Nonparametric stick breaking priors with simple weights

Ramsés H. Mena

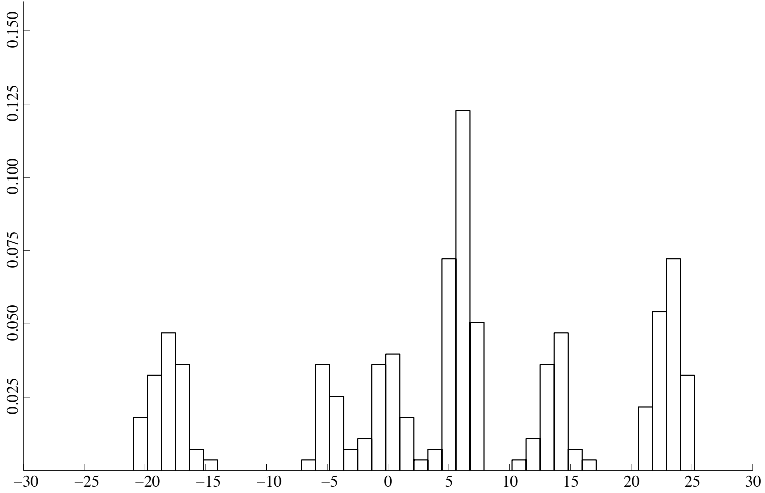
IIMAS-UNAM

(work with Fuentes-García, R., Ruggiero, M. and Walker, S.G.)

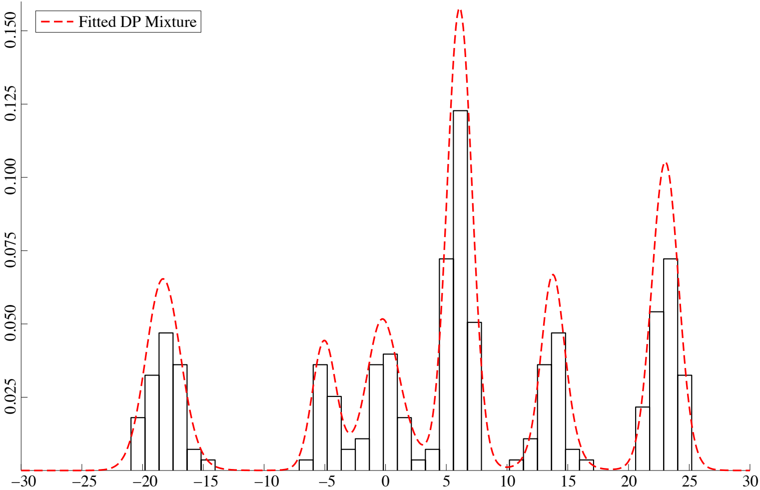
Gatsby Computational Neuroscience Unit

February, 2012

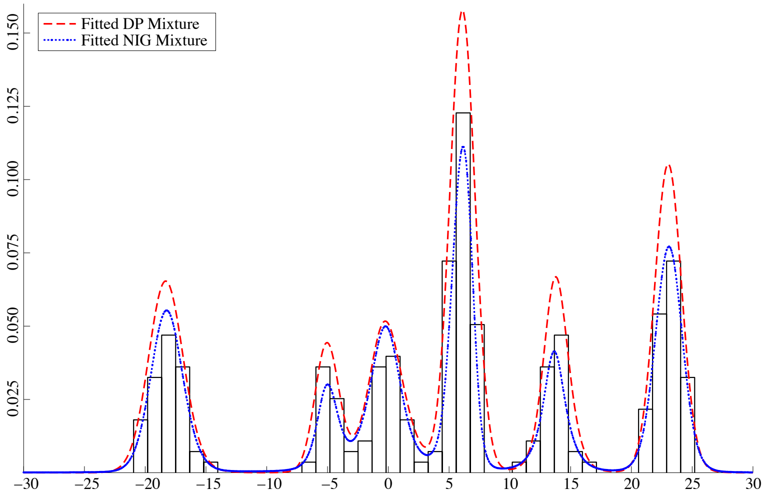
- Suppose we observe the following data



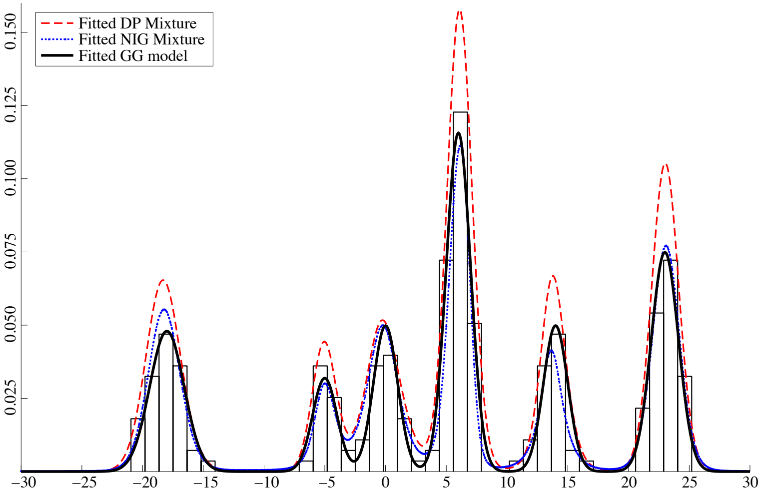
- we could fit of DP mixture $\mathsf{f}(\cdot) = \int_{\mathbb{X}} f(\cdot \mid x)\mu(\mathrm{d}x)$, $\mu \sim \mathcal{D}_{\theta\nu_0}$



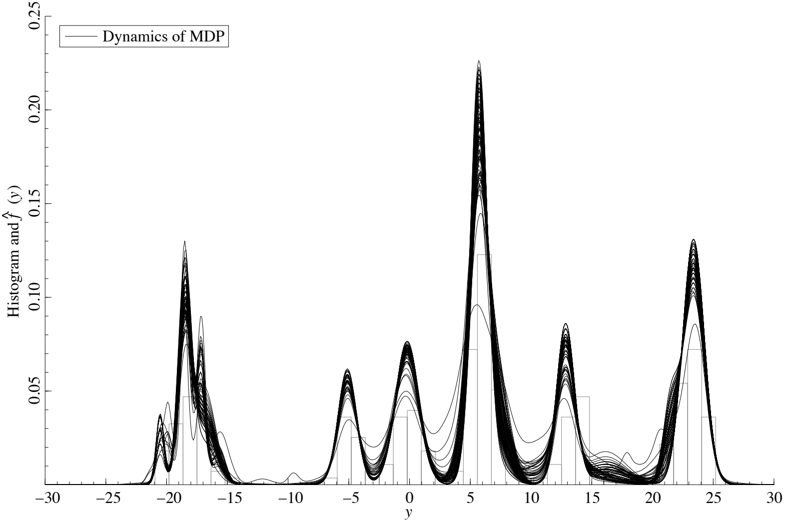
- ...or alternatively a NIG mixture model



- . . . or even a more elaborated GG mixture model



- These estimators are result of a convergent MCMC



→A convergent state of these MCMC estimators typically needs:

- Hyper-parameters specifications in the kernel $f(\cdot \mid x)$ and ν_0
- Randomization of the parameters of RPMs μ
- Techniques to accelerate and attain convergence

→ "General" RPMs partially ease some of these aspects, however there is a tractability issue:

The more general the rpm the less manageable it becomes

Here we present a simplistic approach that addresses some of these issues and explore its applications in depending settings

