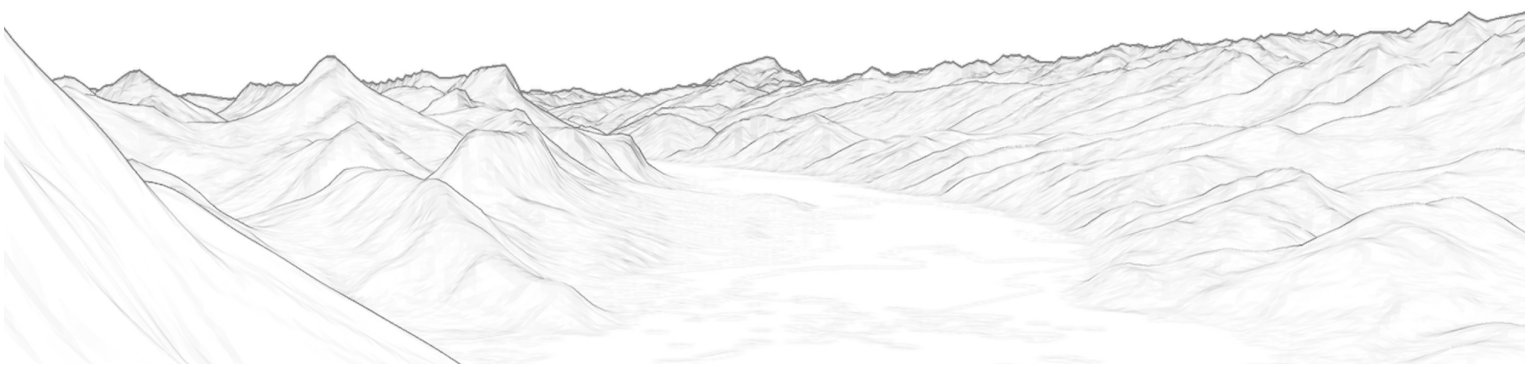
Bayesian nonparametric approaches and clustering

✉ julyan.arbel@inria.fr 🖱 www.julyanarbel.com

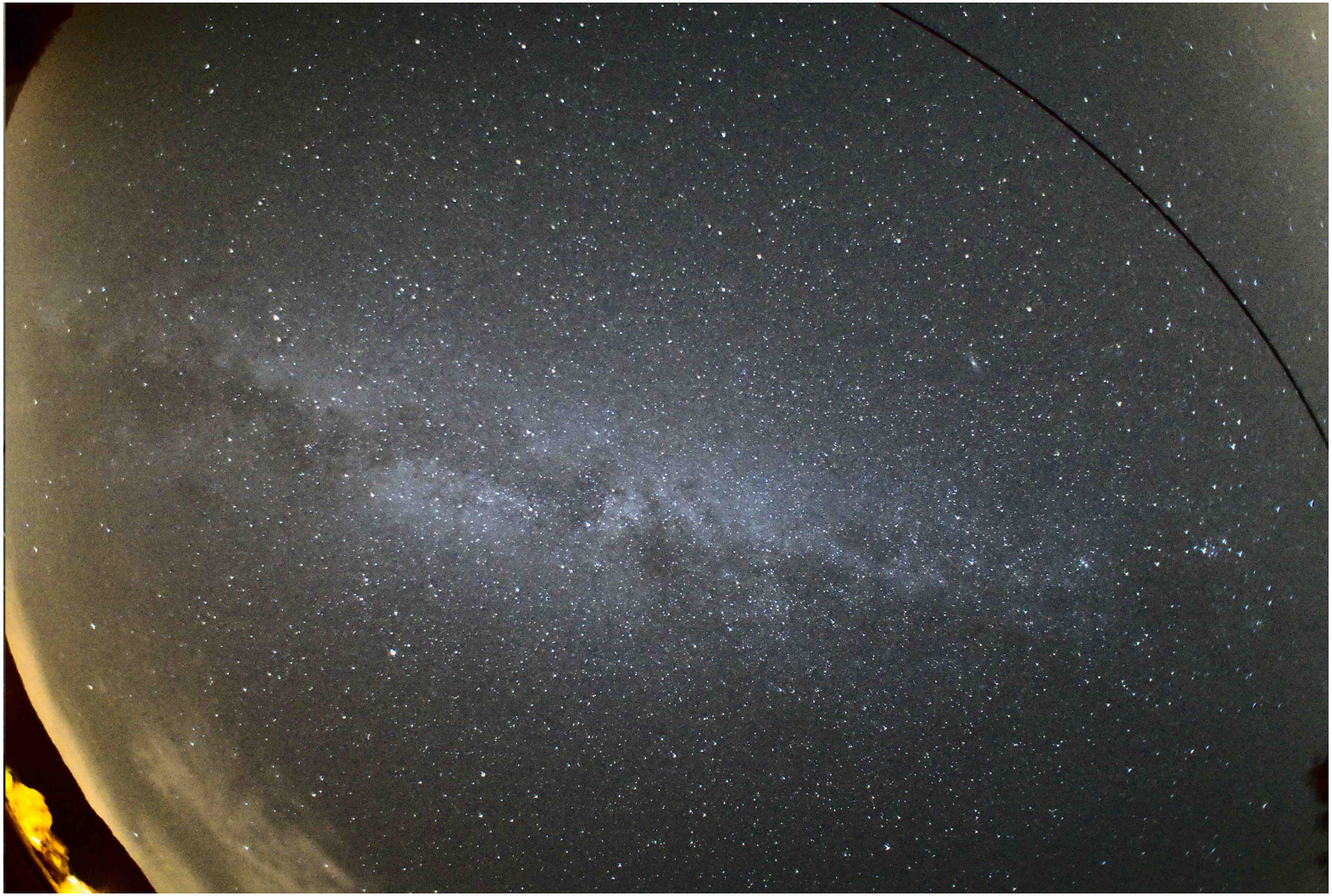
Inria, Grenoble, France

School of Statistics for Astrophysics 2017: Bayesian Methodology
9-13 October 2017, Autrans



What is this talk all about?

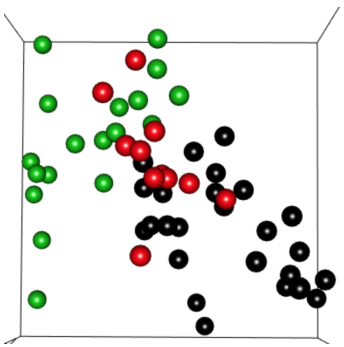
What is this talk all about?



What is this talk all about?

Clustering in a Bayesian nonparametric setting

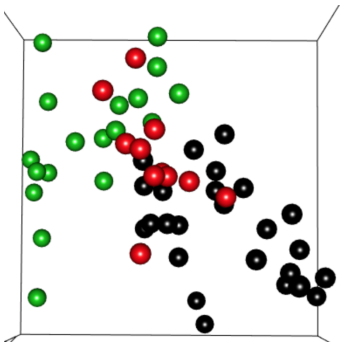
1. study a [toy] astrophysical data set for which clustering is to be performed
2. compare parametric versus nonparametric approaches
3. compare Bayesian versus non Bayesian approaches
4. introduce model-based (unsupervised) clustering
5. run R code all along the way

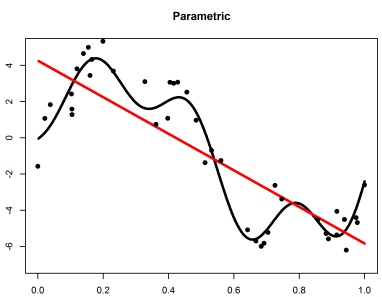


What is this talk all about?

Clustering in a Bayesian nonparametric setting

1. study a [toy] astrophysical data set for which clustering is to be performed
2. compare parametric versus nonparametric approaches
3. compare Bayesian versus non Bayesian approaches
4. introduce model-based (unsupervised) clustering
5. run R code all along the way

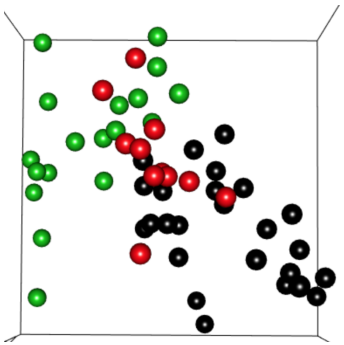


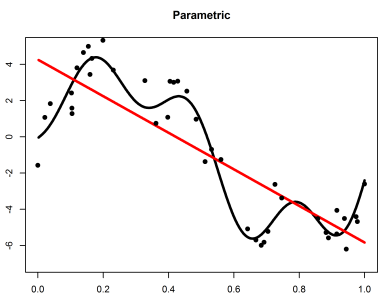


What is this talk all about?

