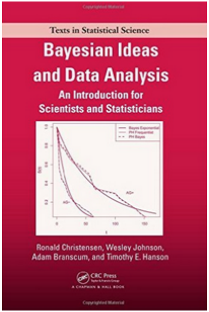
Bayesian Data Analysis

Third Edition

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin

CRC Press

A CHAPMAN & HALL BOOK



Texts in Statistical Science

Bayesian Ideas and Data Analysis

An Introduction for Scientists and Statisticians

Ronald Christensen, Wesley Johnson, Adam Branscum, and Timothy E. Hanson

CRC Press

A CHAPMAN & HALL BOOK

Some references

- Hanson, and Jara, A (2013). Surviving fully Bayesian nonparametric regression models. In *Bayesian Theory and Applications*, Oxford University Press, UK.

- Muller and Rodriguez (2013). *Nonparametric Bayesian Inference*. NSF–CBMS Regional Conference Series in Probability and Statistics, Volume 9.

  `http://projecteuclid.org/euclid.cbms/1362163742`

- Kottas and Rodriguez (2014). Unpublished lecture notes.

  `https://users.soe.ucsc.edu/ thanos/NPB_course_notes.pdf`

- Muller and Quintana (2004). Nonparametric Bayesian data analysis. *Statistical Science* **19**, 95–110.

- Muller and Mitra (2013). Bayesian nonparametric inference – why and how. *Bayesian Analysis* **8**, 269–302.

- Quintana et al. (2022). The dependent Dirichlet process and related Models. *Statistical Science* **37**, 24–41.

- Wade, Inacio, and Petrone (2023). Bayesian dependent mixture models: A predictive comparison and survey

  `https://arxiv.org/abs/2307.16298`

Why Bayesian nonparametrics?

Motivation

- $\hookrightarrow$ The motivation is twofold:
  - $\hookrightarrow$ Allowing for model flexibility.
  - $\hookrightarrow$ Safeguarding against model misspecification.
- $\hookrightarrow$ Bayesian nonparametrics rely on parametric baseline models while allowing for data-driven deviations.
- $\hookrightarrow$ So that we can see if parametric models might actually fit by embedding them in nonparametric families (sensitivity analysis).
- $\hookrightarrow$ Because Bayesian nonparametric modelling is feasible nowadays due to modern MCMC methods.

Why Bayesian nonparametrics?

Motivation

- ↪ The goal here is to provide a rich class of statistical models for data analysis.
- ↪ Data distributions can easily be multimodal or have severe skewness.
- ↪ The normal distribution is still widely used in practice.
- ↪ But the normal family is symmetric and has only two parameters, one corresponding to location and the other corresponding to dispersion.
- ↪ Common practice is to log transform the data. But if after the log, the data still have multimodality and/or skewness, the normal distribution would not be adequate.
- ↪ Other parametric distributions are similarly constrained.

Why Bayesian nonparametrics?

Motivation

- ↪ We thus proceed to discuss broader families that allow for flexibility and robustness beyond what is achievable using parametric models.
- ↪ This goal is accomplished by setting a particular parametric class and expanding it so that there are many more possibilities included.
- ↪ In the parametric Bayesian approach, given a particular dataset, a family of models is selected.
- ↪ For instance, in the normal case, this involves location and scale parameters.
- ↪ We then select prior distributions for these two parameters. This induces a prior on the family of distributions.

Why Bayesian nonparametrics?

Motivation

- ↪ In the nonparametric case, we do the same, except that, instead of having a small number of parameters that characterise the family of distributions for the data, conceptually, there is an infinite number of parameters.
- ↪ But since we live in a finite world, practicality dictates that these models must be truncated to have a possibly large but finite number of parameters. We call them richly parametric as a result.
- ↪ Thus, nonparametric Bayesian modelling requires a prior distribution on the (potentially) infinite dimensional parameters.
- ↪ Technically, this amounts to placing a prior distribution on the space of all distribution functions.

Why Bayesian nonparametrics?

Motivation

↪ We will be considering two approaches:

1. The first involves the specification of a Dirichlet process mixture model for the data distribution (covered today, Tuesday, 12/12).

2. The second approach involves the specification of a mixture of finite Polya trees prior on the space of all distribution functions (covered tomorrow, Wednesday, 13/12).

Finite mixture models

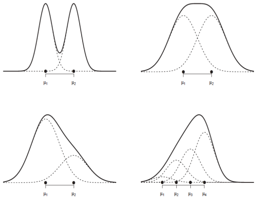
Motivation

- ↪ We start with finite mixture models since, to a certain extent, they provide the background for Dirichlet process mixture models.
- ↪ The natural question is: why mixture models?
- ↪ In a diversity of situations, the complexity of the observed data may render the use of a single parametric distribution insufficient for data modelling.
- ↪ For example, in the context of medical tests, biomarker outcomes for some specific disease may consist of, at least, two subgroups, corresponding to mild diseased and severe diseased individuals.

Finite mixture models

Motivation

- A mixture model assumes that data can be represented by a weighted sum of distributions, with each distribution representing a proportion of the data.
- It is worth nothing that multimodality is not the sole motivation for the use of mixture models. For instance, skewed data can also be handled by mixtures.
- Of course, for this latter case, one could use a skewed distribution, but this would imply that we expect skewness in advance, while with the more general mixture model, we can handle this and other nonstandard features of the data without the need to know them in advance.



source: Komarek, A., 2006, PhD thesis

Finite mixture models

Motivation

↪ In this app, you can play around with the parameters (weights, means, and variances) and check the shape of the resulting mixture (of normals) model.

`https://observablehq.com/@mattiasvillani/normal-mixture`

Finite mixture models

Formulation

- ↪ Throughout, we consider the particular case of mixtures of normal distributions.
- ↪ General mixtures of normal densities can accurately approximate any smooth density on the real line (e.g., Lo 1984).
- ↪ We assume thus that data follow a mixture of K normal distributions, whose density is given by

$$f(y \mid \boldsymbol{\omega}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \omega_k \phi(y \mid \mu_k, \sigma_k^2), \quad \omega_k \geq 0, \quad \sum_{k=1}^{K} \omega_k = 1, \tag{1}$$

  with $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$ and $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$.

- ↪ Hence, each data point would arise from one of K mixture components, with each component having its own mean and variance.

Finite mixture models

Formulation

- ↪ Model (1) is called a location-scale mixture of normals, since both the mean (location) and variance (scale) vary across components.

- ↪ An alternative model would treat the variances across components as equal to each other, that is,

$$f(y \mid \boldsymbol{\omega}, \mu_1, \ldots, \mu_K) = \sum_{k=1}^{K} \omega_k \phi(y \mid \mu_k, \sigma^2).$$

- ↪ Generally, location-scale mixtures of normal distributions produce more accurate results and also correspond to more realistic representations of the data.

Finite mixture models

Formulation

↪ Model (1) can be equivalently written as

$$f(y \mid \boldsymbol{\omega}, \boldsymbol{\theta}) = \int \phi(y \mid \mu, \sigma^2) \mathrm{d}G(\mu, \sigma^2).$$

↪ Here G is a discrete finite mixing distribution given by

$$G(\cdot) = \sum_{k=1}^{K} \omega_k \delta_{(\mu_k, \sigma_k^2)}(\cdot),$$

where δ_a denotes a point mass at a.

↪ The likelihood of the (observed) data is

$$L(\boldsymbol{\omega}, \boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} \omega_k \phi(y_i \mid \mu_k, \sigma_k^2) \right\}.$$

Finite mixture models

Data augmentation

↪ The likelihood is not 'very tractable' (due to its product-sums form) but this can be overcome by data augmentation.

↪ The same trick (data augmentation) is done by frequentists and leads to the well-known Expectation-Maximization algorithm.

↪ To this end, consider the latent (unobserved) random variable $z_i \in \{1, \ldots, K\}$ with $z_i = k$ indicating that observation y_i comes from component k, i.e., from the normal component with mean μ_k and variance σ_k^2.

↪ The mixture model can then be viewed hierarchically; the observed data y_i is modelled conditionally on z_i and z_i is also given a probabilistic specification.