1. Motivation
2. Geometric weights
3. Dependent processes
4. Estimation

1. Motivation
2. Geometric weights
3. Dependent processes
4. Estimation

1. Motivation
2. Geometric weights
3. Dependent processes
4. Estimation

Stick breaking weights

- Any discrete dist. can be represented as

$$P(B) = \sum_{i=1}^{\infty} \mathsf{w}_i \, \delta_{z_i}(B), \quad B \in \mathcal{X}, \qquad \sum_i \mathsf{w}_i = 1$$

- Make the "weights", $(\mathsf{w}_i)_{i\geq 1}$, and "locations", $(z_i)_{i\geq 1}$ random

  $\Rightarrow \mu$ is a Random Prob. Measure (RPM)

- Stick-breaking weights

$$\mathsf{w}_1 = \mathsf{V}_1, \qquad \mathsf{w}_i = \mathsf{V}_i \prod_{j<i} (1 - \mathsf{V}_j), \quad i \geq 2$$

- Let $(\mathsf{V}_i)_{i\geq 1}$ indep. $[0,1]$-valued r.v.'s with $\mathsf{E}[\sum_{i\geq 1} \log(1 - \mathsf{V}_i)] = -\infty$

Stick breaking weights

- Any discrete dist. on a Polish space $(\mathbb{X}, \mathcal{X})$ can be represented as

$$\mu(B) = \sum_{i=1}^{\infty} \mathsf{w}_i \, \delta_{z_i}(B), \quad B \in \mathcal{X}, \qquad \sum_i \mathsf{w}_i = 1 \text{ a.s.}$$

- Make the "weights", $(\mathsf{w}_i)_{i\geq 1}$, and "locations", $(z_i)_{i\geq 1}$ random
  $\Rightarrow \mu$ is a Random Prob. Measure (RPM)

- Stick-breaking weights

$$\mathsf{w}_1 = \mathsf{V}_1, \qquad \mathsf{w}_i = \mathsf{V}_i \prod_{j<i} (1 - \mathsf{V}_j), \quad i \geq 2$$

- Let $(\mathsf{V}_i)_{i\geq 1}$ indep. $[0,1]$-valued r.v.'s with $\mathsf{E}[\sum_{i\geq 1} \log(1 - \mathsf{V}_i)] = -\infty$

Stick breaking weights

- Any discrete dist. on a Polish space $(\mathbb{X}, \mathcal{X})$ can be represented as

$$\mu(B) = \sum_{i=1}^{\infty} \mathsf{w}_i \, \delta_{z_i}(B), \quad B \in \mathcal{X}, \qquad \sum_i \mathsf{w}_i = 1 \text{ a.s.}$$

- Make the "weights", $(\mathsf{w}_i)_{i\geq 1}$, and "locations", $(z_i)_{i\geq 1}$ random
  $\Rightarrow \mu$ is a Random Prob. Measure (RPM)

- Stick-breaking weights

$$\mathsf{w}_1 = \mathsf{V}_1, \qquad \mathsf{w}_i = \mathsf{V}_i \prod_{j<i} (1 - \mathsf{V}_j), \quad i \geq 2$$

- Let $(\mathsf{V}_i)_{i\geq 1}$ indep. $[0,1]$-valued r.v.'s with $\mathsf{E}[\sum_{i\geq 1} \log(1 - \mathsf{V}_i)] = -\infty$

Stick breaking weights

- Any discrete dist. on a Polish space $(\mathbb{X}, \mathcal{X})$ can be represented as

$$\mu(B) = \sum_{i=1}^{\infty} \mathsf{w}_i\, \delta_{z_i}(B), \quad B \in \mathcal{X}, \qquad \sum_i \mathsf{w}_i = 1 \text{ a.s.}$$

- Make the "weights", $(\mathsf{w}_i)_{i\geq 1}$, and "locations", $(z_i)_{i\geq 1}$ random
  $\Rightarrow \mu$ is a Random Prob. Measure (RPM)

- Stick-breaking weights

$$\mathsf{w}_1 = \mathsf{V}_1, \qquad \mathsf{w}_i = \mathsf{V}_i \prod_{j<i}(1 - \mathsf{V}_j), \quad i \geq 2$$

- Let $(\mathsf{V}_i)_{i\geq 1}$ indep. $[0,1]$-valued r.v.'s with $\mathsf{E}[\sum_{i\geq 1} \log(1 - \mathsf{V}_i)] = -\infty$

Dirichlet process $\mathcal{D}_{\theta,\nu_0}$

- Sethuraman (1994)

$$\text{if } \mathsf{V}_i \overset{\text{iid}}{\sim} \mathsf{Be}(1,\theta) \qquad \text{and} \qquad z_i \overset{\text{iid}}{\sim} \nu_0 \quad (\text{indep. of } \mathsf{V}_i\text{'s})$$

- μ follows Ferguson (1973) Dirichlet process ($\mu \sim \mathcal{D}_{\theta,\nu_0}$)

  i.e. a stochastic processes, $\{\mu(B)\}_{B\in\mathcal{X}}$, with finite dim. dist.

$$(\mu(B_1),\ldots,\mu(B_k)) \sim \mathsf{Dirichlet}(\theta\nu_0(B_1),\ldots,\theta\nu_0(B_k))$$

  for all $k \geq 1$ and all partitions $(B_1,\ldots,B_k)$ of $\mathbb{X}$.

Some basic properties of $\mathcal{D}_\alpha$

- $\mathsf{E}[\mu(B)] = \nu_0(B)$, $\mathsf{Var}[\mu(B)] = \frac{\nu_0(B)(1-\nu_0(B))}{\theta+1}$

  $\mathsf{Cov}(\mu(B_2), \mu(B_2)) = \frac{\nu_0(B_1 \cap B_2) - \nu_0(B_1)\nu_0(B_2)}{\theta+1}$

If $X_i \mid \mu \overset{\text{iid}}{\sim} \mu$ and $\mu \sim \mathcal{D}_{\theta,\nu_0}$, hence $X_i \sim \nu_0$, forall $i = 1, 2, \ldots$

$$\mu \mid X_1, \ldots, X_n \sim \mathcal{D}_{\theta\nu_0 + n\mu_n} \quad \text{(Conjugate posterior)}$$

with $\mu_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$

$$\mathsf{E}[\mu \mid X_1, \ldots, X_n] = \frac{\theta}{\theta+n}\nu_0 + \frac{n}{\theta+n}\sum_{i=1}^n \frac{\delta_{X_i}}{n}, \quad \text{(Bayes estimator)}$$

- $\mathcal{D}_{\theta\nu_0}(\mu : \mu \text{ is discrete }) = 1$

Some basic properties of $\mathcal{D}_\alpha$

- $\mathsf{E}[\mu(B)] = \nu_0(B)$, $\mathsf{Var}[\mu(B)] = \frac{\nu_0(B)(1-\nu_0(B))}{\theta+1}$

  $\mathsf{Cov}(\mu(B_2), \mu(B_2)) = \frac{\nu_0(B_1 \cap B_2) - \nu_0(B_1)\nu_0(B_2)}{\theta+1}$

If $X_i \mid \mu \overset{\text{iid}}{\sim} \mu$ and $\mu \sim \mathcal{D}_{\theta,\nu_0}$, hence $X_i \sim \nu_0$, forall $i = 1, 2, \ldots$

$$\mu \mid X_1, \ldots, X_n \sim \mathcal{D}_{\theta\nu_0 + n\mu_n} \quad \text{(Conjugate posterior)}$$

with $\mu_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$

$$\mathsf{E}[\mu \mid X_1, \ldots, X_n] = \frac{\theta}{\theta+n}\nu_0 + \frac{n}{\theta+n}\sum_{i=1}^n \frac{\delta_{X_i}}{n}, \quad \text{(Bayes estimator)}$$