Clustering in a Bayesian nonparametric setting

1. study a [toy] astrophysical data set for which clustering is to be performed
2. compare parametric versus nonparametric approaches
3. compare Bayesian versus non Bayesian approaches
4. introduce model-based (unsupervised) clustering
5. run R code all along the way

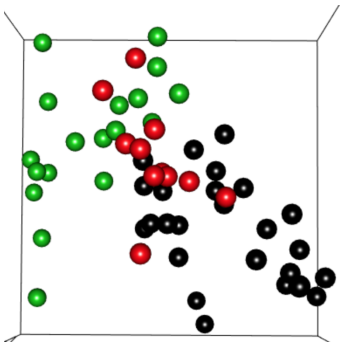


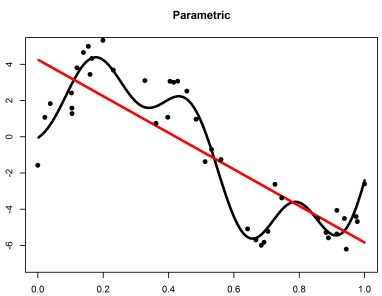


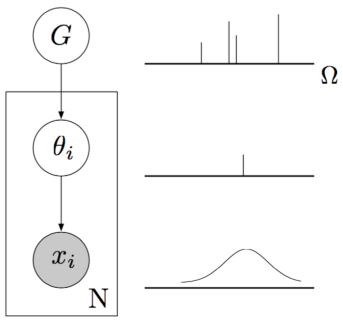
What is this talk all about?

Clustering in a Bayesian nonparametric setting

1. study a [toy] astrophysical data set for which clustering is to be performed
2. compare parametric versus nonparametric approaches
3. compare Bayesian versus non Bayesian approaches
4. introduce model-based (unsupervised) clustering
5. run R code all along the way



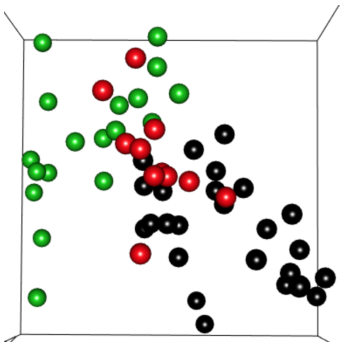


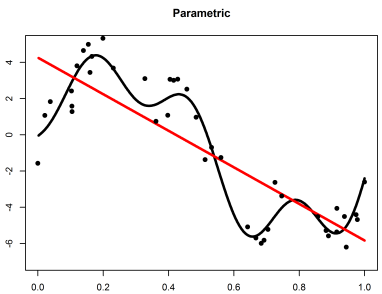


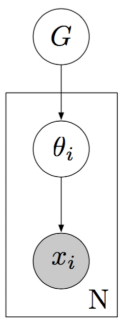
What is this talk all about?

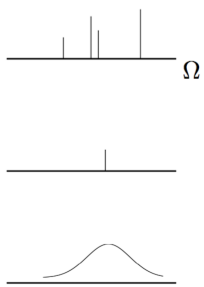
Clustering in a Bayesian nonparametric setting

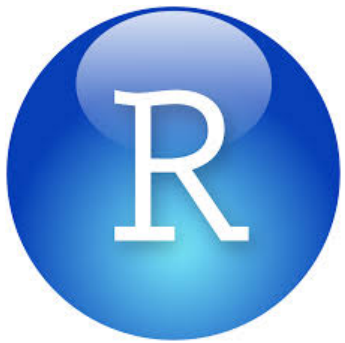
1. study a [toy] astrophysical data set for which clustering is to be performed
2. compare parametric versus nonparametric approaches
3. compare Bayesian versus non Bayesian approaches
4. introduce model-based (unsupervised) clustering
5. run R code all along the way











Acknowledgement and useful links

These slides are inspired by the following introductions to Bayesian nonparametric approaches that I found myself very useful:

- Botond Szabo's tutorial introduction
- Kurt Miller's tutorial introduction
- Peter Orbanz' tutorials webpage, as well as his lecture notes
- Yee Why Teh's tutorial at MLSS 2011
- Mike Jordan's tutorial at NIPS 2005

I also have some handwritten lecture notes available on the website of my Bayesian nonparametric course

Table of Contents

Let's move to R

Table of Contents

Parametric versus nonparametric

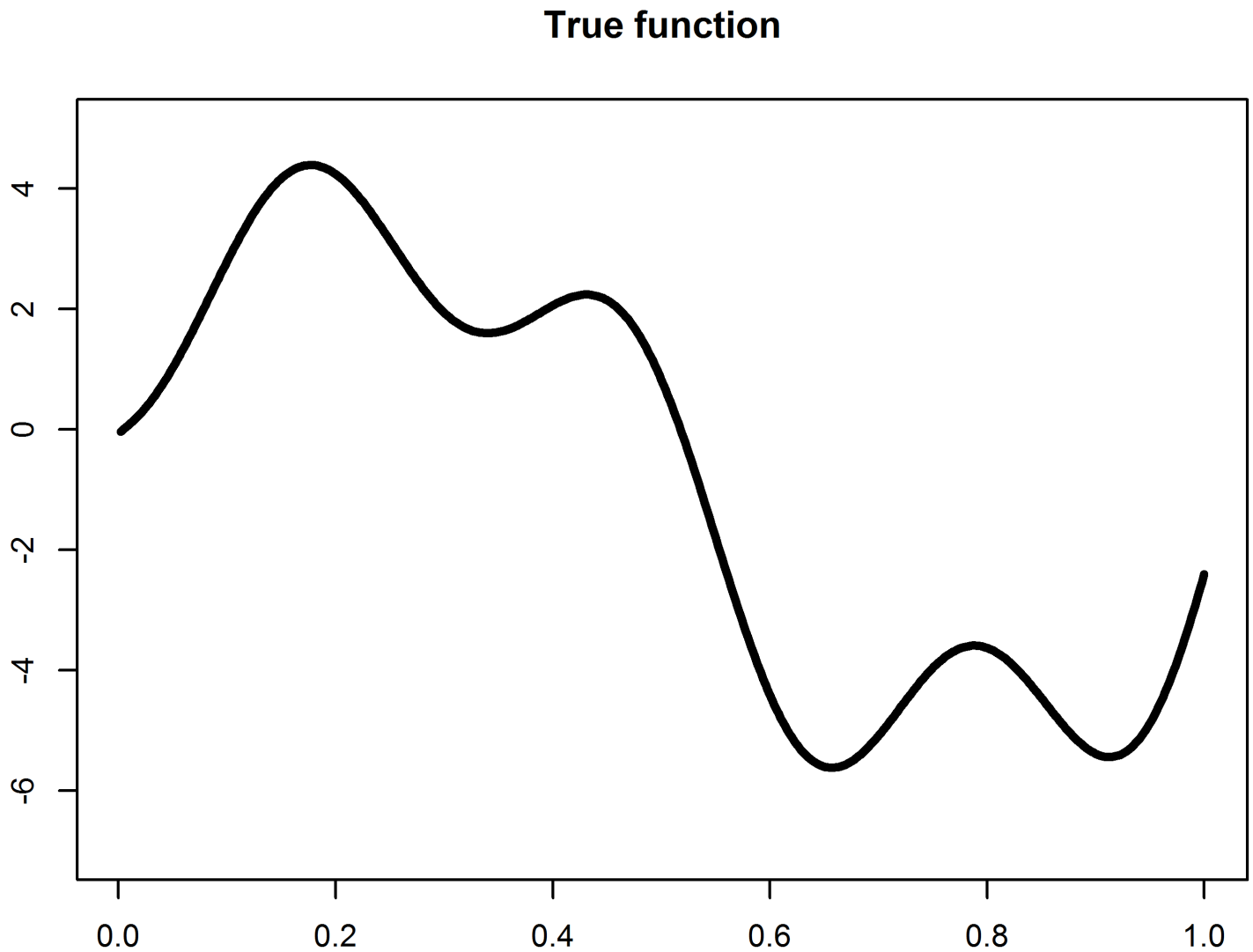
Parametric models

- Finite and fixed number of parameters
- Number of parameters is independent of the dataset