↪ Indeed, we can write

$$\begin{aligned} y_i \mid z_i, \boldsymbol{\theta} &\overset{\text{ind.}}{\sim} \phi(y_i \mid \mu_{z_i}, \sigma_{z_i}^2), \quad i = 1, \ldots, n, \\ P(z_i = k \mid \boldsymbol{\omega}) &= \omega_k, \quad k = 1, \ldots, K. \end{aligned}$$

↪ If we marginalize over z we recover the original mixture formulation.

Finite mixture models

Likelihood

- ↪ By introducing $\mathbf{z} = (z_1, \ldots, z_n)'$ we obtain the complete/augmented data likelihood, which is given by

$$\begin{aligned} L(\boldsymbol{\omega}, \boldsymbol{\theta}, \mathbf{z}; \mathbf{y}) &= \prod_{i=1}^{n} \omega_{z_i} \phi(y_i \mid \mu_{z_i}, \sigma^2_{z_i}) \\ &= \prod_{k=1}^{K} \prod_{i:z_i=k} \omega_k \phi(y_i \mid \mu_k, \sigma^2_k) \\ &= \prod_{k=1}^{K} \omega_k^{n_k} \prod_{i:z_i=k} \phi(y_i \mid \mu_k, \sigma^2_k). \end{aligned} \tag{2}$$

- ↪ Here $n_k = \sum_{i=1}^{n} I(z_i = k)$ counts the number of observations allocated to component k.
- ↪ By introducing the indicators $z_1, \ldots, z_n$ we have transformed the sum that appears in the definition of the density of a mixture into a product over the components.

Finite mixture models

Prior distributions

- ↪ The model specification is completed by specifying the prior distribution for the weights, means, and variances
$$(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \sim p(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2).$$
- ↪ We consider prior independence and thus
$$p(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = p(\boldsymbol{\omega})p(\boldsymbol{\mu})p(\boldsymbol{\sigma}^2).$$
- ↪ We will make use of conjugate priors. Specifically, the weights are assigned a Dirichlet distribution
$$(\omega_1, \ldots, \omega_K) \sim \text{Dir}(\alpha_1, \ldots, \alpha_K).$$
- ↪ In turn, the means and variances of each component are assigned normal and inverse-gamma prior distributions, respectively. That is,
$$\mu_k \sim \text{N}(a_\mu, b_\mu^2), \quad \sigma_k^2 \sim \text{IG}(a_{\sigma^2}, b_{\sigma^2}), \quad k = 1, \ldots, K.$$
- ↪ Note that placing a prior on the collection $(\{\omega_k\}, \{\mu_k, \sigma_k^2\})$ is equivalent to placing a prior on the finite mixing distribution G.

Finite mixture models

Prior distributions

- ↪ The Dirichlet distribution is new to us and deserves a slide. It can be regarded as the multivariate extension of the beta distribution.

- ↪ More information can be found on Wikipedia

  `https://en.wikipedia.org/wiki/Dirichlet_distribution`

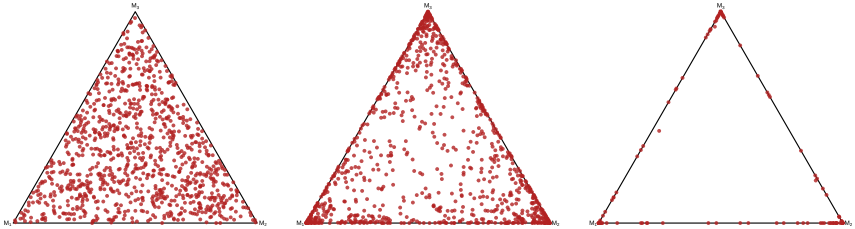
- ↪ A usual choice for the hyperparameters of the DIrichlet distribution, when used as a prior for the components' weights is $\alpha_1 = \ldots = \alpha_K = \alpha$.

- ↪ A small value for α encourages sparsity in the weights and in the extreme case when $\alpha \to 0$, all prior mass is placed on the vertices of the simplex, with all weight on a single component.

Finite mixture models

Prior distributions

$\hookrightarrow$ Left: $(\alpha_1, \alpha_2, \alpha_3) = (1, 1, 1)$. Center: $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.3, 0.3)$. Right: $(\alpha_1, \alpha_2, \alpha_3) = (0.01, 0.01, 0.01)$.



Finite mixture models

Posterior inference

- By combining the likelihood in (2) with this prior specification, the joint posterior distribution is

$$p(\boldsymbol{\omega}, \boldsymbol{\theta}, \mathbf{z} \mid \mathbf{y}) \propto L(\boldsymbol{\omega}, \boldsymbol{\theta}; \mathbf{y}, \mathbf{z}) p(\boldsymbol{\omega}) \prod_{k=1}^{K} p(\mu_k) p(\sigma_k^2) \tag{3}$$

- Although the posterior distribution in (3) does not have a recognisable form, all full conditional distributions have simple conjugate forms.

- Specifically, for the mean and variance of the components we have

$$\mu_k \mid \text{else} \sim \text{N}\left(\frac{a_\mu/b_\mu^2 + \sum_{i:z_i=k} y_i/\sigma_k^2}{1/b_\mu^2 + n_k/\sigma_k^2}, \frac{1}{1/b_\mu^2 + n_k/\sigma_k^2}\right), \tag{4}$$

$$\sigma_k^2 \mid \text{else} \sim \text{IG}\left(a_{\sigma^2} + n_k/2, b_{\sigma^2} + \sum_{i:z_i=k} (y_i - \mu_k)^2/2\right), \tag{5}$$

for $k = 1, \ldots, K$.

Finite mixture models

Posterior inference

$\hookrightarrow$ For $i = 1, \ldots, n$, for the full conditional distribution for z_i we have that

$$\Pr(z_i \mid z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, \boldsymbol{\omega}, \boldsymbol{\theta}, y_i) \propto p(y_i \mid z_i, \boldsymbol{\theta}) \Pr(z_i \mid \boldsymbol{\omega})$$

and hence

$$\Pr(z_i = k \mid z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, \boldsymbol{\omega}, \boldsymbol{\theta}, y_i) = \frac{\omega_k \phi(y_i \mid \mu_k, \sigma_k^2)}{\sum_{l=1}^{K} \omega_l \phi(y_i \mid \mu_l, \sigma_l^2)}, \quad k = 1, \ldots, K. \tag{6}$$

$\hookrightarrow$ The density function of the Dirichlet distribution is given by

$$p(\boldsymbol{\omega}) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \omega_k^{\alpha_k - 1}, \quad B(\alpha_1, \ldots, \alpha_K) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)},$$

and therefore the full conditional of the weights is given by

$$\boldsymbol{\omega} \mid \text{else} \sim \text{Dir}(\alpha_1 + n_1, \ldots, \alpha_K + n_K).$$

Finite mixture models

Gibbs sampler algorithm

1. Set initial values $\boldsymbol{\omega}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$.

2. For $s = 1, \ldots, S$

   $\hookrightarrow$ For $i = 1, \ldots, n$ and $k = 1, \ldots, K$, sample $z_i^{(s)}$ from its full conditional in (6).

   $\hookrightarrow$ Conditional on $\mathbf{z}^{(s)}$, update $\boldsymbol{\omega}$

$$\boldsymbol{\omega}^{(s)} \sim \text{Dirichlet}\left(\alpha_1 + n_1^{(s)}, \ldots, \alpha_K + n_K^{(s)}\right), \quad n_1^{(s)} = \sum_{i=1}^{n} z_i^{(s)}.$$

   $\hookrightarrow$ Conditional on $\mathbf{z}^{(s)}$, update $\mu_k^{(s)}$ and $(\sigma_k^{(s)})^2$, for $k = 1, \ldots, K$, from (4) and (5), respectively.

Finite mixture models

How to choose K?

- ↪ A 'drawback' of finite mixture models is that we must choose the number of components K, which is a non-trivial task in general.

- ↪ One possibility is placing a prior on K (e.g., a truncated Poisson or a discrete uniform distribution), which leads to a model that changes dimension (number of parameters) with K.