- $\mathcal{D}_{\theta\nu_0}(\mu : \mu \text{ is discrete }) = 1$

Some basic properties of $\mathcal{D}_\alpha$

- $\mathsf{E}[\mu(B)] = \nu_0(B)$, $\quad\mathsf{Var}[\mu(B)] = \frac{\nu_0(B)(1-\nu_0(B))}{\theta+1}$

  $\mathsf{Cov}(\mu(B_2), \mu(B_2)) = \frac{\nu_0(B_1\cap B_2)-\nu_0(B_1)\nu_0(B_2)}{\theta+1}$

If $X_i \mid \mu \overset{\text{iid}}{\sim} \mu$ and $\mu \sim \mathcal{D}_{\theta,\nu_0}$, hence $X_i \sim \nu_0$, forall $i = 1, 2, \ldots$

$$\mu \mid X_1, \ldots, X_n \sim \mathcal{D}_{\theta\nu_0 + n\mu_n} \quad \text{(Conjugate posterior)}$$

with $\mu_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$

$$\mathsf{E}[\mu \mid X_1, \ldots, X_n] = \frac{\theta}{\theta+n}\nu_0 + \frac{n}{\theta+n}\sum_{i=1}^n \frac{\delta_{X_i}}{n}, \quad \text{(Bayes estimator)}$$

- $\mathcal{D}_{\theta\nu_0}(\mu : \mu \text{ is discrete }) = 1$

Some basic properties of $\mathcal{D}_\alpha$

- $\mathsf{E}[\mu(B)] = \nu_0(B)$, $\mathsf{Var}[\mu(B)] = \frac{\nu_0(B)(1-\nu_0(B))}{\theta+1}$

  $\mathsf{Cov}(\mu(B_2), \mu(B_2)) = \frac{\nu_0(B_1 \cap B_2) - \nu_0(B_1)\nu_0(B_2)}{\theta+1}$

If $X_i \mid \mu \overset{\text{iid}}{\sim} \mu$ and $\mu \sim \mathcal{D}_{\theta,\nu_0}$, hence $X_i \sim \nu_0$, forall $i = 1, 2, \ldots$

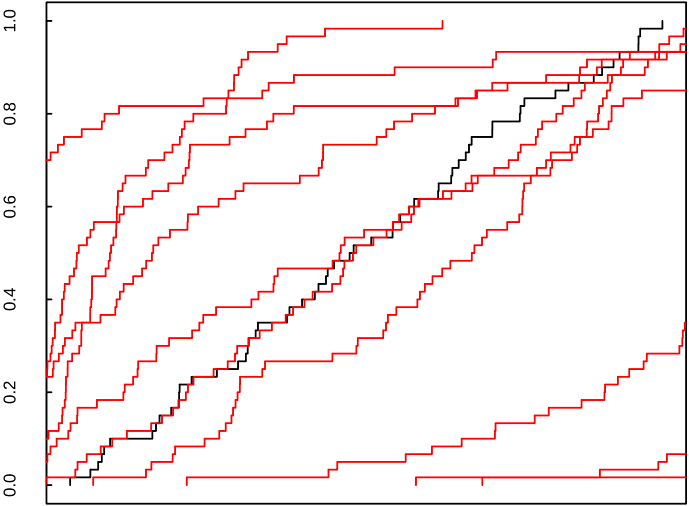
$$\mu \mid X_1, \ldots, X_n \sim \mathcal{D}_{\theta\nu_0 + n\mu_n} \quad \text{(Conjugate posterior)}$$

with $\mu_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$

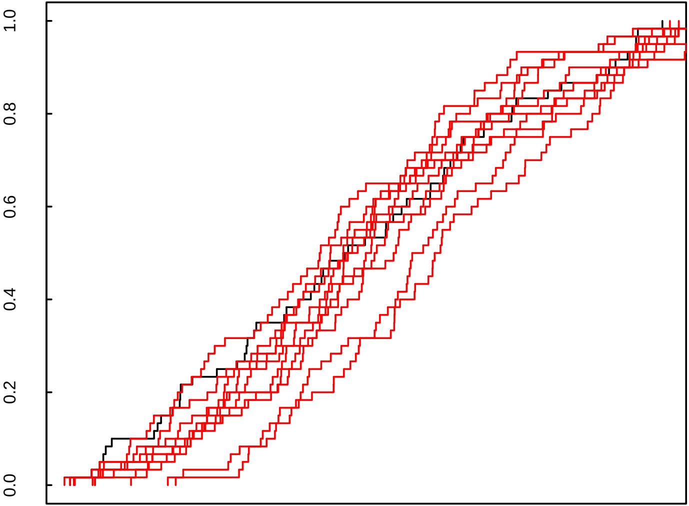
$$\mathsf{E}[\mu \mid X_1, \ldots, X_n] = \frac{\theta}{\theta+n}\nu_0 + \frac{n}{\theta+n}\sum_{i=1}^n \frac{\delta_{X_i}}{n}, \quad \text{(Bayes estimator)}$$

- $\mathcal{D}_{\theta\nu_0}(\mu : \mu \text{ is discrete }) = 1$

Precision parameter θ



Precision parameter θ



θ can be seen as a precision param.

Clustering induced by $\mathcal{D}_\alpha$

- Since $\mathcal{D}_\alpha$ a.s. discrete, $P(X_i = X_j) > 0$ for $i \neq j$
- $(X_1, \ldots, X_n)$ can be encoded to $(X_1^*, \ldots, X_{K_n}^*)$ unique values
- with random frequencies $(N_1, \ldots, N_{K_n})$, i.e. $\sum_{i=1}^{K_n} N_i = n$
- The support of $(N_1, \ldots, N_{K_n})$ is in bijection with

$$\mathscr{P}_{[n]} := \text{Set of all partitions of } \{1, \ldots, n\}$$

- Selecting $\mathcal{D}_\alpha$ induces an Exchangeable Partition Probability Function (EPPF) –Ewens (1972) and Antoniak (1974)–

Clustering induced by $\mathcal{D}_\alpha$

- Since $\mathcal{D}_\alpha$ a.s. discrete, $P(X_i = X_j) > 0$ for $i \neq j$
- $(X_1, \ldots, X_n)$ can be encoded to $(X_1^*, \ldots, X_{K_n}^*)$ unique values
- with random frequencies $(N_1, \ldots, N_{K_n})$, i.e. $\sum_{i=1}^{K_n} N_i = n$
- The support of $(N_1, \ldots, N_{K_n})$ is in bijection with

$$\mathscr{P}_{[n]} := \text{Set of all partitions of } \{1, \ldots, n\}$$

- Selecting $\mathcal{D}_\alpha$ induces an Exchangeable Partition Probability Function (EPPF) –Ewens (1972) and Antoniak (1974)–

Clustering induced by $\mathcal{D}_\alpha$

- Since $\mathcal{D}_\alpha$ a.s. discrete, $P(X_i = X_j) > 0$ for $i \neq j$
- $(X_1, \ldots, X_n)$ can be encoded to $(X_1^*, \ldots, X_{K_n}^*)$ unique values
- with random frequencies $(N_1, \ldots, N_{K_n})$, i.e. $\sum_{i=1}^{K_n} N_i = n$
- The support of $(N_1, \ldots, N_{K_n})$ is in bijection with

$$\mathscr{P}_{[n]} := \text{Set of all partitions of } \{1, \ldots, n\}$$

- Selecting $\mathcal{D}_\alpha$ induces an Exchangeable Partition Probability Function (EPPF) –Ewens (1972) and Antoniak (1974)–

$$\mathbb{P}(\text{Obs. in } k \text{ groups with freq. } n_1, \ldots, n_k) = \frac{\theta^k}{(\theta)_n} \prod_{j=1}^{k} (n_j - 1)!$$