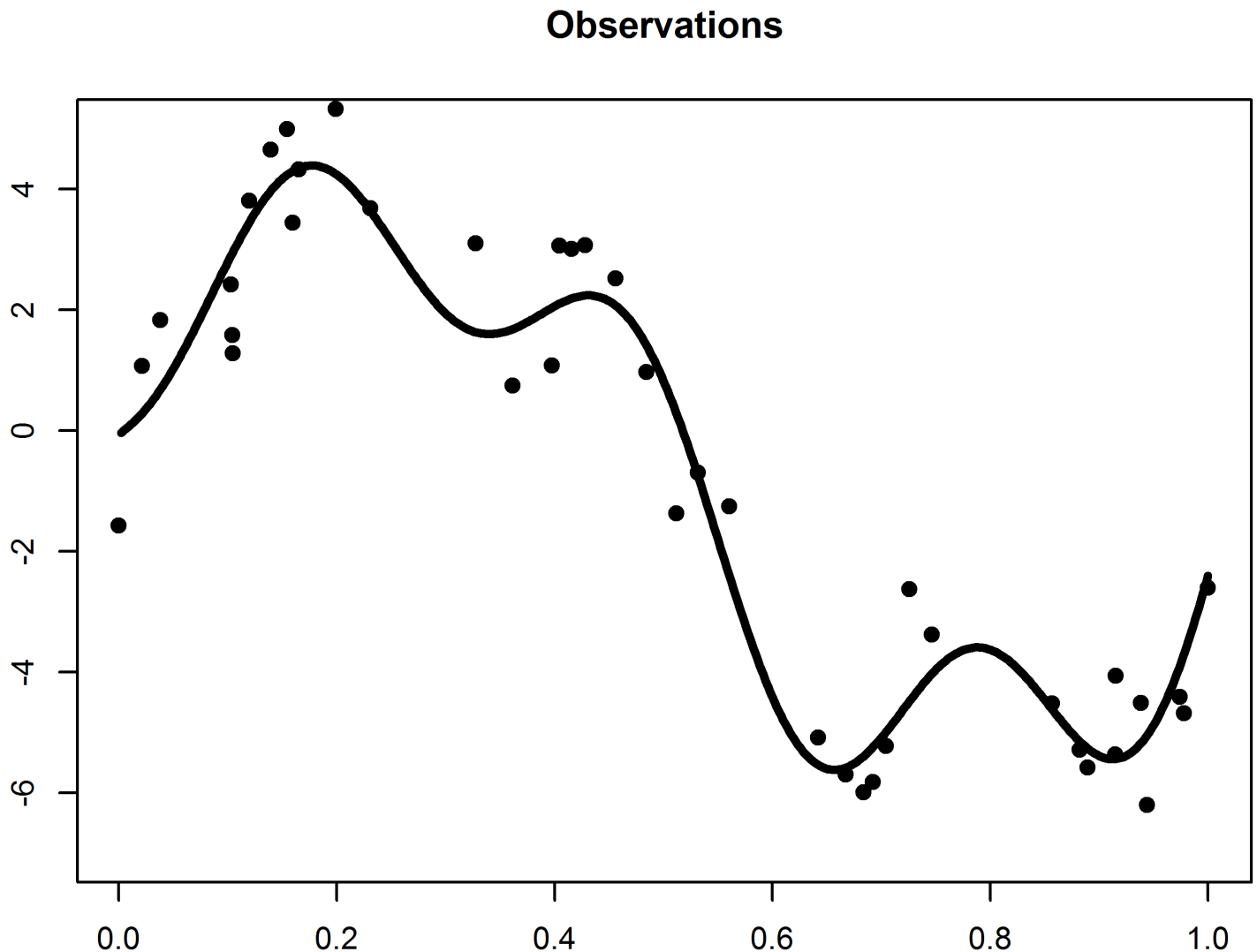
Nonparametric models

- Do have parameters
- Can be understood as having an infinite number of parameters
- Can be understood as having a random number of parameters
- Number of parameters can grow with the dataset