- ↪ One approach to fit such model is to use reversible jump type of algorithms, but these type of algorithms tend to be difficult to implement efficiently in practice.

Finite mixture models

How to choose K?

- Another possibility is to fit the model for different values of K and assess the adequacy of the fit through, for instance, the LPML and WAIC criteria. Predictive checks can also aid in the choice of K.

- Placing a K too small is much worst than placing a K too large.

- For instance, one cannot get a trimodal density out of a mixture of two components but one can essentially ignore extra mixture terms and get bimodal densities from mixtures of three or more components.

Finite mixture models

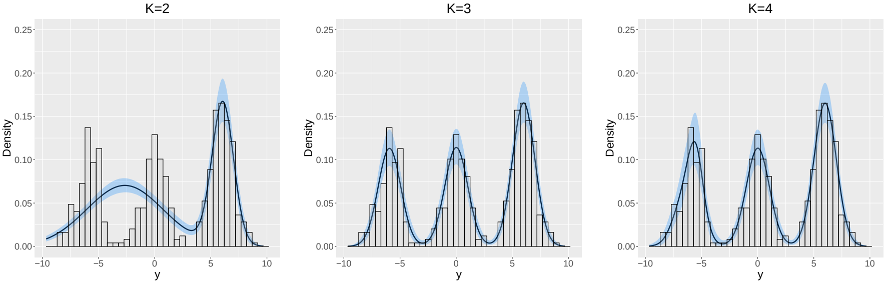
How to choose K?

- $\hookrightarrow$ An alternative approach is to use the so-called overfitted mixtures (Rousseau and Mengersen, 2011).
- $\hookrightarrow$ The idea behind overfitted mixtures is to saturate the model with a large number of components, which can be regarded as an upper bound on the number of occupied mixture components or clusters.
- $\hookrightarrow$ The problem with a large number of mixture components is that different components that are very similar and hence redundant may be introduced, leading to a degrading of the model performance.
- $\hookrightarrow$ Some form of sparsity is therefore essential in order to effectively regularise and prune the extra, redundant,components.
- $\hookrightarrow$ With this in mind, Rousseau and Mengersen (2011) propose a prior distribution for the weights that is still a Dirichlet distribution but the values of $\alpha_1, \ldots, \alpha_K$ are specified in such a way that the resulting distribution favours either emptying or merging the extra redundant components.

Finite mixture models

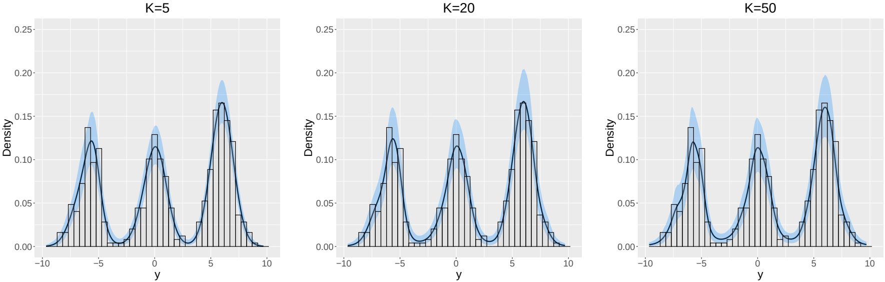
Example

True data generating distribution: $0.3\phi(y \mid -6, 1) + 0.3\phi(y \mid 0, 1) + 0.4\phi(y \mid 6, 1)$, $n = 500$.



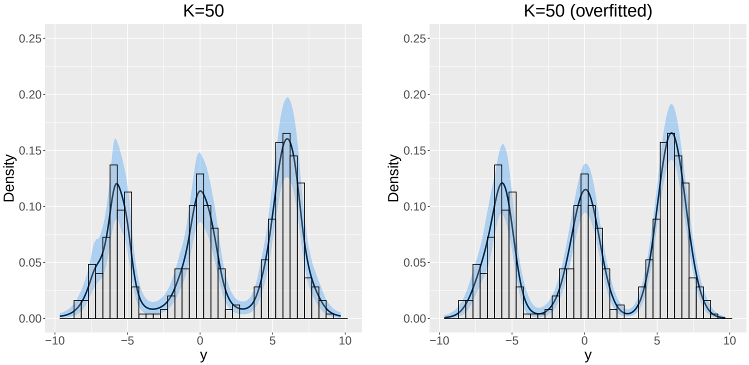
Finite mixture models

Example



Finite mixture models

Example



Finite mixture models

Example

- Let us now compute the LPML and WAIC for the different values of K. Of course we know K is 3 but we want to see how these two criteria perform when we know the truth.
- Remember that for LPML the higher its value, the better, while for the WAIC is the other way around.

	LPML	WAIC
$K = 2$	-1379	2758
$K = 3$	-1260	2521
$K = 4$	-1260	2519
$K = 5$	-1260	2525
$K = 20$	-1266	2532
$K = 50$	-1279	2558
$K = 50$ (overfitted)	-1264	2528

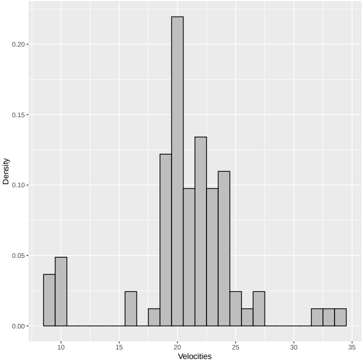
- Clearly, based on these two criteria, the model with $K = 2$ would be excluded. Models with $K = 3$ and $K = 4$ are practically indistinguishable (from the WAIC/LPML point of view). When differences in the model selection criteria are small (e.g., less than 5) we should opt for the most parsimonious model.
- Of course, these results are only based on one simulation and to formally judge the performance of such criteria in picking up the right number of components, one would need to simulate a large number of datasets. You can try changing the seed twice and see what happens.

Finite mixture models

Example: galaxy data

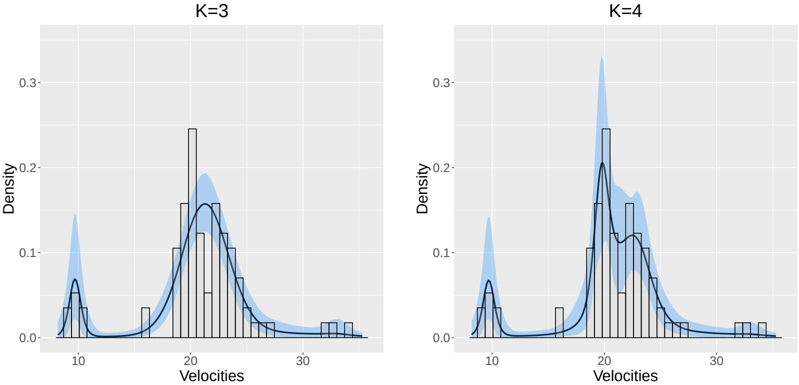
- A classic example in the literature of finite mixture models is the galaxy data, containing the velocities of 82 galaxies.

- The dataset is available, among other packages, in the `MASS` package.