Clustering induced by $\mathcal{D}_\alpha$

- Since $\mathcal{D}_\alpha$ a.s. discrete, $P(X_i = X_j) > 0$ for $i \neq j$
- $(X_1, \dots, X_n)$ can be encoded to $(X_1^*, \dots, X_{K_n}^*)$ unique values
- with random frequencies $(N_1, \dots, N_{K_n})$, i.e. $\sum_{i=1}^{K_n} N_i = n$
- The support of $(N_1, \dots, N_{K_n})$ is in bijection with

$$\mathscr{P}_{[n]} := \text{Set of all partitions of } \{1, \dots, n\}$$

- Selecting $\mathcal{D}_\alpha$ induces an Exchangeable Partition Probability Function (EPPF) –Ewens (1972) and Antoniak (1974)–

$$\Pi_k^{(n)}(n_1, \dots, n_k) = \frac{\theta^k}{(\theta)_n} \prod_{j=1}^{k} (n_j - 1)!$$

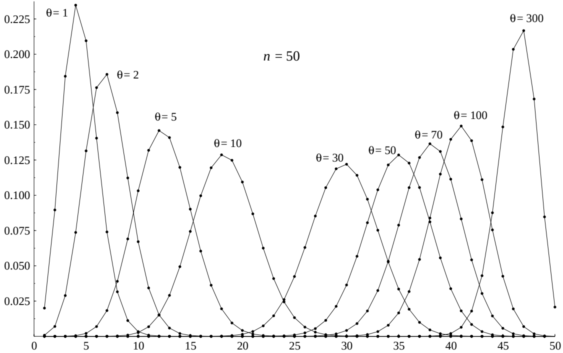
Clustering induced by $\mathcal{D}_\alpha$

- Summing over all posible partitions for fixed k

$$\mathbb{P}(K_n = k) = \frac{\theta^k}{(\theta)_n} |s(n,k)|$$

where $s(n,k)$ for $n \geq k \geq 1$ Stirling numbers of the first type.

The precision param. θ also controls the grouping.
Too informative!



Distribution of K_n, when $\mu \sim \mathcal{D}_{(\theta\nu_0)}$

BNP mixtures

For continuous data use μ-mixtures

BNP mixture models

$$Y_i \mid \mathsf{f} \overset{\text{iid}}{\sim} \mathsf{f} \quad \text{where} \quad \mathsf{f}(\cdot) = \int_{\mathbb{X}} f(\cdot \mid x)\mu(\mathrm{d}x)$$

$\mathsf{f}(\cdot)$ random density (Lo 84': $\mathsf{Q} = \mathcal{D}_\alpha$)

Density estimation & Clustering problems

BNP mixtures

For continuous data use μ-mixtures

BNP mixture models

$$
\begin{aligned}
Y_i \mid X_i &\overset{\text{ind}}{\sim} f(Y_i \mid X_i) \qquad i \geq 1 \ (e.g.\ \mathsf{f}(\cdot)\ \text{Leb. density}) \\
X_i \mid \mu &\overset{\text{iid}}{\sim} \mu \\
\mu &\sim \mathsf{Q} \qquad (e.g.\ \text{a discrete RPM})
\end{aligned}
$$

Equivalently

$$Y_i \mid \mathsf{f} \overset{\text{iid}}{\sim} \mathsf{f} \quad \text{where} \quad \mathsf{f}(\cdot) = \int_{\mathbb{X}} f(\cdot \mid x)\mu(\mathrm{d}x)$$

$\mathsf{f}(\cdot)$ random density (Lo 84': $\mathsf{Q} = \mathcal{D}_\alpha$)

Density estimation & Clustering problems

BNP mixtures

For continuous data use μ-mixtures

BNP mixture models

$$Y_i \mid X_i \overset{\text{ind}}{\sim} f(Y_i \mid X_i) \qquad i \geq 1 \ (e.g.\ \mathsf{f}(\cdot) \text{ Leb. density})$$

$$X_i \mid \mu \overset{\text{iid}}{\sim} \mu$$

$$\mu \sim \mathsf{Q} \qquad (e.g. \text{ a discrete RPM})$$

Equivalently

$$Y_i \mid \mathsf{f} \overset{\text{iid}}{\sim} \mathsf{f} \quad \text{where} \quad \mathsf{f}(\cdot) = \int_{\mathbb{X}} f(\cdot \mid x)\mu(\mathrm{d}x)$$

$\mathsf{f}(\cdot)$ random density (Lo 84': $\mathsf{Q} = \mathcal{D}_\alpha$)

Density estimation & Clustering problems

BNP mixtures

For continuous data use μ-mixtures

BNP mixture models

$$Y_i \mid X_i \overset{\text{ind}}{\sim} f(Y_i \mid X_i) \qquad i \geq 1 \ (e.g.\ \mathsf{f}(\cdot) \text{ Leb. density})$$

$$X_i \mid \mu \overset{\text{iid}}{\sim} \mu$$

$$\mu \sim \mathsf{Q} \qquad (e.g. \text{ a discrete RPM})$$

Equivalently

$$Y_i \mid \mathsf{f} \overset{\text{iid}}{\sim} \mathsf{f} \quad \text{where} \quad \mathsf{f}(\cdot) = \int_{\mathbb{X}} f(\cdot \mid x)\mu(\mathrm{d}x)$$

$\mathsf{f}(\cdot)$ random density (Lo 84': $\mathsf{Q} = \mathcal{D}_\alpha$)

Density estimation & Clustering problems

BNP mixtures: Density estimation

A Bayes density estimator, *e.g.*

$$\mathsf{E}\left[\mathsf{f}(y) \mid Y^{(n)}\right] = \sum_{k=1}^{n} \int_{\mathbb{X}} f(y \mid x) \sum_{\mathbf{p}_k \in \mathscr{P}_{[n]}^k} \mathsf{E}\left[\mu(\mathrm{d}x) \mid x_{1:k}^*\right] \mathbb{P}[x_{1:k}^* \in \mathbf{p}_k \mid Y^{(n)}]$$

where $x_{1:k}^* = (x_1^*, \ldots, x_k^*)$ and $\mathbf{p}_k \in \mathscr{P}_{[n]}^k$

- $\mathsf{E}\left[\mu(\mathrm{d}x) \mid x_{1:k}^*\right]$ denotes the predictive
  - For large n virtually impossible to evaluate exactly
  - The need of MCMC methods is evident

BNP mixtures: Density estimation

A Bayes density estimator, *e.g.*

$$\mathsf{E}\left[\mathsf{f}(y) \mid Y^{(n)}\right] = \sum_{k=1}^{n} \int_{\mathbb{X}} f(y \mid x) \sum_{\mathbf{p}_k \in \mathscr{P}_{[n]}^k} \mathsf{E}\left[\mu(\mathrm{d}x) \mid x_{1:k}^*\right] \mathbb{P}[x_{1:k}^* \in \mathbf{p}_k \mid Y^{(n)}]$$

where $x_{1:k}^* = (x_1^*, \ldots, x_k^*)$ and $\mathbf{p}_k \in \mathscr{P}_{[n]}^k$

- $\mathsf{E}\left[\mu(\mathrm{d}x) \mid x_{1:k}^*\right]$ denotes the predictive
  - For large n virtually impossible to evaluate exactly
  - The need of MCMC methods is evident