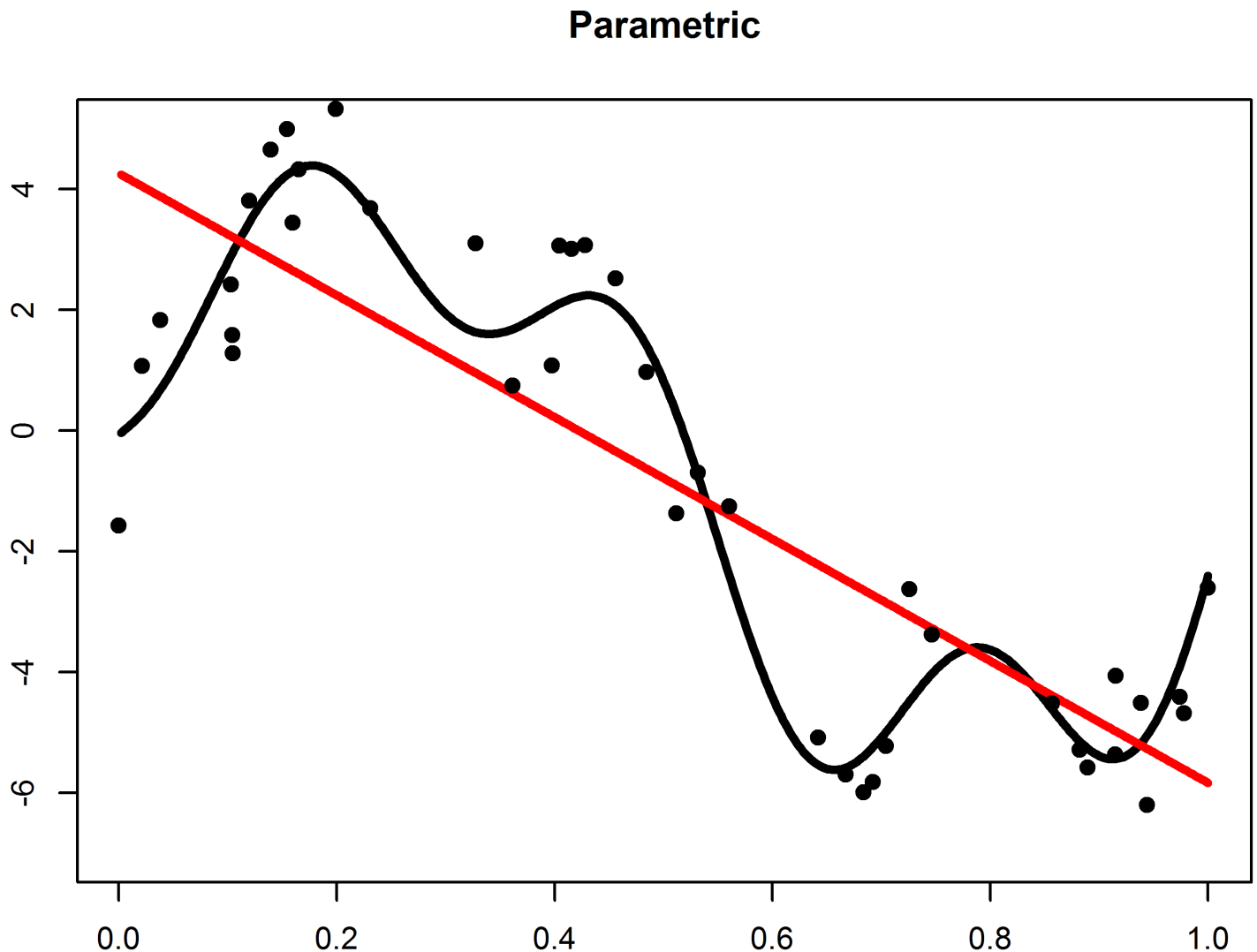
Underlying function



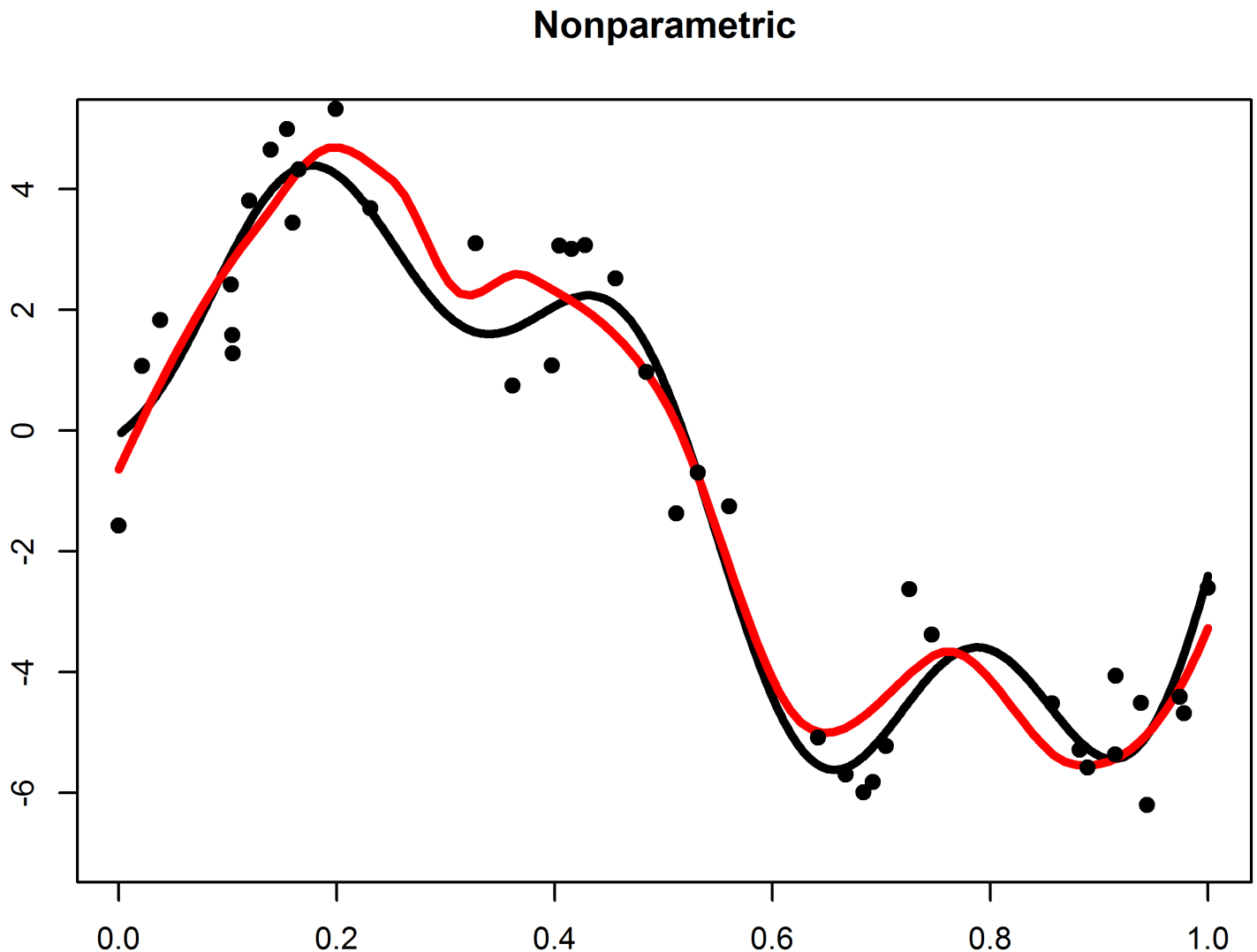
Data



Parametric fitting



Non-parametric fitting

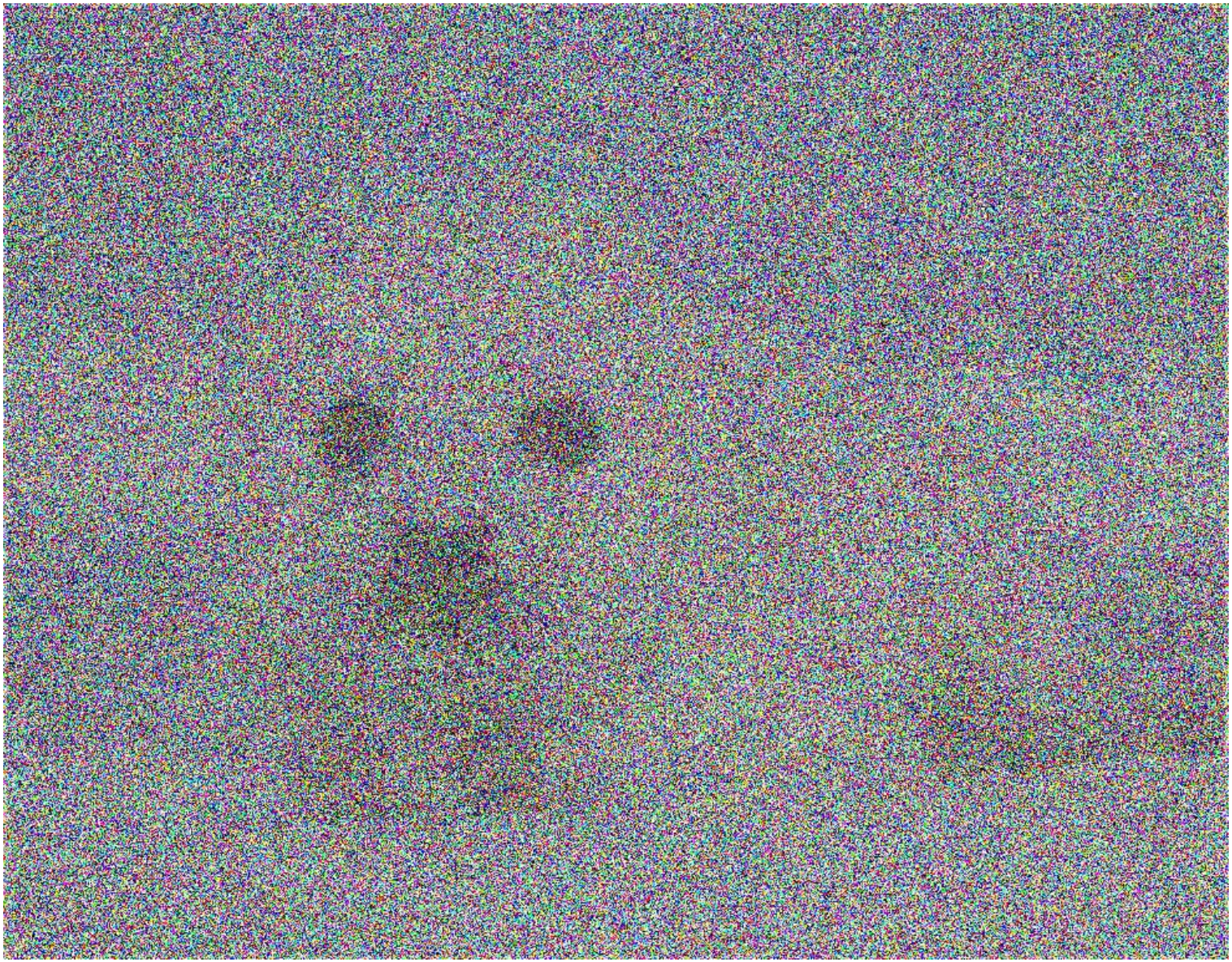


Parametric vs. Nonparametric models

Complexity of the model $\{P_\theta : \theta \in \Theta\}$:

Models:	**Parametric**	**Nonparametric**
Dimension:	Finite dimensional Θ.	Infinite dimensional Θ.
Advantages:	Easier to handle and make interpretations of the results.	Less chance for misspecifications.
	Computationally faster.	More flexible.
Disadvantages:	Without strong belief in the particular structure of the model not reliable.	Computationally and analytically challenging.
Examples:	Poisson (number of car crashes, typos in a book).	Density, regression function estimation.
	Normal distribution (grades of students, height, weight, foot-size of people).	Clustering (unknown cluster size and number).

Noisy picture



Parametric

Nonparametric

Table of Contents

A Model-based Approach to Clustering

We must specify one thing:

- The likelihood term (how data is affected by the parameters):

$$p(X|\phi)$$

Typically Gaussian distribution for $p(X|\theta)$, but not necessarily ; can be skewed, heavy tailed, etc.

A parametric non Bayesian approach

Gaussian Mixture Models with K mixtures

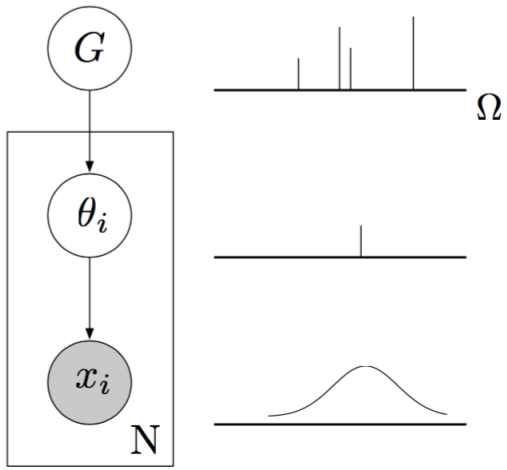
$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

δ_{ϕ_k} is a point mass at ϕ_k.
G is to be understood as a K faceted-dice. The Gaussian mixture likelihood is:

$$p(X|\pi, \phi) = \sum_{k=1}^{K} \pi_k p(x|\phi_k)$$

Then

$$\begin{aligned} \theta_i &\sim G \\ x_i &\sim p(x|\theta_i) \end{aligned}$$



A Bayesian Approach to Clustering

We must specify two things:

- The likelihood term (how data is affected by the parameters):

$$p(X|\phi)$$

  Typically Gaussian distribution for $p(X|\theta)$, but not necessarily ; can be skewed, heavy tailed, etc.

- The prior (the prior distribution on the parameters):

$$p(\phi)$$

We will start by introducing a parametric prior $p(\theta)$, then move to several nonparametric extensions.

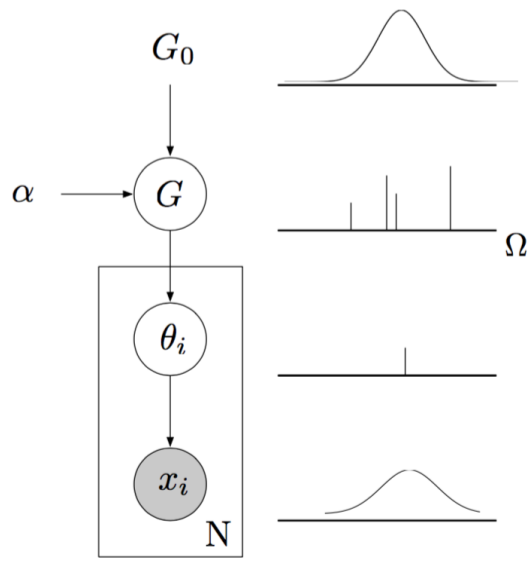
A Bayesian parametric approach

Bayesian Gaussian Mixture Models with K mixtures
We need a distribution over the measure (aka dice) G, that is a distribution over weights or classes $\pi = (\pi_1, \ldots, \pi_K)$ and over mean and covariance (for 2-dimensional data) $\phi_k = (\mu_k, \Sigma_k)$

- $\pi \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$
- $(\mu_k, \Sigma_k) \sim \text{Normal} \times \text{Inverse-Wishart}$

This makes $G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$ a random dice!

$$
\begin{aligned}
\phi_k &\sim G_0 \\
\pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
G &= \sum_{i=1}^{K} \pi_k \delta_{\phi_k} \\
\theta_i &\sim G \\
x_i &\sim p(x|\theta_i)
\end{aligned}
$$



Posterior distribution

In the Bayesian setting, we don't focus only on the maximum likelihood parameters, but in the full posterior instead:

$$p(\pi, \phi|X) \propto p(X|\pi, \phi)p(\pi, \phi)$$

It is not analytically tractable in general, and one resort to approximate inference:

- Markov Chain Monte Carlo (MCMC) methods
- Variational approximations

Choosing K

There are several options for choosing K

- Information criteria for model selection: AIC, BIC, or cross-validation, etc
- Put a prior on K
- Let K get large... possibly infinite.