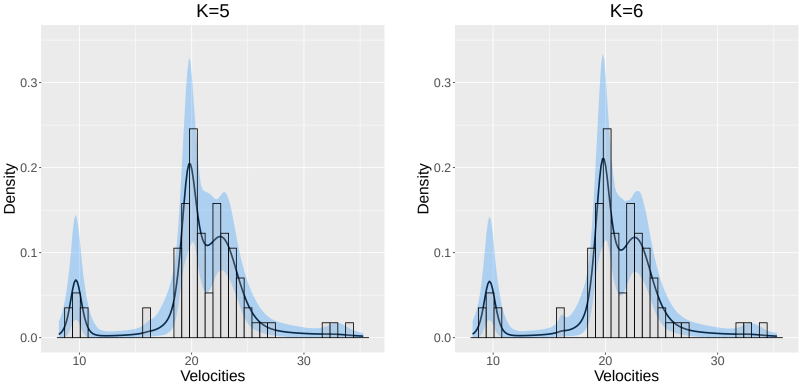
Finite mixture models

Example: galaxy data



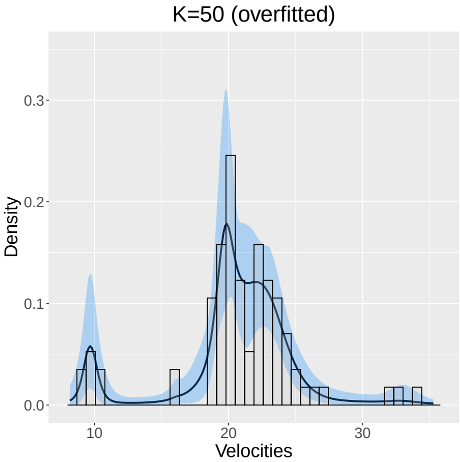
Finite mixture models

Example: galaxy data



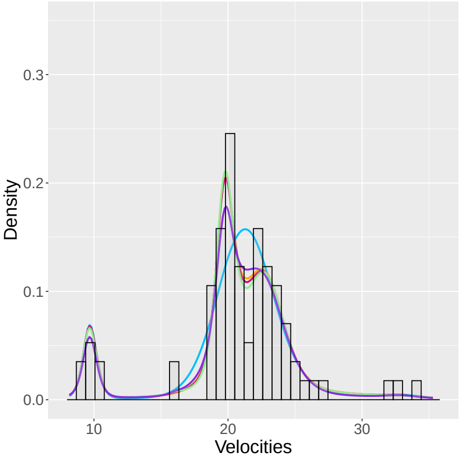
Finite mixture models

Example: galaxy data



Finite mixture models

Example: galaxy data



$K = 3$: Blue line. $K = 4$: Orange line. $K = 5$: Pink line. $K = 6$: green line. $K = 50$ (overfitted): purple line

Finite mixture models

Example: galaxy data

$\hookrightarrow$ Let us compute the LPML/WAIC for each fitted model.

	LPML	WAIC
$K = 3$	-214	428
$K = 4$	-212	424
$K = 5$	-213	425
$K = 6$	-213	426
$K = 50$ (overfitted)	-218	436

$\hookrightarrow$ Both criteria seem to favour the model with 4, 5 or 6 components.

Finite mixture models

Identifiability issues

- A probability model is identifiable if and only if different values of the parameters generate different probability distributions of the observable variables.

- Due to identifiability issues, such as the so-called label switching problem (Jasra et al. 2005), it makes a difference whether there is interest in making inferences about the mixture component-specific parameters and clustering.

- The label switching problem (also known as label ambiguity) refers to the fact that there is nothing in the likelihood to distinguish mixture component k from mixture component k'.

- Permuting the K labels in any of $K!$ ways results in the same model for the data.

Finite mixture models

Identifiability issues

- As a concrete example, in the $K = 2$ case, consider

$$p_1 = 0.3, \quad \mu_1 = 1, \quad p_2 = 0.7, \quad \mu_2 = 1.5, \quad \sigma_1^2 = \sigma_2^2 = 1, \qquad \text{(Scenario A)}.$$

- Then, the model is equivalent to one with

$$p_1 = 0.7, \quad \mu_1 = 1.5, \quad p_2 = 0.3, \quad \mu_2 = 1, \quad \sigma_1^2 = \sigma_2^2 = 1, \qquad \text{(Scenario B)}.$$

- If one is only interested in density estimation, then everything is fine, because as illustrated in the example

$$\begin{aligned} f_{\text{A}}(y) &= 0.3\phi(y \mid 1, 1) + 0.7\phi(y \mid 1.5, 1) \\ &= 0.7\phi(y \mid 1.5, 1) + 0.3\phi(y \mid 1, 1) = f_{\text{B}}(y). \end{aligned}$$

- Of course, if we are 'only' interested in estimating the density, the label switching poses no problem.

Finite mixture models

Identifiability issues

- ↪ As stated by Peter Green (Chapter 1, Handbook of Mixture Analysis, 2018), *"In an MCMC run, the lack of coherence between the numberings can apply separately at each iteration, so commonly switchings are observed".*

- ↪ Numerous solutions to this problem have been (and still are) proposed, but there is no universal approach.

- ↪ One way to overcome this issue is to enforce identifiability in the prior. For instance, in one dimensional problems, we can order the component means, that is, $\mu_1 < \mu_2 < \ldots < \mu_K$.

Dirichlet process mixtures

DP prior

- ↪ As we have just seen, choosing the number K of mixture components is not trivial. And letting K to be random (by placing a prior on it) is computationally complex.

- ↪ Overfitted mixtures are a good solution; however, they were introduced only after the development of Dirichlet process mixtures.

- ↪ Another alternative is to use a Dirichlet process (DP) prior (Ferguson 1973, 1974) for the mixing distribution G, resulting in a Dirichlet process mixture (DPM) model.

- ↪ The parametric discrete distribution for G is therefore replaced with a DP prior that supports all mixing distributions.

- ↪ The resulting model is

$$f(y) \equiv f(y \mid G) = \int k(y \mid \boldsymbol{\theta}) \mathrm{d}G(\boldsymbol{\theta}), \quad G \sim \mathrm{DP}(\alpha, G_0).$$

- ↪ In the context of DPM of normals, $k(y \mid \boldsymbol{\theta}) = \phi(y \mid \boldsymbol{\theta})$, $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Dirichlet process mixtures

DP prior

- ↪ But what does it to write $G \sim \mathrm{DP}(\alpha, G_0)$?

- ↪ The random distribution G is said to follow a Dirichlet process with parameters α and G_0 if, for any finite partition $A_1, \ldots, A_k$ of the sample space, the probability vector $[G(A_1), \ldots, G(A_k)]$ follows a Dirichlet distribution

$$[G(A_1), \ldots, G(A_k)] \sim \mathrm{Dir}[\alpha G_0(A_1), \ldots, \alpha G_0(A_k)].$$

Dirichlet process mixtures

DP prior

- The definition of the Dirichlet process and the properties of the Dirichlet distribution imply that for any subset A of the sample space

$$G(A) \sim \text{Beta}\{\alpha G_0(A), \alpha(1 - G_0(A))\}.$$

- Thus,

$$\mathbb{E}\{G(A)\} = G_0(A), \quad \text{Var}\{G(A)\} = \frac{G_0(A)(1 - G_0(A))}{\alpha + 1}.$$

- Hence, from the above, we have
  - G is centred on G_0 (G_0 is also referred as baseline distribution). G_0 encapsulates any prior knowledge that might be known about G.
  - $\alpha > 0$ is a precision parameter controlling the variance of the process. If α is large, G is highly concentrated around G_0.
- Therefore, the DP prior is centred on a parametric model through the specification of G_0, while allowing α to control uncertainty in this choice.

Dirichlet process mixtures

Conjugacy of the DP prior

- The DP prior is closed under sampling (Ferguson, 1973). That is, the posterior distribution is also a DP.

- Assume $y_1, \ldots, y_n \mid G \overset{\text{iid}}{\sim} G$ and $G \sim \text{DP}(\alpha, G_0)$. Then

$$G \mid y_1, \ldots, y_n \sim \text{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{y_i}\right)$$

- The posterior mean is then

$$E\{G(y) \mid y_1, \ldots, y_n\} = \frac{\alpha}{\alpha + n} G_0(y) + \frac{n}{\alpha + n} G_n(y),$$

where $G_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \leq y)$ is the empirical cdf.

- Thus, the posterior mean is a weighted average between the centring distribution and the empirical cdf.