BNP mixtures: Density estimation

A Bayes density estimator, *e.g.*

$$\mathsf{E}\left[\mathsf{f}(y) \mid Y^{(n)}\right] = \sum_{k=1}^{n} \int_{\mathbb{X}} f(y \mid x) \sum_{\mathbf{p}_k \in \mathscr{P}_{[n]}^k} \mathsf{E}\left[\mu(\mathrm{d}x) \mid x_{1:k}^*\right] \mathbb{P}[x_{1:k}^* \in \mathbf{p}_k \mid Y^{(n)}]$$

where $x_{1:k}^* = (x_1^*, \ldots, x_k^*)$ and $\mathbf{p}_k \in \mathscr{P}_{[n]}^k$

- $\mathsf{E}\left[\mu(\mathrm{d}x) \mid x_{1:k}^*\right]$ denotes the predictive
  - ▷ For large n virtually impossible to evaluate exactly
  - ▷ The need of MCMC methods is evident

BNP mixtures: Density estimation

A Bayes density estimator, *e.g.*

$$\mathsf{E}\left[\mathsf{f}(y) \mid Y^{(n)}\right] = \sum_{k=1}^{n} \int_{\mathbb{X}} f(y \mid x) \sum_{\mathbf{p}_k \in \mathscr{P}_{[n]}^k} \mathsf{E}\left[\mu(\mathrm{d}x) \mid x_{1:k}^*\right] \mathbb{P}[x_{1:k}^* \in \mathbf{p}_k \mid Y^{(n)}]$$

where $x_{1:k}^* = (x_1^*, \ldots, x_k^*)$ and $\mathbf{p}_k \in \mathscr{P}_{[n]}^k$

- $\mathsf{E}\left[\mu(\mathrm{d}x) \mid x_{1:k}^*\right]$ denotes the predictive
  - ▷ For large n virtually impossible to evaluate exactly
  - ▷ The need of MCMC methods is evident

BNP mixtures: Posterior distribution on $\mathscr{P}_{[n]}$

- Posterior clustering under BNP mixture (or clustering likelihood!)

$$\mathbb{P}[\mathbf{p}_k \mid Y^{(n)}] \propto \Pi_k^{(n)}(n_1, \ldots, n_k) \prod_{j=1}^{k} \int_{\mathbb{X}} \prod_{i \in \mathcal{J}_j} f(y_i \mid x_i) \nu_0(\mathrm{d}x_i)$$

where as before $\mathbf{p}_k \in \mathscr{P}_{[n]}^k$ and $\mathcal{J}_j := \{i : X_i = X_j^*\}, \quad j = 1, \ldots, k$

▷ No longer exchangeable due to effect of $f(\cdot \mid x)$ the y's

- Summing over all the partitions for fixed k we obtain the posterior on the number of groups of size $k = 1, \ldots, n$

$$\mathbb{P}[K_n = k \mid Y^{(n)}] = \sum_{\mathbf{p}_k \in \mathcal{P}_{[n]}^k} \mathbb{P}[\mathbf{p}_k \mid Y^{(n)}]$$

BNP mixtures: Posterior distribution on $\mathscr{P}_{[n]}$

- Posterior clustering under BNP mixture (or clustering likelihood!)

$$\mathbb{P}[\mathbf{p}_k \mid Y^{(n)}] \propto \Pi_k^{(n)}(n_1, \ldots, n_k) \prod_{j=1}^{k} \int_{\mathbb{X}} \prod_{i \in \mathcal{J}_j} f(y_i \mid x_i)\nu_0(\mathrm{d}x_i)$$

where as before $\mathbf{p}_k \in \mathscr{P}_{[n]}^k$ and $\mathcal{J}_j := \{i : X_i = X_j^*\}, \quad j = 1, \ldots, k$

▷ No longer exchangeable due to effect of $f(\cdot \mid x)$ the y's

- Summing over all the partitions for fixed k we obtain the posterior on the number of groups of size $k = 1, \ldots, n$

$$\mathbb{P}[K_n = k \mid Y^{(n)}] = \sum_{\mathbf{p}_k \in \mathcal{P}_{[n]}^k} \mathbb{P}[\mathbf{p}_k \mid Y^{(n)}]$$

BNP mixtures: Posterior distribution on $\mathscr{P}_{[n]}$

- Posterior clustering under BNP mixture (or clustering likelihood!)

$$\mathbb{P}[\mathbf{p}_k \mid Y^{(n)}] \propto \Pi_k^{(n)}(n_1, \ldots, n_k) \prod_{j=1}^{k} \int_{\mathbb{X}} \prod_{i \in \mathcal{J}_j} f(y_i \mid x_i) \nu_0(\mathrm{d}x_i)$$

where as before $\mathbf{p}_k \in \mathscr{P}_{[n]}^k$ and $\mathcal{J}_j := \{i : X_i = X_j^*\}, \quad j = 1, \ldots, k$

▷ No longer exchangeable due to effect of $f(\cdot \mid x)$ the y's

- Summing over all the partitions for fixed k we obtain the posterior on the number of groups of size $k = 1, \ldots, n$

$$\mathbb{P}[K_n = k \mid Y^{(n)}] = \sum_{\mathbf{p}_k \in \mathcal{P}_{[n]}^k} \mathbb{P}[\mathbf{p}_k \mid Y^{(n)}]$$

BNP mixtures: Toy example (10 data points)

- $f(y \mid \theta) = \mathsf{N}(y \mid \mu, \lambda^{-1})$, $\mu \sim \mathcal{D}_{\theta\nu_0}$

  $\nu_0(\mathrm{d}\mu, \mathrm{d}\lambda) = \mathsf{N}(\mu \mid 0, \frac{10}{\lambda})\mathsf{Exp}(\lambda \mid 1)\mathrm{d}\mu\mathrm{d}\lambda$

▷ $\mathbf{p}_2 = \{\{y_1, \dots, y_4\}, \{y_5, \dots, y_{10}\}\}$

→ integer partition $(n_1, n_2) = (4, 6)$

▷ If $\theta = 1$ posterior mode is at $\mathbf{p}_2$ with $\mathbb{P}[\mathbf{p}_2 \mid y^{(n)}] = 0.332$

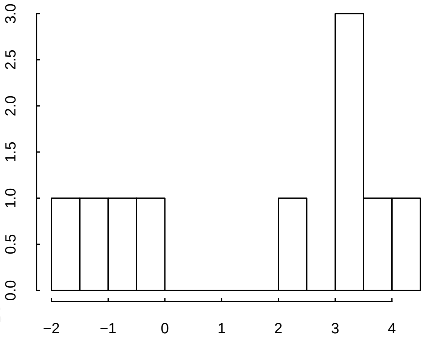
▷ Posterior on #groups: mode at $k = 3$

with $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.39$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.37$

▷ If $\theta = 0.5$: $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.31$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.59$ – $\mathsf{E}(K_{10}) = 2.1$ –

▷ If $\theta = 5$: $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.80$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.02$ – $\mathsf{E}(K_{10}) = 5.8$ –

Need to randomize (put a prior) on θ for $\mathcal{D}_{\theta P_0}$



BNP mixtures: Toy example (10 data points)

- $f(y \mid \theta) = \mathsf{N}(y \mid \mu, \lambda^{-1})$, $\mu \sim \mathcal{D}_{\theta\nu_0}$

  $\nu_0(\mathrm{d}\mu, \mathrm{d}\lambda) = \mathsf{N}(\mu \mid 0, \frac{10}{\lambda})\mathsf{Exp}(\lambda \mid 1)\mathrm{d}\mu\mathrm{d}\lambda$