A Bayesian nonparametric approach

Bayesian Nonparametric Gaussian Mixture Models

We now move to G being an infinite sum $G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$

We need a distribution over this infinite dice G, that is exactly what the Dirichlet process does. It is parameterized by the precision parameter α and the base measure G_0

- $\pi = (\pi_1, \pi_2, \ldots) \sim \text{GEM}(\alpha)$
- $\phi_k \sim G_0$

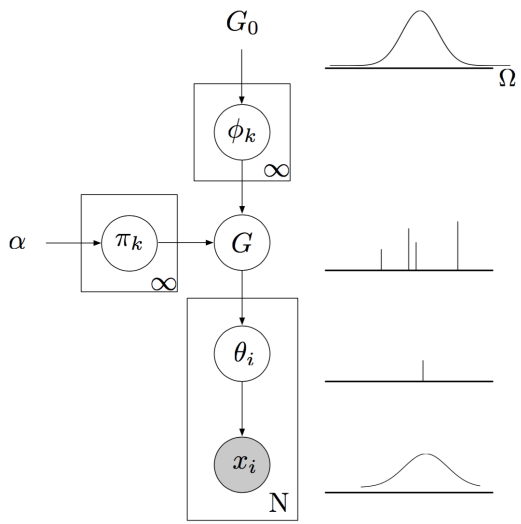


Table of Contents

Dirichlet process

A central Bayesian nonparametric model (Ferguson, 1973)

Definition (Dirichlet process)

A Dirichlet process on the space $\mathcal{Y}$ is a random process P such that there exist α (precision parameter) and G_0 (base/centering distribution) such that for any finite partition $\{A_1, \ldots, A_d\}$ of $\mathcal{Y}$, the random vector $(P(A_1), \ldots, P(A_d))$ is Dirichlet distributed

$$(P(A_1), \ldots, P(A_d)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_d))$$

Notation: $P \sim DP(\alpha, G_0)$

Dirichlet process

A central Bayesian nonparametric model (Ferguson, 1973)

Definition (Dirichlet process)

A Dirichlet process on the space $\mathcal{Y}$ is a random process P such that there exist α (precision parameter) and G_0 (base/centering distribution) such that for any finite partition $\{A_1, \ldots, A_d\}$ of $\mathcal{Y}$, the random vector $(P(A_1), \ldots, P(A_d))$ is Dirichlet distributed

$$(P(A_1), \ldots, P(A_d)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_d))$$

Notation: $P \sim DP(\alpha, G_0)$

Dirichlet process

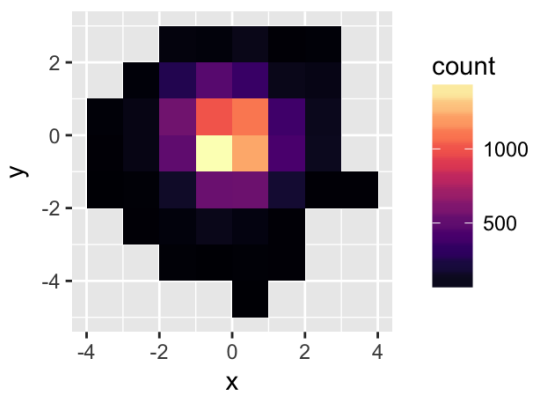
A central Bayesian nonparametric model (Ferguson, 1973)

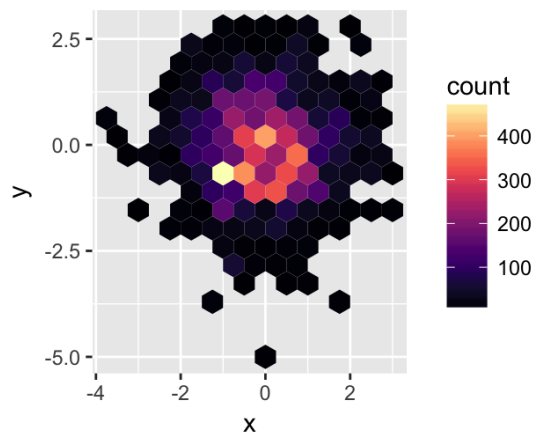
Definition (Dirichlet process)

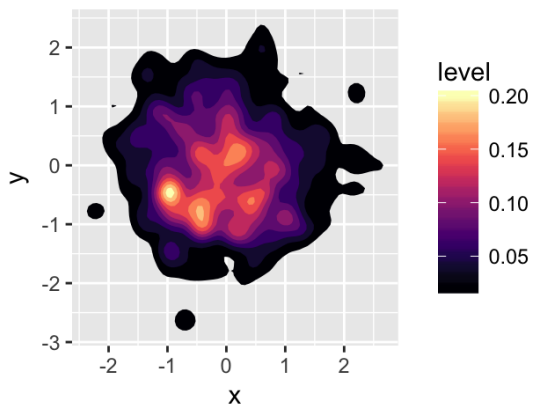
A Dirichlet process on the space $\mathcal{Y}$ is a random process P such that there exist α (precision parameter) and G_0 (base/centering distribution) such that for any finite partition $\{A_1, \ldots, A_d\}$ of $\mathcal{Y}$, the random vector $(P(A_1), \ldots, P(A_d))$ is Dirichlet distributed

$$(P(A_1), \ldots, P(A_d)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_d))$$

Notation: $P \sim DP(\alpha, G_0)$







Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}$$

- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $p_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} Beta(1, \alpha)$,

Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}$$

- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $p_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} Beta(1, \alpha)$,

Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}$$

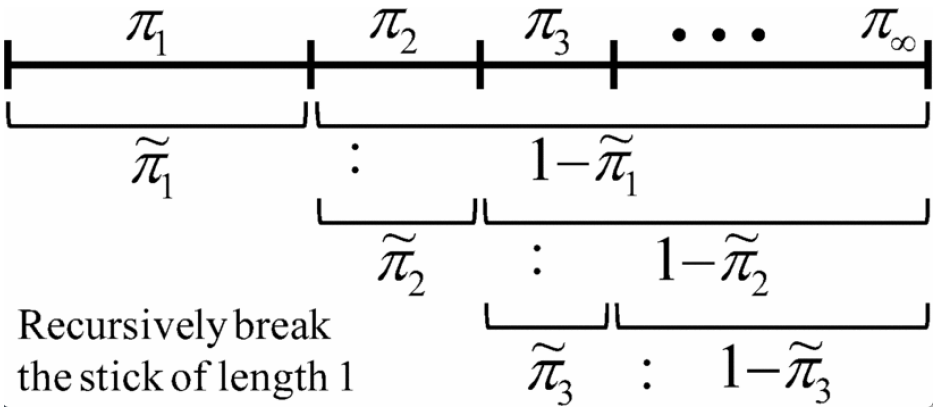
- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $p_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} \mathit{Beta}(1, \alpha)$,

Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}$$

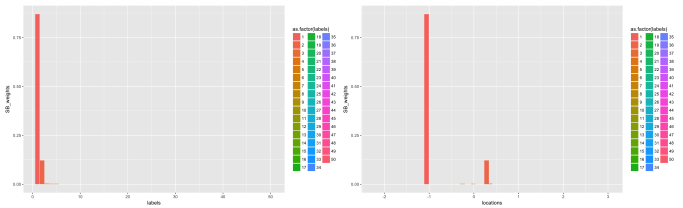
- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $p_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} \mathit{Beta}(1, \alpha)$,



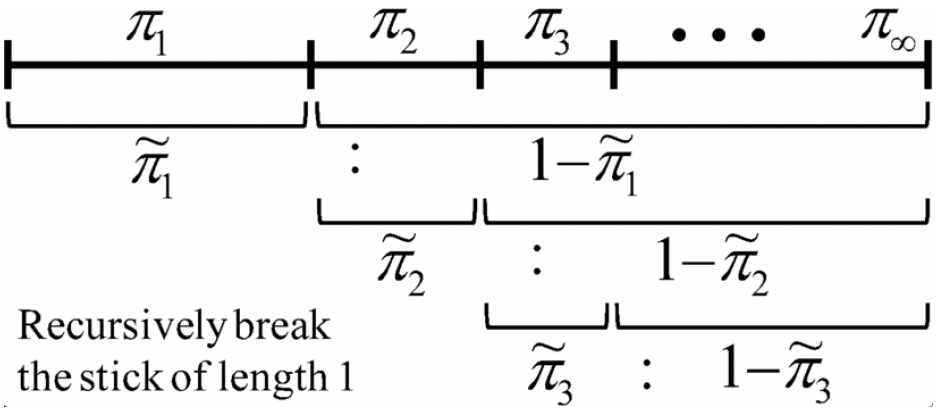
Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}$$



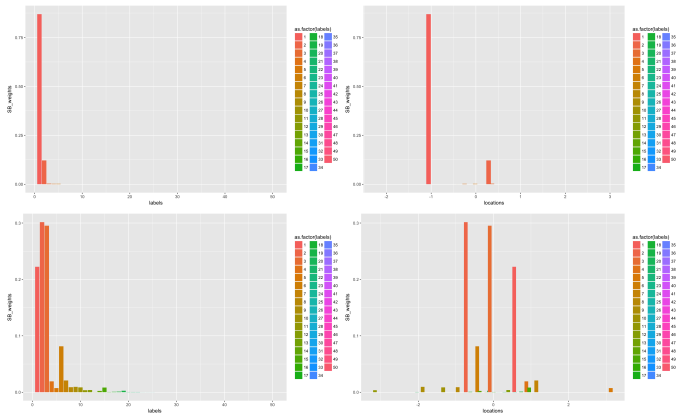
- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $p_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} \mathit{Beta}(1,\alpha)$,