Dirichlet process mixtures

Conjugacy of the DP prior

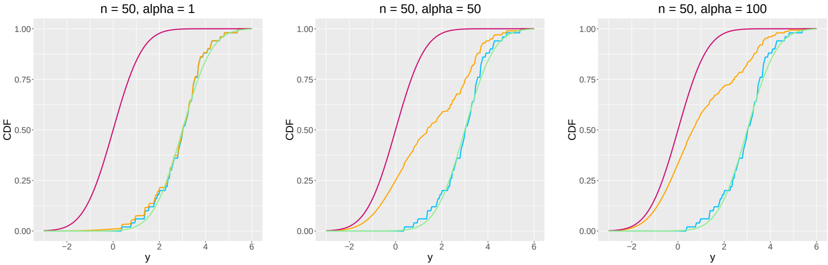
- For small α relative to *n*, little weight is placed on the centring distribution G_0.
- For large α relative to *n*, little weight is placed on the data.
- From Athanasios Kottas notes it can be read:

  > *"Hence, α can be viewed as a measure of faith in the prior guess G_0 measured in units of number of observations (thus, $\alpha = 1$ indicates strength of belief in G_0 worth one observation)."*

Dirichlet process mixtures

Conjugacy of the DP prior

$\hookrightarrow$ Estimating the posterior mean under a DP prior using simulated data. The true distribution generating the data is $N(3, 1)$, while the centring is a $N(0, 1)$ distribution.

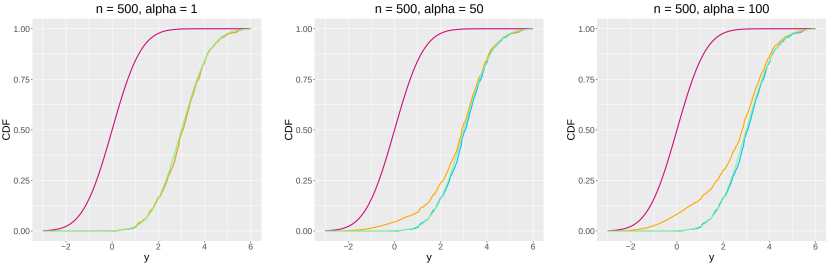


ECDF: Blue line. $\mathbb{E}(G \mid y)$: Orange line. $\mathbb{E}(G) = G_0$: Pink line. TRUE: green line.

Dirichlet process mixtures

Conjugacy of the DP prior

- ↪ Estimating the posterior mean under a DP prior using simulated data. The true distribution generating the data is $N(3, 1)$, while the centring is a $N(0, 1)$ distribution.



ECDF: Blue line. $\mathbb{E}(G \mid y)$: Orange line. $\mathbb{E}(G) = G_0$: Pink line. TRUE: green line.

Dirichlet process mixtures

Conjugacy of the DP prior/ Bayesian bootstrap

- When $\alpha \to 0$, the limiting posterior distribution is

$$G \mid y_1, \ldots, y_n \sim \text{DP}\left(n, \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}\right),$$

and thus

$$E(G \mid y_1, \ldots, y_n) = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}$$

- This limiting posterior distribution is known as the Bayesian bootstrap (Rubin 1981, Gasparini 1995).

- A Bayesian bootstrap estimator for G can be computed as

$$G = \sum_{i=1}^{n} p_i\delta_{y_i}, \quad (p_1, \ldots, p_n) \sim \text{Dir}(n; 1, \ldots, 1). \tag{7}$$

- Again, from (7) it follows that

$$E(G \mid y_1, \ldots, y_n) = E\left(\sum_{i=1}^{n} p_i\delta_{y_i}\right) = \sum_{i=1}^{n} E(p_i)\delta_{y_i} = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}$$

Dirichlet process mixtures

Conjugacy of the DP prior/ Bayesian bootstrap

- ↪ Recall that in the frequentist bootstrap (Efron, 1979) $p_i \in \{0, 1/n, \ldots, n/n\}$ corresponding to the number of times y_i appears in a bootstrap sample.

- ↪ Thus, the weights in the Bayesian bootstrap are smoother than those from Efron's frequentist bootstrap.

- ↪ This is justified by the fact that in the BB the weights arise from a Dirichlet (continuous distribution) while the weights in the classical bootstrap have a discrete distribution.

- ↪ Note that in the BB the data should be regarded as fixed, so that we do not resample from it.

Dirichlet process mixtures

Stick-breaking representation of the DP

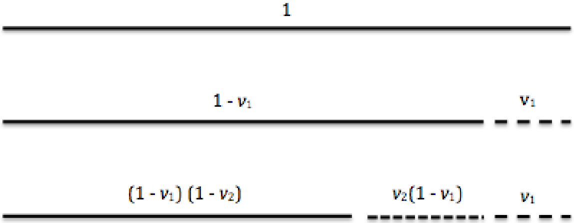
- Although useful, the definition of the DP does not provide an intuition for what realisations of a DP actually look like.
- Undoubtedly, the most useful definition of the DP is its constructive definition (Sethuraman, 1994), according to which G has an almost sure representation of the form

$$G(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\boldsymbol{\theta}_k}(\cdot). \tag{8}$$

- In (8), we have
  - $\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} G_0$, $k = 1, 2, \ldots$
  - $\omega_1 = v_1$, and for $k \geq 2$, $\omega_k = v_k \prod_{l<k}(1 - v_l)$, with $v_k \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$ (independently of the θ_k) $\Rightarrow$ Stick-breaking construction.

Dirichlet process mixtures

Stick-breaking representation of the DP



↪ $\omega_1 = v_1$, and for $k \geq 2$, $\omega_k = v_k \prod_{l<k}(1 - v_l)$, with $v_k \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$ (independently of the θ_k) $\Rightarrow$ Stick-breaking construction.

Dirichlet process mixtures

Stick-breaking representation of the DP

- ↪ As the stick-breaking construction proceeds, the stick gets shorter and shorter and the lengths allocated to higher index atoms decrease stochastically, with the rate of decrease depending on α.

- ↪ Since $v_k \sim \text{Beta}(1, \alpha)$, then

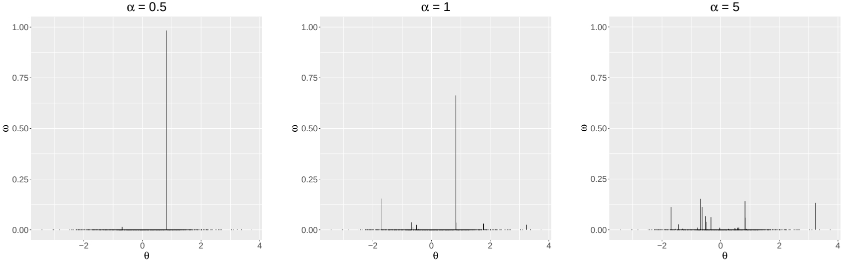
$$E(v_k) = \frac{1}{1+\alpha},$$

so that values of α close to zero lead to high weight on the first few atoms, with the remaining atoms being assigned small probabilities.

Dirichlet process mixtures

Stick-breaking representation of the DP

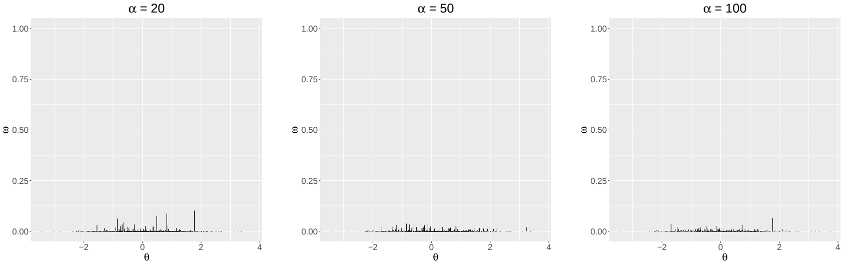
↪ Centring distribution is the standard normal (based on 1000 draws).



Dirichlet process mixtures

Stick-breaking representation of the DP

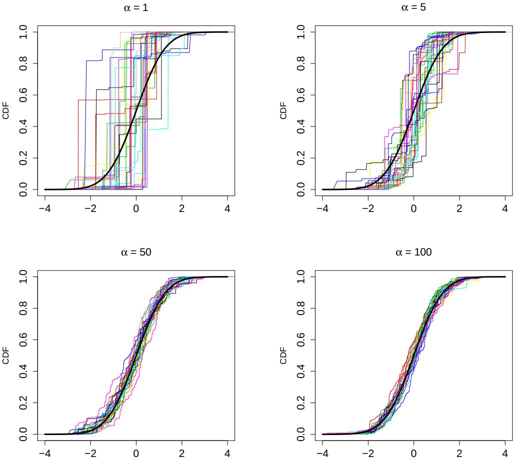
↪ Centring distribution is the standard normal (based on 1000 draws).