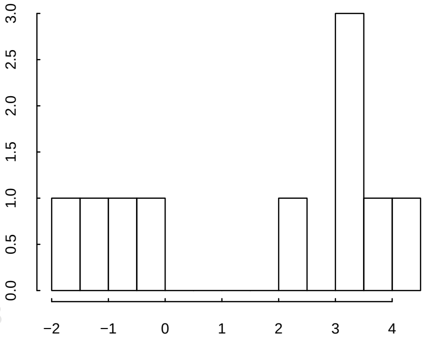
▷ $\mathbf{p}_2 = \{\{y_1, \ldots, y_4\}, \{y_5, \ldots, y_{10}\}\}$

→ integer partition $(n_1, n_2) = (4, 6)$

▷ If $\theta = 1$ posterior mode is at $\mathbf{p}_2$

with $\mathbb{P}[\mathbf{p}_2 \mid y^{(n)}] = 0.332$

▷ Posterior on #groups: mode at $k = 3$



with $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.39$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.37$

▷ If $\theta = 0.5$: $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.31$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.59$ – $\mathsf{E}(K_{10}) = 2.1$ –

▷ If $\theta = 5$: $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.80$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.02$ – $\mathsf{E}(K_{10}) = 5.8$ –

Need to randomize (put a prior) on θ for $\mathcal{D}_{\theta P_0}$

BNP mixtures: Toy example (10 data points)

- $f(y \mid \theta) = \mathsf{N}(y \mid \mu, \lambda^{-1})$, $\mu \sim \mathcal{D}_{\theta\nu_0}$

  $\nu_0(\mathrm{d}\mu, \mathrm{d}\lambda) = \mathsf{N}(\mu \mid 0, \frac{10}{\lambda})\mathsf{Exp}(\lambda \mid 1)\mathrm{d}\mu\mathrm{d}\lambda$

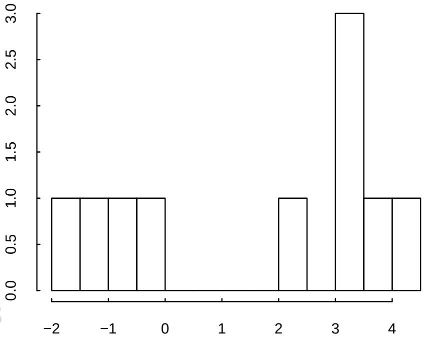
▷ $\mathbf{p}_2 = \{\{y_1, \ldots, y_4\}, \{y_5, \ldots, y_{10}\}\}$

→ integer partition $(n_1, n_2) = (4, 6)$

▷ If $\theta = 1$ posterior mode is at $\mathbf{p}_2$

with $\mathbb{P}[\mathbf{p}_2 \mid y^{(n)}] = 0.332$

▷ Posterior on #groups: mode at $k = 3$



with $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.39$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.37$

▷ If $\theta = 0.5$: $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.31$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.59$ – $\mathsf{E}(K_{10}) = 2.1$ –

▷ If $\theta = 5$: $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.80$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.02$ – $\mathsf{E}(K_{10}) = 5.8$ –

Need to randomize (put a prior) on θ for $\mathcal{D}_{\theta P_0}$

BNP mixtures: Toy example (10 data points)

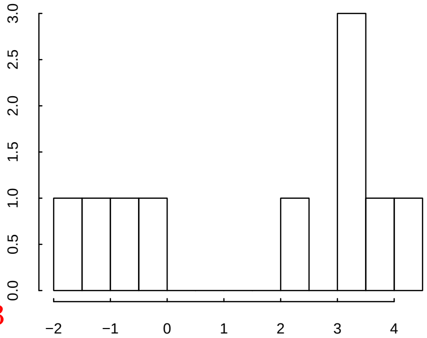
- $f(y \mid \theta) = \mathsf{N}(y \mid \mu, \lambda^{-1})$, $\mu \sim \mathcal{D}_{\theta\nu_0}$
  $\nu_0(\mathrm{d}\mu, \mathrm{d}\lambda) = \mathsf{N}(\mu \mid 0, \frac{10}{\lambda})\mathsf{Exp}(\lambda \mid 1)\mathrm{d}\mu\mathrm{d}\lambda$

▷ $\mathbf{p}_2 = \{\{y_1, \ldots, y_4\}, \{y_5, \ldots, y_{10}\}\}$

→ integer partition $(n_1, n_2) = (4, 6)$

▷ If $\theta = 1$ posterior mode is at $\mathbf{p}_2$ with $\mathbb{P}[\mathbf{p}_2 \mid y^{(n)}] = 0.332$

▷ Posterior on #groups: mode at $k = 3$



with $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.39$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.37$

▷ If $\theta = 0.5$: $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.31$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.59$ – $\mathsf{E}(K_{10}) = 2.1$ –

▷ If $\theta = 5$: $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.80$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.02$ – $\mathsf{E}(K_{10}) = 5.8$ –

Need to randomize (put a prior) on θ for $\mathcal{D}_{\theta P_0}$

BNP mixtures: Toy example (10 data points)

- $f(y \mid \theta) = \mathsf{N}(y \mid \mu, \lambda^{-1})$, $\mu \sim \mathcal{D}_{\theta\nu_0}$

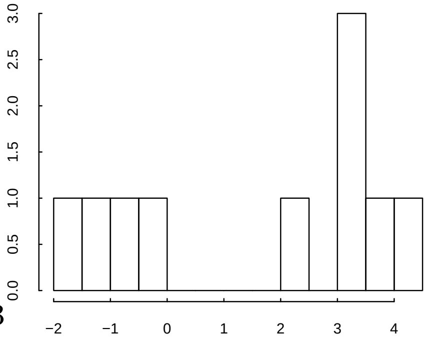
  $\nu_0(\mathrm{d}\mu, \mathrm{d}\lambda) = \mathsf{N}(\mu \mid 0, \frac{10}{\lambda})\mathsf{Exp}(\lambda \mid 1)\mathrm{d}\mu\mathrm{d}\lambda$

▷ $\mathbf{p}_2 = \{\{y_1, \ldots, y_4\}, \{y_5, \ldots, y_{10}\}\}$

→ integer partition $(n_1, n_2) = (4, 6)$

▷ If $\theta = 1$ posterior mode is at $\mathbf{p}_2$ with $\mathbb{P}[\mathbf{p}_2 \mid y^{(n)}] = 0.332$

▷ Posterior on #groups: mode at $k = 3$



with $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.39$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.37$

▷ If $\theta = 0.5$: $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.31$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.59$ – $\mathsf{E}(K_{10}) = 2.1$ –

▷ If $\theta = 5$: $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.80$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.02$ – $\mathsf{E}(K_{10}) = 5.8$ –

Need to randomize (put a prior) on θ for $\mathcal{D}_{\theta P_0}$

BNP mixtures: Toy example (10 data points)

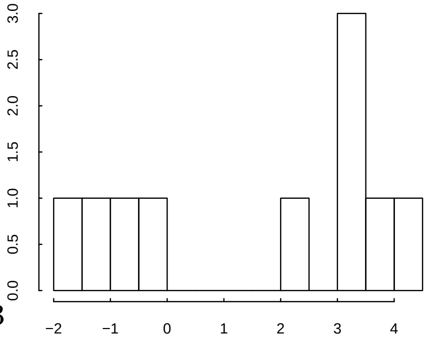
- $f(y \mid \theta) = \mathsf{N}(y \mid \mu, \lambda^{-1})$, $\mu \sim \mathcal{D}_{\theta\nu_0}$
  $\nu_0(\mathrm{d}\mu, \mathrm{d}\lambda) = \mathsf{N}(\mu \mid 0, \frac{10}{\lambda})\mathsf{Exp}(\lambda \mid 1)\mathrm{d}\mu\mathrm{d}\lambda$