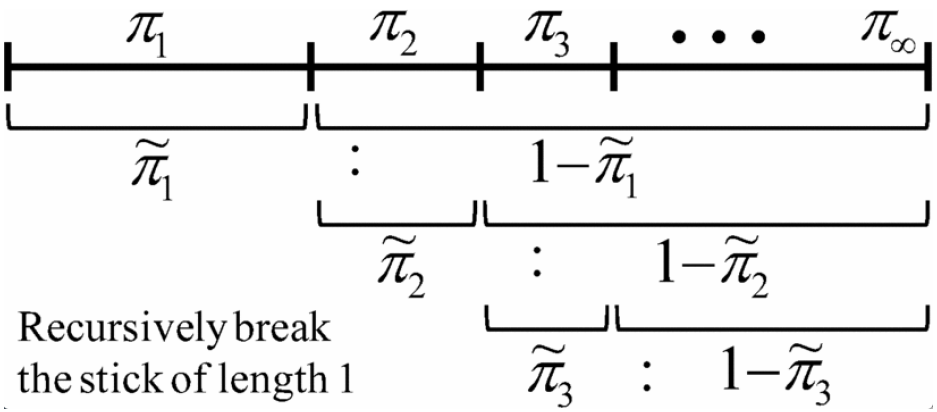
Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}$$



- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $p_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} Beta(1, \alpha)$,

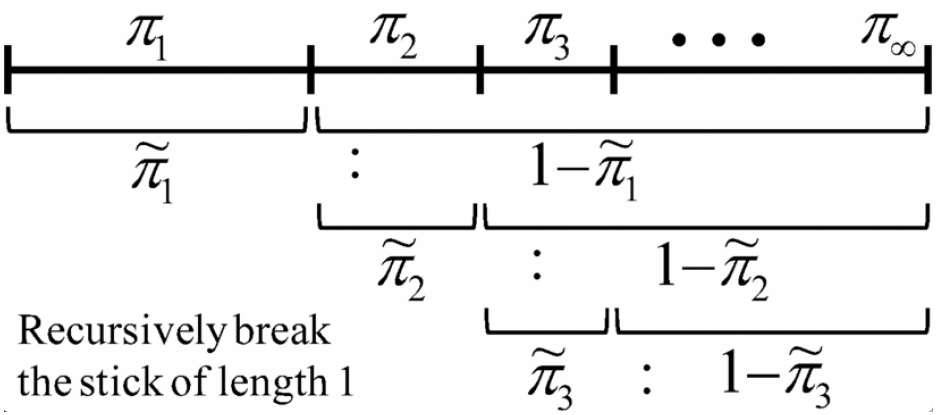


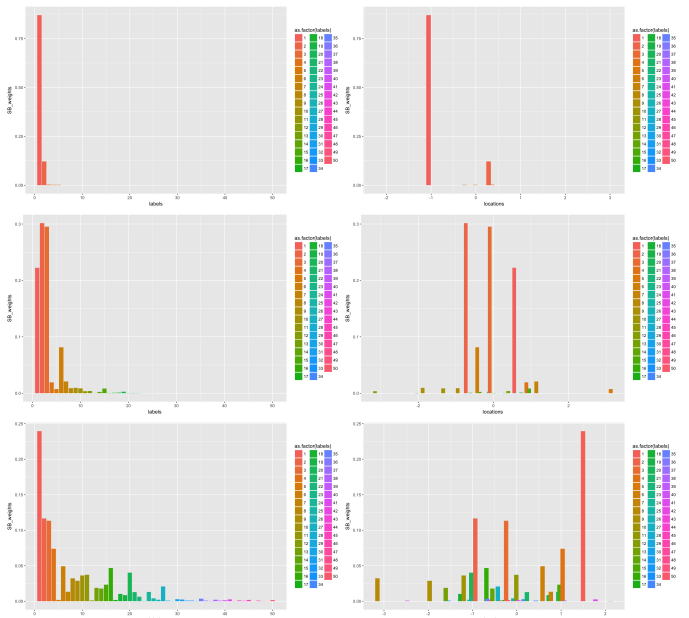
Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}$$

- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $p_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} \mathit{Beta}(1, \alpha)$,



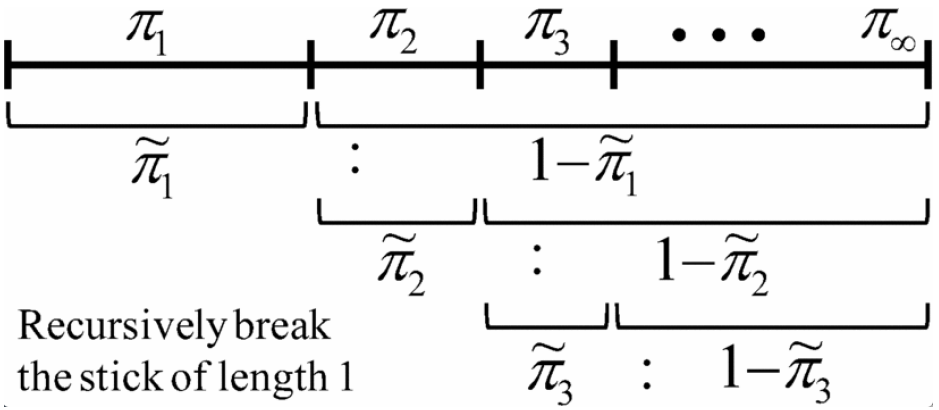


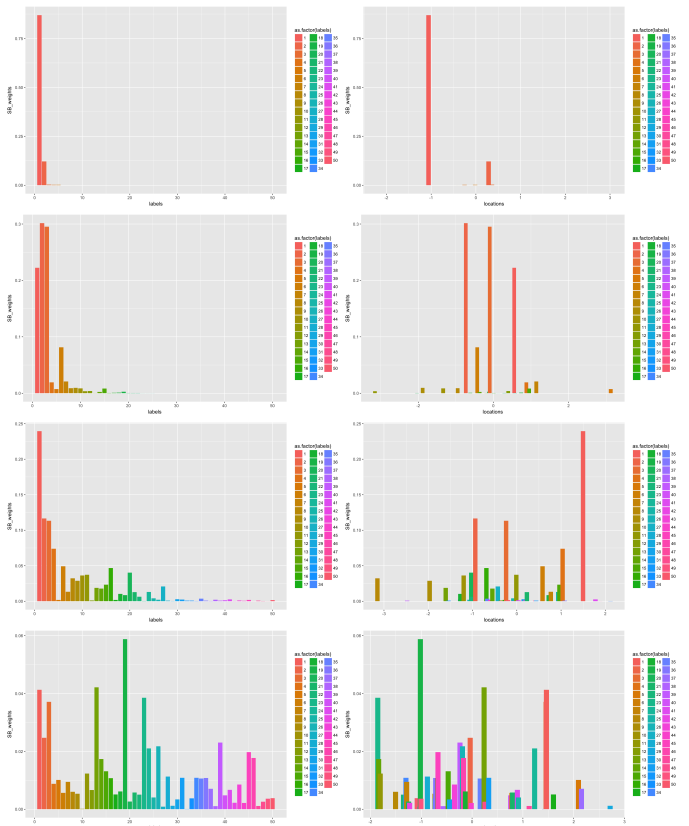
Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}$$

- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $p_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} \mathit{Beta}(1, \alpha)$,





Chinese restaurant process

Consider discrete data $X_1, \dots X_n | P \overset{\text{iid}}{\sim} P$, and $P \sim \mathcal{Q}$
Features $k_n \leq n$ unique values $X_1^*, \dots, X_{k_n}^*$ with resp. frequencies $n_1, \dots, n_{k_n}$

Discrete random probability measures are characterized by predictive distr.

Dirichlet process by Ferguson (1973): $P \sim DP(\alpha, G_0)$

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, X_1, \dots X_n] = \frac{\alpha}{\alpha + n} G_0(.) + \frac{1}{\alpha + n} \sum_{j=1}^{k_n} n_j \delta_{X_j^*}(.)$$

Log rate for number of clusters $k_n \asymp \alpha \log n$

Product form exchangeable partition probability function

$$p(n_1, \dots, n_{k_n}) = \alpha^{k_n} \frac{\Gamma(\alpha)}{\Gamma(\alpha + k_n)} \prod_{j=1}^{k_n} (n_j - 1)!$$

Chinese restaurant process

Consider discrete data $X_1, \dots X_n | P \overset{\text{iid}}{\sim} P$, and $P \sim \mathcal{Q}$
Features $k_n \leq n$ unique values $X_1^*, \dots, X_{k_n}^*$ with resp. frequencies $n_1, \dots, n_{k_n}$

Discrete random probability measures are characterized by predictive distr.