Dirichlet process mixtures

Stick-breaking representation of the DP

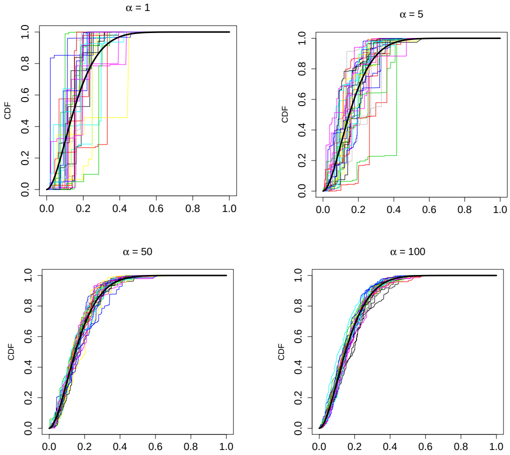
↪ 30 trajectories from a DP with $G_0 = \mathrm{N}(0,1)$ (based on 1000 draws).



Dirichlet process mixtures

Stick-breaking representation of the DP

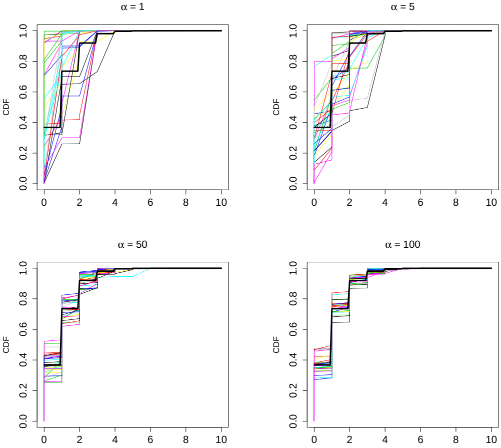
↪ 30 trajectories from a DP with $G_0 = \text{Beta}(2, 10)$ (based on 1000 draws).



Dirichlet process mixtures

Stick-breaking representation of the DP

$\hookrightarrow$ 30 trajectories from a DP with $G_0 = \text{Pois}(1)$ (based on 1000 draws).



Dirichlet process mixtures

DPM model

- The DP generates flexible albeit discrete distributions and so it is not suitable to model continuous data.

- Due to this fact, and as we have already anticipated, it is commonly used as a prior for the mixing distribution in mixture models.

- The Dirichlet process mixture model can be specified as

$$f(y) \equiv f(y \mid G) = \int k(y \mid \boldsymbol{\theta}) \mathrm{d}G(\boldsymbol{\theta}).$$

- The mixture density can be discrete or continuous, univariate or multivariate, depending on the nature of the kernel $k(y \mid \boldsymbol{\theta})$.

- The corresponding mixture distribution is

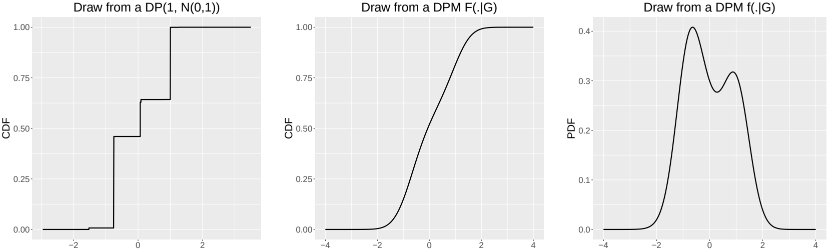
$$F(y) = \int K(y \mid \boldsymbol{\theta}) \mathrm{d}G(\boldsymbol{\theta}),$$

where $K(\cdot \mid \boldsymbol{\theta})$ is the cdf corresponding to the density/mass function $k(\cdot \mid \boldsymbol{\theta})$.

Dirichlet process mixtures

DPM model

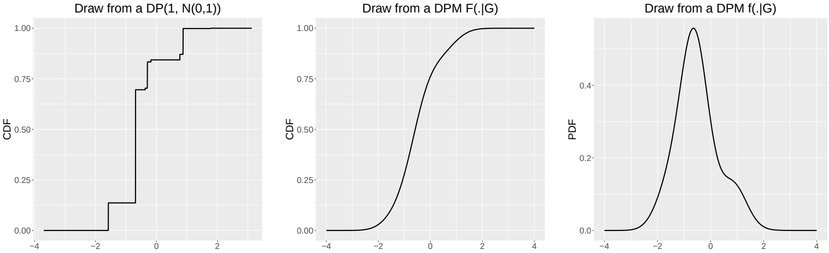
- ↪ Realisation from a DP($\alpha = 1, G_0 = \mathrm{N}(0, 1)$) (left) and associated cumulative distribution function (center) and probability density function (right)) for a location DP mixture of normal kernels with standard deviation 0.5.



Dirichlet process mixtures

DPM model

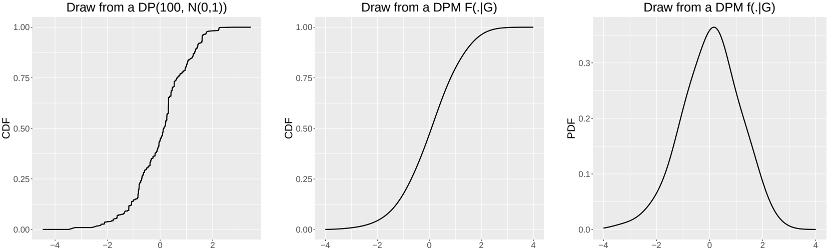
↪ Realisation from a DP($\alpha = 1, G_0 = \mathrm{N}(0,1)$) (left) and associated cumulative distribution function (center) and probability density function (right)) for a location DP mixture of normal kernels with standard deviation 0.5.



Dirichlet process mixtures

DPM model

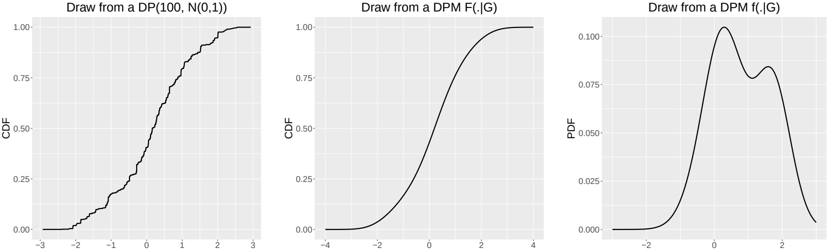
↪ Realisation from a DP($\alpha = 100$, $G_0 = \mathrm{N}(0, 1)$) (left) and associated cumulative distribution function (center) and probability density function (right)) for a location DP mixture of normal kernels with standard deviation 0.5.



Dirichlet process mixtures

DPM model

- ↪ Realisation from a DP($\alpha = 100$, $G_0 = \text{N}(0,1)$) (left) and associated cumulative distribution function (center) and probability density function (right)) for a location DP mixture of normal kernels with standard deviation 0.5.



Dirichlet process mixtures

DPM model

- Hereby, our main focus is on Dirichlet process mixtures of normals

$$f(y) = \int \phi(y \mid \mu, \sigma^2)\mathrm{d}G(\mu, \sigma^2), \quad G \sim \mathrm{DP}(\alpha, G_0). \tag{9}$$

- The centring distribution G_0 is defined on $\mathbb{R} \times \mathbb{R}^+$.
- Due to conjugacy reasons, G_0 is usually taken to be the normal-inverse-gamma distribution, that is

$$G_0 \equiv \mathrm{N}(a_\mu, b_\mu^2)\mathrm{IG}(a_{\sigma^2}, b_{\sigma^2}).$$

- To allow for extra flexibility, hyperpriors can be placed on a_μ, b_μ^2, a_{σ^2}, b_{σ^2}.
- The stick-breaking representation of the DP allows us to write (9) as the following countably infinite mixture of normals

$$f(y) = \sum_{l=1}^{\infty} \omega_l \phi(y \mid \mu_l, \sigma_l^2). \tag{10}$$

- Note that equation (10) resembles the finite mixture model considered earlier but with the important difference that the number of mixture components is set to infinite and the weights now follow a stick-breaking construction.