▷ $\mathbf{p}_2 = \{\{y_1, \ldots, y_4\}, \{y_5, \ldots, y_{10}\}\}$

→ integer partition $(n_1, n_2) = (4, 6)$

▷ If $\theta = 1$ posterior mode is at $\mathbf{p}_2$ with $\mathbb{P}[\mathbf{p}_2 \mid y^{(n)}] = 0.332$

▷ Posterior on #groups: mode at $k = 3$



with $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.39$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.37$

▷ If $\theta = 0.5$: $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.31$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.59$ – $\mathsf{E}(K_{10}) = 2.1$ –

▷ If $\theta = 5$: $\mathbb{P}[K_{10} = 3 \mid y^{(n)}] = 0.80$ & $\mathbb{P}[K_{10} = 2 \mid y^{(n)}] = 0.02$ – $\mathsf{E}(K_{10}) = 5.8$ –

Need to randomize (put a prior) on θ for $\mathcal{D}_{\theta P_0}$

A simplified RPM: Geometric weights

- Given that for the $\mathcal{D}_{\theta\nu_0}$ a randomization of θ is needed we could instead consider the simplified RPM

$$\mu(B) = \sum_{i=1}^{\infty} \mathsf{E}[\mathsf{w}_i]\delta_{z_i}(B) = \sum_{i=1}^{\infty} \lambda(1-\lambda)^{i-1}\delta_{z_i}(B)$$

where $\lambda = (\theta+1)^{-1}$ and $\lambda \sim \mathsf{Be}(a,b)$, *i.e.* with geometric weights.

▷ Namely, a DP with the randomness of the weights removed!

▷ This RPM has ordered weights!

▷ Still has full support wrt weak topology

A simplified RPM: Geometric weights

- Given that for the $\mathcal{D}_{\theta\nu_0}$ a randomization of θ is needed we could instead consider the simplified RPM

$$\mu(B) = \sum_{i=1}^{\infty} \mathsf{E}[\mathsf{w}_i]\delta_{z_i}(B) = \sum_{i=1}^{\infty} \lambda(1-\lambda)^{i-1}\delta_{z_i}(B)$$

where $\lambda = (\theta + 1)^{-1}$ and $\lambda \sim \mathsf{Be}(a, b)$, *i.e.* with geometric weights.

▷ Namely, a DP with the randomness of the weights removed!

▷ This RPM has ordered weights!

▷ Still has full support wrt weak topology

A simplified RPM: Geometric weights

- Given that for the $\mathcal{D}_{\theta\nu_0}$ a randomization of θ is needed we could instead consider the simplified RPM

$$\mu(B) = \sum_{i=1}^{\infty} \mathsf{E}[\mathsf{w}_i]\delta_{z_i}(B) = \sum_{i=1}^{\infty} \lambda(1-\lambda)^{i-1}\delta_{z_i}(B)$$

where $\lambda = (\theta + 1)^{-1}$ and $\lambda \sim \mathsf{Be}(a, b)$, *i.e.* with geometric weights.

▷ Namely, a DP with the randomness of the weights removed!

▷ This RPM has ordered weights!

▷ Still has full support wrt weak topology

A simplified RPM: Geometric weights

- Given that for the $\mathcal{D}_{\theta\nu_0}$ a randomization of θ is needed we could instead consider the simplified RPM

$$\mu(B) = \sum_{i=1}^{\infty} \mathsf{E}[\mathsf{w}_i]\delta_{z_i}(B) = \sum_{i=1}^{\infty} \lambda(1-\lambda)^{i-1}\delta_{z_i}(B)$$

where $\lambda = (\theta + 1)^{-1}$ and $\lambda \sim \mathsf{Be}(a, b)$, *i.e.* with geometric weights.

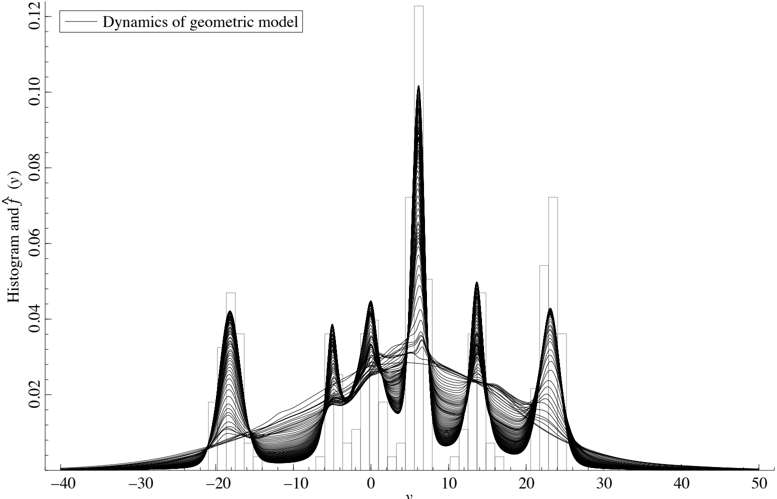
▷ Namely, a DP with the randomness of the weights removed!

▷ This RPM has ordered weights!

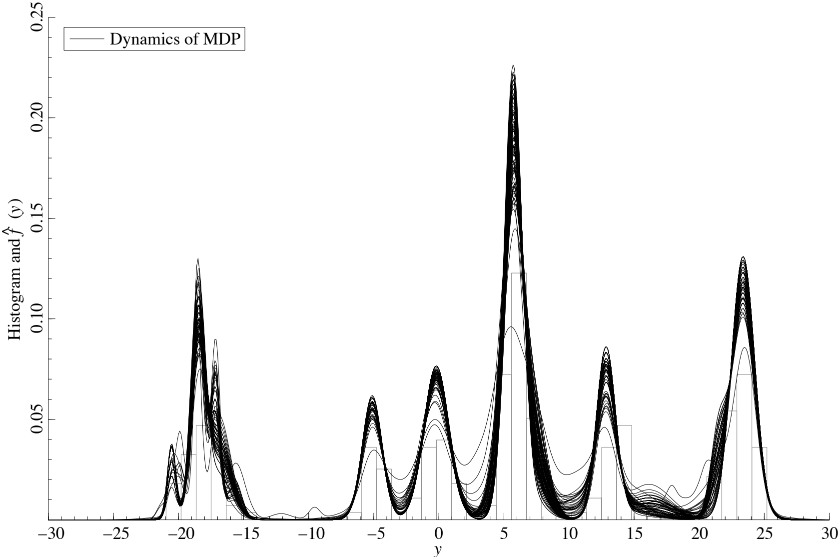
▷ Still has full support wrt weak topology

100 iter. BNP mixture model based on geom. weights

$$f(y) = \int_{\mathbb{X}} f(y \mid z)\mu(\mathrm{d}z) = \sum_{l \geq 1} \lambda(1-\lambda)^{l-1} f(y \mid \theta_l)$$



DP mixture


Dynamics of MDP
Histogram and $\hat{f}(y)$
0.25
0.20
0.15
0.10
0.05
–30
–25
–20
–15
–10
–5
0
5
10
15
20
25
30
y

Properties

So why is that it works so well?

Weights are ordered

But let us find an alternative explanation for it!

Properties

So why is that it works so well?

Weights are ordered

But let us find an alternative explanation for it!