Pitman–Yor process by Pitman & Yor (1997): $P \sim PY(\sigma, \alpha, G_0)$, $\sigma \in (0, 1)$

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, X_1, \dots X_n] = \frac{\alpha + \sigma k_n}{\alpha + n} G_0(.) + \frac{1}{\alpha + n} \sum_{j=1}^{k_n} (n_j - \sigma) \delta_{X_j^*}(.)$$

Power law rate for number of clusters $k_n \asymp S n^{\sigma}$

Product form exchangeable partition probability function

$$p(n_1, \dots, n_{k_n}) = \frac{\prod_{i=1}^{k_n - 1} (\alpha + i\sigma)}{(\alpha + 1)_{(n-1)}} \prod_{j=1}^{k_n} (1 - \sigma)_{(n_j - 1)}$$

Chinese restaurant process

Consider discrete data $X_1, \dots X_n | P \overset{\text{iid}}{\sim} P$, and $P \sim \mathcal{Q}$
Features $k_n \leq n$ unique values $X_1^*, \dots, X_{k_n}^*$ with resp. frequencies $n_1, \dots, n_{k_n}$

Discrete random probability measures are characterized by predictive distr.

Gibbs-type processes by Pitman (2003): $P \sim Gibbs(\sigma, (V_{n,k})_{n,k}, G_0)$, $\sigma < 1$

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, X_1, \dots X_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} G_0(.) + \frac{V_{n+1,k_n}}{V_{n,k_n}} \sum_{j=1}^{k_n} (n_j - \sigma) \delta_{X_j^*}(.)$$

Rate for number of clusters $k_n \asymp$
- K random variable a.s. finite if $\sigma < 0$
- $\alpha \log n$ if $\sigma = 0$
- Sn^σ if $\sigma \in (0,1)$, (S random variable).

Product form exchangeable partition probability function

$$p(n_1, \dots, n_{k_n}) = V_{n,k_n} \prod_{j=1}^{k_n} (1-\sigma)_{(n_j - 1)}$$

Beyond the DP from predictive function viewpoint

A discrete random probability measure P can be classified in 3 main categories according to $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n]$

1) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, \text{model parameters})$
$\iff$ depends on n but not on k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ Dirichlet process (Ferguson, 1973);

2) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, \text{model parameters})$
$\iff$ depends on n and k_n but not on $(n_1, \ldots, n_{k_n})$
$\iff$ Gibbs-type prior (Pitman, 2003);

3) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, (n_1, \ldots, n_{k_n}), \text{model parameters})$
$\iff$ depends on n, k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ tractability issues

Beyond the DP from predictive function viewpoint

A discrete random probability measure P can be classified in 3 main categories according to $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n]$

1) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, \text{model parameters})$
$\iff$ depends on n but not on k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ Dirichlet process (Ferguson, 1973);

2) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, \text{model parameters})$
$\iff$ depends on n and k_n but not on $(n_1, \ldots, n_{k_n})$
$\iff$ Gibbs-type prior (Pitman, 2003);

3) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, (n_1, \ldots, n_{k_n}), \text{model parameters})$
$\iff$ depends on n, k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ tractability issues

Beyond the DP from predictive function viewpoint

A discrete random probability measure P can be classified in 3 main categories according to $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n]$

1) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, \text{model parameters})$
$\iff$ depends on n but not on k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ Dirichlet process (Ferguson, 1973);

2) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, \text{model parameters})$
$\iff$ depends on n and k_n but not on $(n_1, \ldots, n_{k_n})$
$\iff$ Gibbs-type prior (Pitman, 2003);

3) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, (n_1, \ldots, n_{k_n}), \text{model parameters})$
$\iff$ depends on n, k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ tractability issues

Beyond the DP from predictive function viewpoint

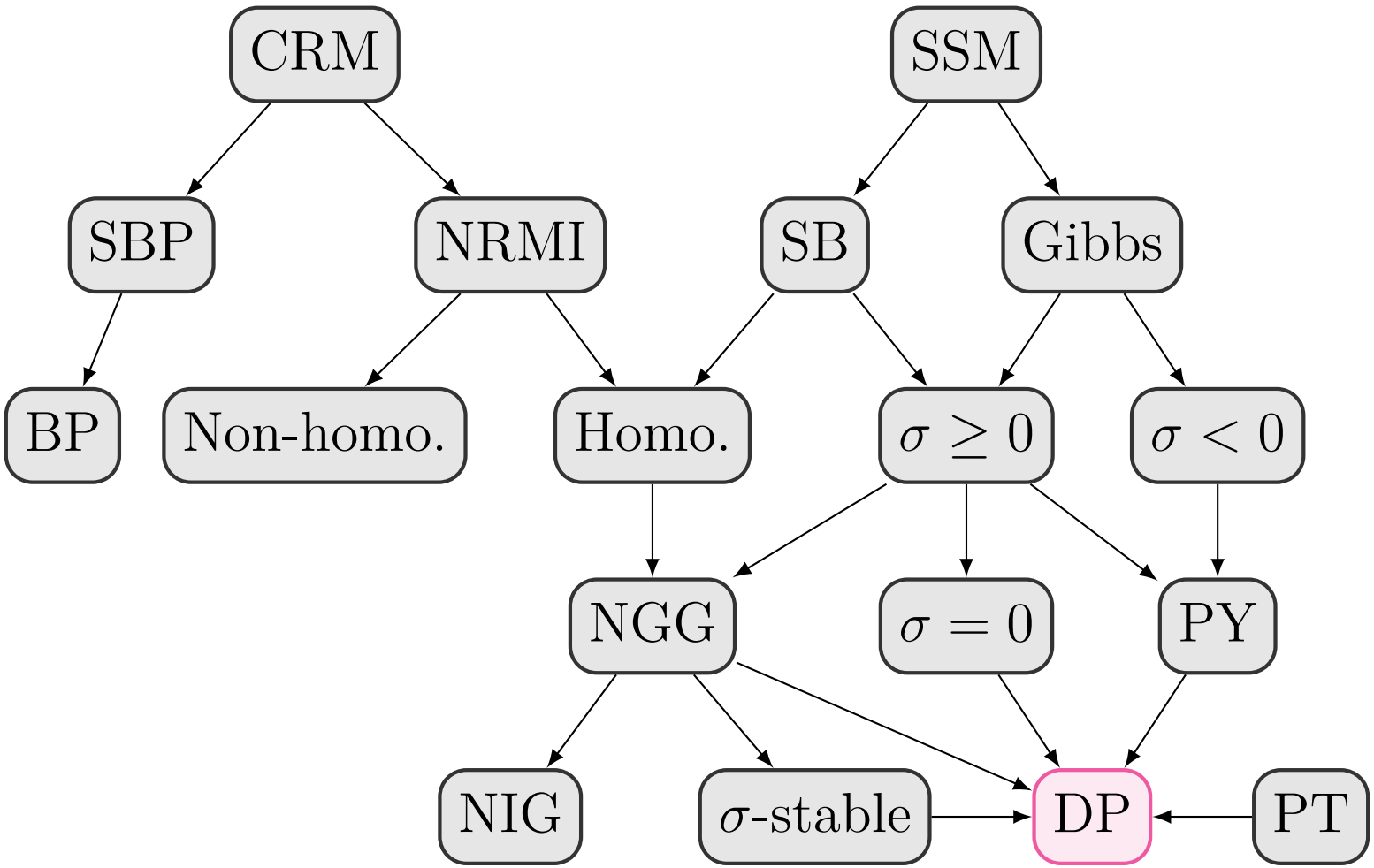
A discrete random probability measure P can be classified in 3 main categories according to $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n]$

1) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, \text{model parameters})$
$\iff$ depends on n but not on k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ Dirichlet process (Ferguson, 1973);

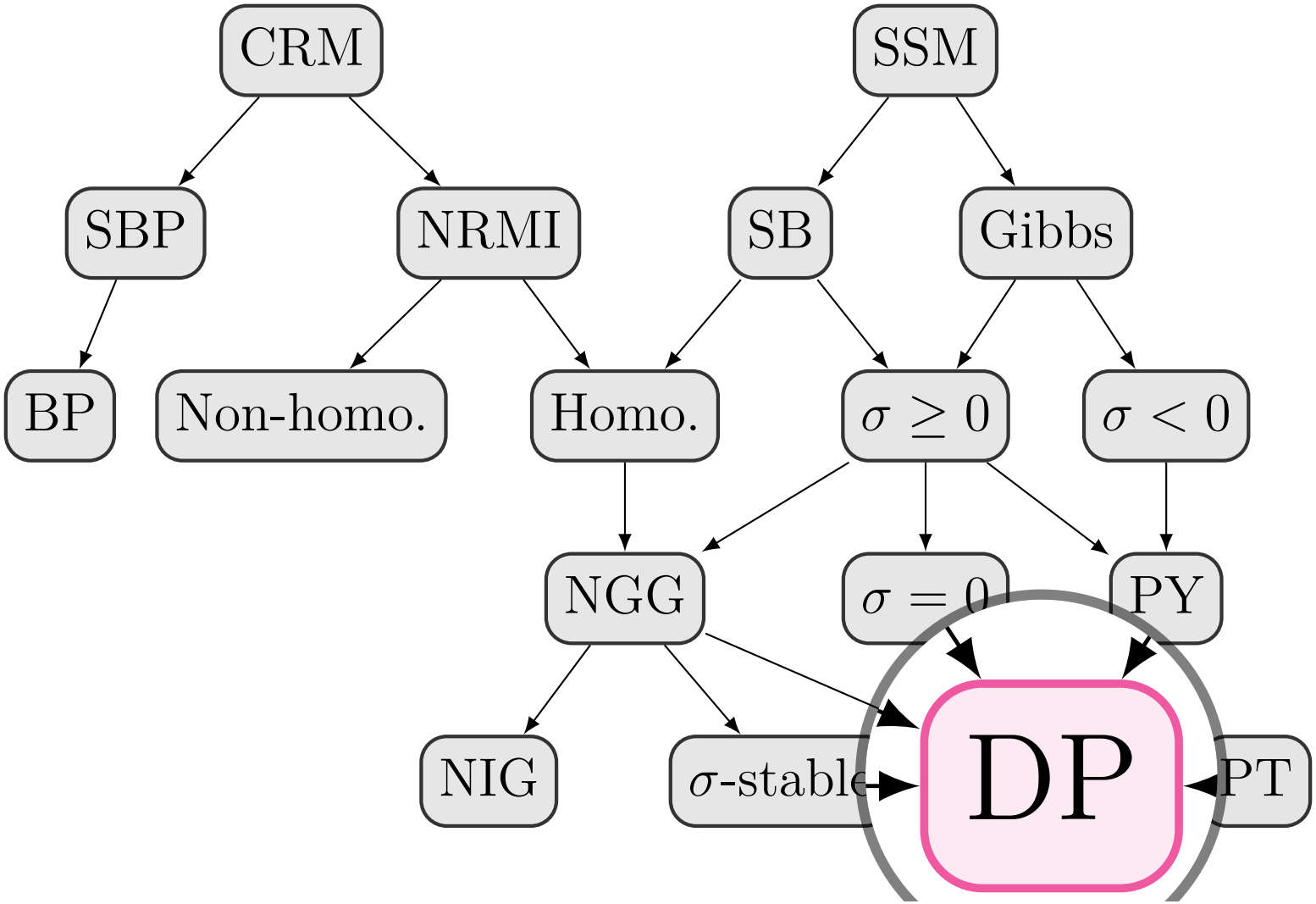
2) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, \text{model parameters})$
$\iff$ depends on n and k_n but not on $(n_1, \ldots, n_{k_n})$
$\iff$ Gibbs-type prior (Pitman, 2003);

3) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, (n_1, \ldots, n_{k_n}), \text{model parameters})$
$\iff$ depends on n, k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ tractability issues

Tree of discrete random probability measures



Tree of discrete random probability measures



Tree of discrete random probability measures

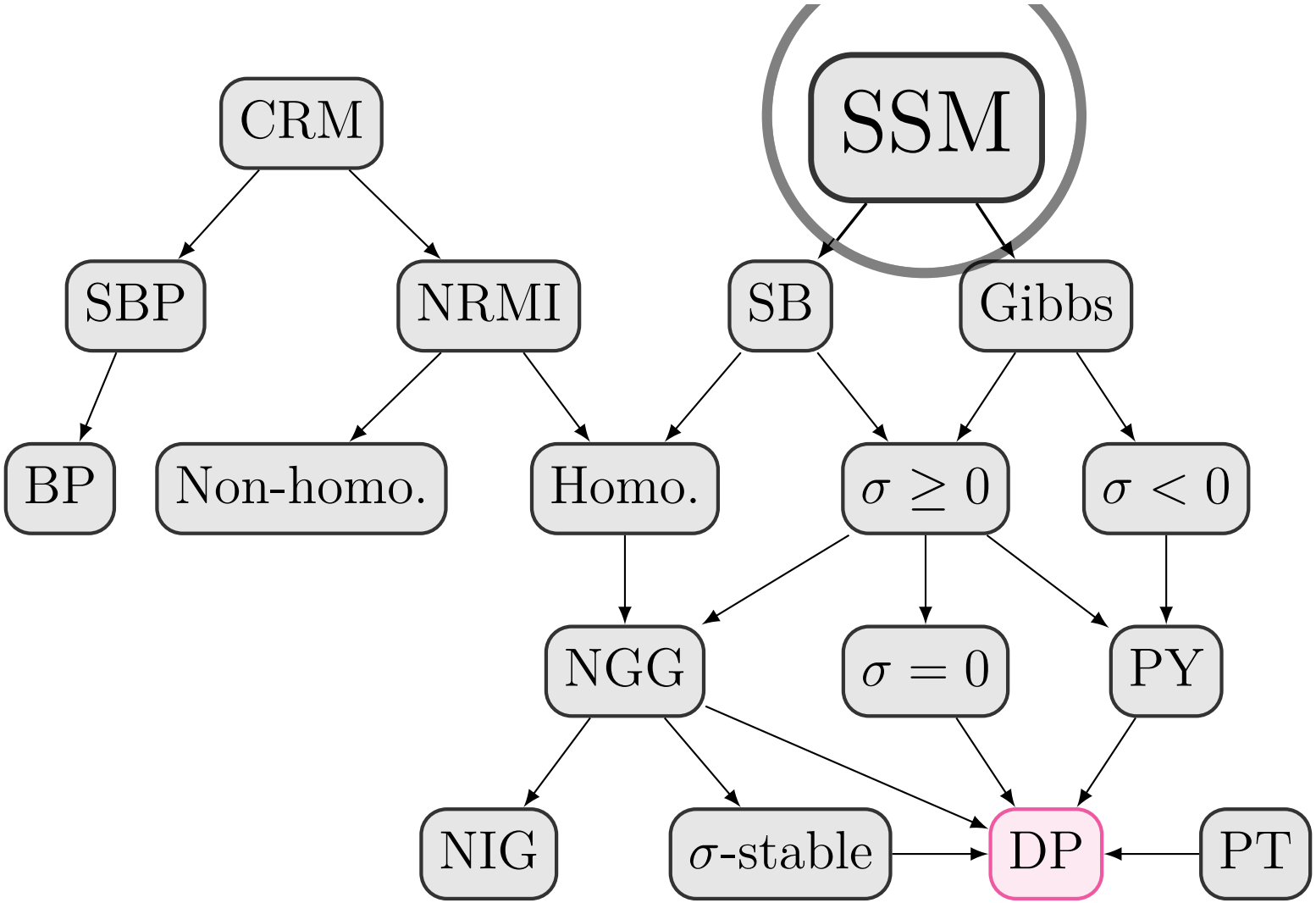


Table of Contents

Sampling from the posterior distribution

This could be an exercise... otherwise, we can you code or packages that do precisely that!

- DPpackage
- BNPdensity
- function MCMCcoolMulti in the src folder. Based on ongoing work: Arbel, J., Corradin, R., Nipoti, B. (2017). Robust Bayesian nonparametric methods for density estimation and clustering on the phase-space. In preparation.

Deriving an optimal cluster

From decision theory: a Bayes estimator minimizes a posterior expected loss.

$$\hat{a}_L = \arg \inf_{a \in A} \mathbb{E}_{\pi(\theta)}[L_a(\theta)]. \tag{1}$$

The posterior expected loss of clustering c', denoted by $L(c')$, is obtained by **averaging the loss with respect to posterior weight**

$$L(c') = \sum_{c \in \mathcal{A}_n} L(c, c') p(c|\mathbf{x}),$$

and the decision is taken by choosing the best

$$\hat{c} = \arg \min_{c' \in \mathcal{A}_n} \sum_{c \in \mathcal{A}_n} L(c, c') p(c|\mathbf{x}) \tag{2}$$

- 0-1 loss: leads to mode a posteriori (MAP). Negative results show that the MAP is inconsistent.
- a better loss from information theory, called Variation of information (VI). See Wade and Ghahramani in the ref folder.

Simplest loss: L_{0-1}

I start with the 0–1 loss just for mathematical tractability. However, I do not necessarily expect that it works with it. We have

$$L_{0-1}(c') = \sum_{c \in \mathcal{A}_n} L_{0-1}(c, c') p(c|\mathbf{x}) = \sum_{c \in \mathcal{A}_n,\, c \neq c'} p(c|\mathbf{x}),$$

which is to say that the expected loss of c' is **all the posterior mass except that of c'.** So that it is easily minimized at the value c' which has **maximum** posterior weight:

$$\hat{c} = \arg \min_{c' \in \mathcal{A}_n} L_{0-1}(c') = \arg \max_{c' \in \mathcal{A}_n} p(c'|\mathbf{x}).$$

Variation of information

Meila (2007) introduces the **variation of information** (VI) for cluster comparison, which is constructed from information theory and compares the information in two clusterings with the information shared between the two clusterings. More formally, the VI is defined as

$$\begin{aligned}\mathrm{VI}(c, \widehat{c}) &= \mathrm{H}(c) + \mathrm{H}(\widehat{c}) - 2\mathcal{I}(c, \widehat{c}) \\ &= -\sum_{i=1}^{k_N} \frac{n_{i+}}{N} \log\left(\frac{n_{i+}}{N}\right) - \sum_{j=1}^{\widehat{k}_N} \frac{n_{+j}}{N} \log\left(\frac{n_{+j}}{N}\right) - 2\sum_{i=1}^{k_N}\sum_{j=1}^{\widehat{k}_N} \frac{n_{ij}}{N} \log\left(\frac{n_{ij}N}{n_{i+}n_{+j}}\right),\end{aligned}$$

The first two terms represent the entropy of the two clusterings, which measures the uncertainty in bits of the cluster allocation of a unknown randomly chosen data point given a particular clustering of the data points. The last term is the mutual information between the two clusterings and measures the reduction in the uncertainty of the cluster allocation of a data point in c when we are told its cluster allocation in $\widehat{c}$.