Dirichlet process mixtures

Posterior inference

- Posterior inference can be conducted using different kinds of MCMC strategies.
- Most popular MCMC approaches are based on:
  - Algorithms that make use of the Polya urn representation of the DP and that sample the marginal distribution of the mixing distribution that arises after marginalising G over its DP prior.
  - Blocked Gibbs samplers that rely on truncation approximations to G.
  - Retrospective sampling, which is a modification of the blocked Gibbs sampler that allows one to add components adaptively, if needed, but not to eliminate any, as the MCMC progresses.
- Throughout the course the blocked Gibbs sampler will be detailed.

Dirichlet process mixtures

Blocked Gibbs sampler

- The blocked Gibbs sampler relies on truncating the stick-breaking representation to a finite number of components, say L, that is

$$G_L(\cdot) = \sum_{l=1}^{L} \omega_l \delta_{(\mu_l, \sigma_l^2)}(\cdot).$$

- The atoms (μ_l, σ_l^2) are iid G_0, i.e., $(\mu_l, \sigma_l^2) \overset{\text{iid}}{\sim} \text{N}(a_\mu, b_\mu^2)\text{IG}(a_{\sigma^2}, b_{\sigma^2})$, $l = 1, \ldots, L$.

- The weights arise through a truncated stick-breaking construction

$$\omega_1 = v_1, \quad \text{for } l \geq 2 \;\; \omega_l = v_l \prod_{m<l}(1 - v_m), \quad v_l \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha), \;\; l = 1, \ldots, L-1, \text{ and } v_L = 1$$

Dirichlet process mixtures

Blocked Gibbs sampler

- Using the truncated version G_L of G, the mixture density can be expressed as

$$f(y) = \sum_{l=1}^{L} \omega_l \phi(y \mid \mu_l, \sigma_l^2),$$

  with ω_l generated from the truncated stick-breaking representation, whereas $\mu_l \overset{\text{iid}}{\sim} \text{N}(a_\mu, b_\mu^2)$, and $\sigma_k^2 \overset{\text{iid}}{\sim} \text{IG}(a_{\sigma^2}, b_{\sigma^2})$.

- A valid concern with the blocked Gibbs sampler is that by truncating the stick-breaking representation we are effectively fitting a finite (and hence parametric) mixture model.

- Quoting Dunson (2011):

> "*For example, if we let* $L = 25$ *as a truncation level, a natural question is how this is better or intrinsically different than fitting a finite mixture model with 25 components. One answer is that L is not the number of components occupied by the subjects in your sample but is instead an upper bound on the number of subjects.*

Dirichlet process mixtures

Blocked Gibbs sampler

- ↪ Note that, given the v's, the weights are deterministic and $v_l \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$, for $l = 1, \ldots, L-1$.

- ↪ The value of α can either be fixed or a prior distribution placed on it. Usually, for conjugacy reasons, $\alpha \sim \Gamma(a_\alpha, b_\alpha)$.

- ↪ Also, the value of α has a direct relationship with the number of occupied mixture components, say L^*.

- ↪ In fact, Liu et al. (1996) showed that for moderate to large sample sizes, the conditional prior mean and variance of the number of active components are, respectively,

$$\mathbb{E}(L^* \mid \alpha) \approx \alpha \log\left(\frac{\alpha + n}{\alpha}\right), \quad \text{var}(L^* \mid \alpha) \approx \alpha \left\{\log\left(\frac{\alpha + n}{\alpha}\right) - 1\right\}.$$

- ↪ These results allow to use genuine prior information to derive values for α, or a_α and b_α.

- ↪ In practice, it is common to either set $\alpha = 1$ or to use $\alpha \sim \Gamma(1, 1)$, $\alpha \sim \Gamma(2, 2)$, or $\alpha \sim \Gamma(2, 4)$, or any other values that encourage, a prior, a small number of occupied components.

Dirichlet process mixtures

Blocked Gibbs sampler

- ↪ An appropriate value of L can be determined by considering properties of the high-order ω_l values in the infinite tick-breaking representation.

- ↪ Ishwaran and Zarepour (2000) shown that

$$\mathbb{E}\left(\sum_{l=L+1}^{\infty}\omega_l \mid \alpha\right) = \left(\frac{\alpha}{1+\alpha}\right)^L.$$

- ↪ For example, setting $\alpha = 1$ and $L = 10$ leads to $\mathbb{E}(\sum_{l=L+1}^{\infty}\omega_l) \approx 0.00098$, so the truncation error is about 0.1%.

- ↪ Furthermore, the information provided earlier on the prion mean and variance of the number of occupied components can also be used to guide the choice of L.

- ↪ For instance, it might seem reasonable to set $L > \mathbb{E}(L^* \mid \alpha) + 2\sqrt{\text{var}(L^* \mid \alpha)}$.

Dirichlet process mixtures

Blocked Gibbs sampler

- ↪ As it was the case for the finite mixture model, derivation of the full conditionals for Gibbs sampling involves the data-augmented likelihood.

- ↪ The means, variances, and latent component indicators are sampled in an identical manner to the finite mixture model.

- ↪ The main difference is that, unlike in the finite mixture model, uncertainty in the component weights is shifted to $\mathbf{v}$, the inputs into the construction of the stick-breaking weights.

Dirichlet process mixtures

Blocked Gibbs sampler

↪ The full conditional distributions are

$$\mu_l \mid \text{else} \sim \text{N}\left(\frac{a_\mu/b_\mu^2 + \sum_{i:z_i=l} y_i/\sigma_l^2}{1/b_\mu^2 + n_l/\sigma_l^2}, \frac{1}{1/b_\mu^2 + n_l/\sigma_l^2}\right), \tag{11}$$

$$\sigma_l^2 \mid \text{else} \sim \text{IG}\left(a_{\sigma^2} + n_l/2, b_{\sigma^2} + \sum_{i:z_i=l} (y_i - \mu_l)^2/2\right), \tag{12}$$

for $l = 1, \ldots, L$.

↪ For $i = 1, \ldots, n$, the full conditional distribution for z_i is

$$\Pr(z_i = l \mid \text{else}) = \frac{\omega_l \phi(y_i \mid \mu_l, \sigma_l^2)}{\sum_{l=1}^{L} \omega_l \phi(y_i \mid \mu_l, \sigma_l^2)}, \quad l = 1, \ldots, L.$$

Dirichlet process mixtures

Blocked Gibbs sampler

$\hookrightarrow$ For $l = 1, \ldots, L-1$, update the inputs of the stick-breaking weights from

$$v_l \mid \text{else} \sim \text{Beta}\left(n_k + 1, \alpha + \sum_{m=l+1}^{L} n_m\right). \tag{13}$$

$\hookrightarrow$ Letting $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$, the resulting full conditional for α is

$$\alpha \mid \text{else} \sim \text{Gamma}\left(a_\alpha + N - 1, b_\alpha - \sum_{k=1}^{N-1} \log(1 - v_k)\right). \tag{14}$$

Dirichlet process mixtures

Blocked Gibbs sampler

1. Set initial values $\alpha^{(0)}$, $\boldsymbol{\omega}^{(0)}$, and $\boldsymbol{\theta}^{(0)}$.

2. For $s = 1, \ldots, S$:

   ↪ For $i = 1, \ldots, n$, sample the latent component indicator from its full conditional distribution

$$\Pr(z_i^{(s)} = l \mid \text{else}) = \frac{\omega_l^{(s)} \phi(y_i \mid \mu_l^{(s)}, \sigma_l^{2(s)})}{\sum_{l=1}^{L} \omega_l^{(s)} \phi(y_i \mid \mu_l^{(s)}, \sigma_l^{2(s)})}, \quad l = 1, \ldots, L.$$

   ↪ Simulate stick-breaking inputs $\mathbf{v}^{(t)}$ from equation (13).

   ↪ Given $\mathbf{v}^{(s)}$, compute $\boldsymbol{\omega}^{(s)}$ using the stick-breaking construction.

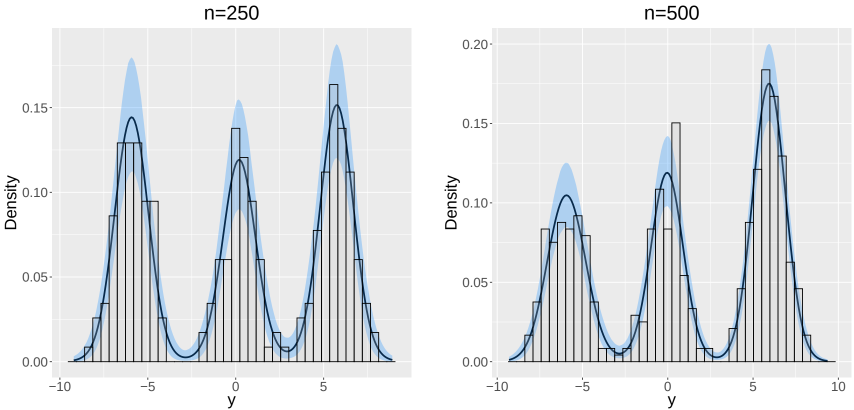
   ↪ Conditional on $\mathbf{z}^{(s)}$, update $\mu_l^{(s)}$ and $(\sigma_l^{(s)})^2$, for $l = 1, \ldots, L$, from (11) and (12), respectively.

   ↪ Update the precision parameter α from (14).

Dirichlet process mixtures

Example

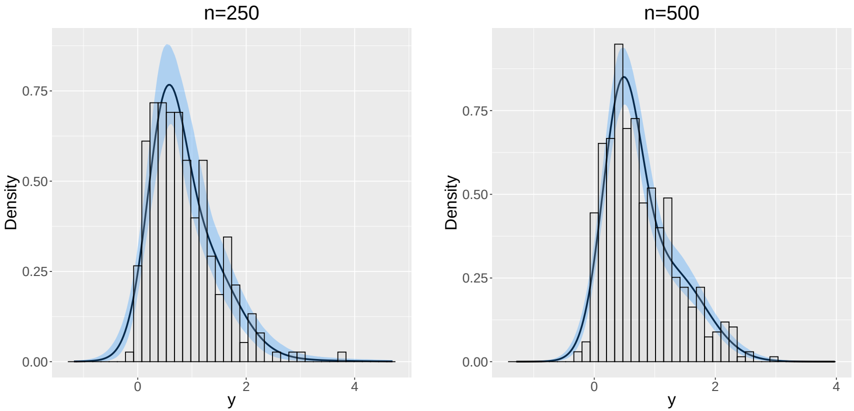
↪ True data generating mechanism: $0.3\phi(y \mid -6, 1) + 0.3\phi(y \mid 0, 1) + 0.4\phi(y \mid 6, 1)$.



Dirichlet process mixtures

Example

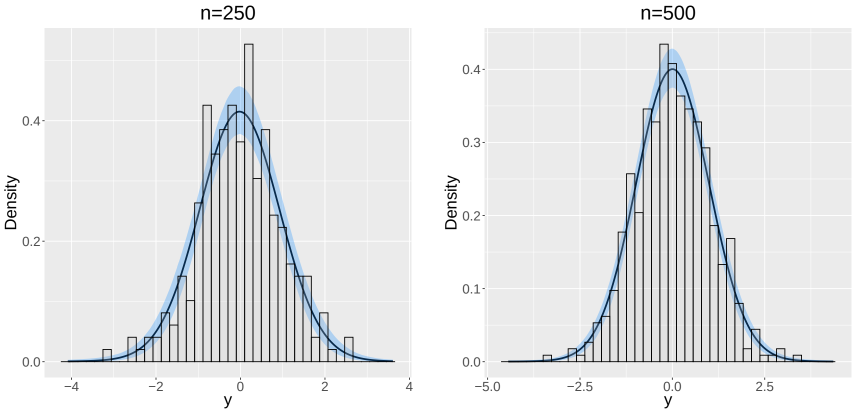
↪ True data generating mechanism: $\phi_{\text{SN}}(y \mid \mu = 0, \sigma^2 = 1, \lambda = 8)$.



Dirichlet process mixtures

Example

↪ True data generating mechanism: $\phi(y \mid 0, 1)$.



Dirichlet process mixtures

Application to evaluating the accuracy of biomarkers

↪ The receiver operating characteristic (ROC) curve is, arguably, the most popular tool used for evaluating the discriminatory ability of continuous-outcome diagnostic tests.

↪ The ROC curve displays the false positive fraction (FPF) against the true positive fraction (TPF) for all possible threshold values that can be used to dichotomise the test result.

↪ The ROC curve thus provides a global description of the trade-off between the FPF and the TPF of the test as the threshold changes.

↪ Let Y be the outcome of the diagnostic test and as D the binary variable indicating the presence ($D = 1$) or absence ($D = 0$) of disease.

↪ By a slight abuse of notation, we use the subscripts D and $\bar{D}$ to denote (random) quantities conditional on, respectively, $D = 1$ and $D = 0$. For example, Y_D and $Y_{\bar{D}}$ denote the test outcomes in the diseased and nondiseased populations, respectively.

Dirichlet process mixtures

Application to evaluating the accuracy of biomarkers

- In the case of a continuous-outcome diagnostic test, the classification is usually made by comparing the test result Y against a threshold c.

- If the outcome is equal or above the threshold, $Y \geq c$, the subject will be diagnosed as diseased. On the other hand, if the test result is below the threshold, $Y < c$, he or she will be classified as nondiseased.

- The ROC curve is then defined as the set of all possible pairs of false positive fractions, $\text{FPF}(c) = \Pr(Y \geq c \mid D = 0) = \Pr(Y_{\bar{D}} \geq c)$, and true positive fractions, $\text{TPF}(c) = \Pr(Y \geq c \mid D = 1) = \Pr(Y_D \geq c)$, that can be obtained by varying the threshold value c, i.e.,
$$\{(\text{FPF}(c), \text{TPF}(c)) : c \in \mathbb{R}\}.$$

- It is common to represent the ROC curve as $\{(p, \text{ROC}(p)) : p \in [0, 1]\}$, where
$$p = \text{FPF}(c) = 1 - F_{\bar{D}}(c), \quad \text{ROC}(p) = 1 - F_D\left\{F_{\bar{D}}^{-1}(1-p)\right\},$$
with $F_{\bar{D}}(y) = \Pr(Y_{\bar{D}} \leq y)$ and $F_D(y) = \Pr(Y_D \leq y)$ denoting the cumulative distribution function (CDF) of Y in the nondiseased and diseased groups, respectively.

Dirichlet process mixtures

Application to evaluating the accuracy of biomarkers

↪ The most widely used is the area under the ROC curve (AUC), defined as

$$\text{AUC} = \int_0^1 \text{ROC}\,(p)\,\text{d}p.$$

↪ The AUC takes values between 0.5, in the case of an uninformative test that classifies individuals no better than chance, and 1.0 for a perfect test.

↪ Estimating the ROC curve/AUC it is a matter of estimating the constituent distribution functions in each group.

↪ We have that, under a DPM of normals in each group, we can write

$$F_D(y) = \sum_{k=1}^{L_D} \omega_{Dk} \Phi(y \mid \mu_{Dk}, \sigma^2_{Dk}),$$

and similarly in the nondiseased group.

Dirichlet process mixtures

Application to evaluating the accuracy of biomarkers

- We will apply our methods to a synthetic dataset mimicking a real dataset about the use of the body mass index (BMI) as a way to detect the presence (1) or absence (0) of two or more cardiovascular disease (CVD) risk factors.

- Following previous studies, the CVD risk factors considered include raised triglycerides, reduced HD-cholesterol, raised blood pressure, and raised fasting plasma glucose.

- We will be using the `R` package `ROCnReg` to conduct the analysis. Please see the results in the associated supplementary material file.