Properties

So why is that it works so well?

Weights are ordered

But let us find an alternative explanation for it!

MCMC methods: via slice sampler (Walker 07')

$$\mathsf{f}(y) = \sum_{i=1}^{\infty} \mathsf{w}_i f(y \mid z_i) \qquad (*)$$

▷ Infinite summation becomes a problem since w_i's are not ordered

- Augment (*) through a uniform latent variable

$$f(y,u) = \sum_{j=1}^{\infty} \mathbb{I}(u < \mathsf{w}_j) f(y \mid z_i)$$

- Given u the set $A_u := \{j : \mathsf{w}_j > u\}$ is finite.

The infinite summation disappear since the summation in

$$f(y \mid u) = \frac{1}{\# A_u} \sum_{j \in A_u} f(y \mid z_i) \qquad \text{is finite}$$

MCMC methods: via slice sampler (Walker 07')

$$\mathsf{f}(y) = \sum_{i=1}^{\infty} \mathsf{w}_i f(y \mid z_i) \qquad (*)$$

▷ Infinite summation becomes a problem since w_i's are not ordered

- Augment (*) through a uniform latent variable

$$f(y,u) = \sum_{j=1}^{\infty} \mathbb{I}(u < \mathsf{w}_j) f(y \mid z_i)$$

- Given u the set $A_u := \{j : \mathsf{w}_j > u\}$ is finite.

The infinite summation disappear since the summation in

$$f(y \mid u) = \frac{1}{\# A_u} \sum_{j \in A_u} f(y \mid z_i) \qquad \text{is finite}$$

MCMC methods: via slice sampler (Walker 07')

$$\mathsf{f}(y) = \sum_{i=1}^{\infty} \mathsf{w}_i f(y \mid z_i) \qquad (*)$$

▷ Infinite summation becomes a problem since w_i's are not ordered

- Augment (*) through a uniform latent variable

$$f(y,u) = \sum_{j=1}^{\infty} \mathbb{I}(u < \mathsf{w}_j) f(y \mid z_i)$$

- Given u the set $A_u := \{j : \mathsf{w}_j > u\}$ is finite.

The infinite summation disappear since the summation in

$$f(y \mid u) = \frac{1}{\# A_u} \sum_{j \in A_u} f(y \mid z_i) \qquad \text{is finite}$$

Random set A_u

- So A_u is a finite subset of the set of positive integers
- For the DP weights the A_u typically generates set of integers with gaps, e.g. $\{2, 5, 16, 40, 200, 3029\}$
- But given that the representation

$$\mu(B) = \sum_{i=1}^{\infty} \mathsf{w}_i \, \delta_{z_i}(B), \quad B \in \mathcal{X}$$

  includes a infinite number of locations z_i's
- The same mass could be attained with a set $\{1, 2, 3, 4, 5, 6\}$
- No need for the gaps!!

Random set A_u

- So A_u is a finite subset of the set of positive integers
- For the DP weights the A_u typically generates set of integers with gaps, e.g. $\{2, 5, 16, 40, 200, 3029\}$
- But given that the representation

$$\mu(B) = \sum_{i=1}^{\infty} \mathsf{w}_i \, \delta_{z_i}(B), \quad B \in \mathcal{X}$$

includes a infinite number of locations z_i's
- The same mass could be attained with a set $\{1, 2, 3, 4, 5, 6\}$
- No need for the gaps!!

Random set A_u

- So A_u is a finite subset of the set of positive integers
- For the DP weights the A_u typically generates set of integers with gaps, e.g. $\{2, 5, 16, 40, 200, 3029\}$
- But given that the representation

$$\mu(B) = \sum_{i=1}^{\infty} \mathsf{w}_i \, \delta_{z_i}(B), \quad B \in \mathcal{X}$$

includes a infinite number of locations z_i's

- The same mass could be attained with a set $\{1, 2, 3, 4, 5, 6\}$
- No need for the gaps!!

Random set A_u

- So A_u is a finite subset of the set of positive integers
- For the DP weights the A_u typically generates set of integers with gaps, e.g. $\{2, 5, 16, 40, 200, 3029\}$
- But given that the representation

$$\mu(B) = \sum_{i=1}^{\infty} \mathsf{w}_i \, \delta_{z_i}(B), \quad B \in \mathcal{X}$$

  includes a infinite number of locations z_i's
- The same mass could be attained with a set $\{1, 2, 3, 4, 5, 6\}$
- No need for the gaps!!

Random set A_u

- So A_u is a finite subset of the set of positive integers
- For the DP weights the A_u typically generates set of integers with gaps, e.g. $\{2, 5, 16, 40, 200, 3029\}$
- But given that the representation

$$\mu(B) = \sum_{i=1}^{\infty} \mathsf{w}_i \, \delta_{z_i}(B), \quad B \in \mathcal{X}$$

includes a infinite number of locations z_i's
- The same mass could be attained with a set $\{1, 2, 3, 4, 5, 6\}$
- No need for the gaps!!

A different construction of the weights

Consider the random density defined by

$$f(y \mid A) = \frac{1}{\#A} \sum_{j \in A} f(y \mid z_i)$$

with A a finite random subset of $\mathbb{N}_+$

- Here we look at $A = \{1, \ldots, N\}$ with $N \sim q_N$ so

$$f(y \mid N) = \frac{1}{N} \sum_{j=1}^{N} f(y \mid z_i)$$

which marginalizing corresponds to

$$f(y) = \sum_{i=1}^{\infty} \left\{ \frac{1}{N} \sum_{l=1}^{N} f(y \mid z_i) \right\} q_N$$

A different construction of the weights

Consider the random density defined by

$$f(y \mid A) = \frac{1}{\#A} \sum_{j \in A} f(y \mid z_i)$$

with A a finite random subset of $\mathbb{N}_+$

- Here we look at $A = \{1, \ldots, N\}$ with $N \sim q_N$ so

$$f(y \mid N) = \frac{1}{N} \sum_{j=1}^{N} f(y \mid z_i)$$

which marginalizing corresponds to

$$f(y) = \sum_{i=1}^{\infty} \left\{ \frac{1}{N} \sum_{l=1}^{N} f(y \mid z_i) \right\} q_N$$

A different construction of the weights

This can be seen as a BNP mixture with weights

$$\mathsf{w}_i = \sum_{N=i}^{\infty} \frac{q_N}{N}$$

q_N a prob. mass function on $\mathbb{N}_+$

- Weights are ordered!

For example if q_N is a $\mathsf{Neg-Bin}(r, \lambda)$ we get

$$\mathsf{w}_i = \frac{1}{i}\binom{i+r-2}{r-1}\lambda^r(1-\lambda)^{i-1}{}_2\mathsf{F}_1(1, i+r-1; i+1; \lambda)$$

which for $r = 2$ we recover the geometric case

$$\mathsf{w}_i = \lambda(1-\lambda)^{i-1}$$

A different construction of the weights

This can be seen as a BNP mixture with weights

$$\mathsf{w}_i = \sum_{N=i}^{\infty} \frac{q_N}{N}$$

q_N a prob. mass function on $\mathbb{N}_+$

- Weights are ordered!

For example if q_N is a $\mathsf{Neg-Bin}(r, \lambda)$ we get

$$\mathsf{w}_i = \frac{1}{i}\binom{i+r-2}{r-1}\lambda^r(1-\lambda)^{i-1}{}_2\mathsf{F}_1(1, i+r-1; i+1; \lambda)$$

which for $r = 2$ we recover the geometric case

$$\mathsf{w}_i = \lambda(1-\lambda)^{i-1}$$

A different construction of the weights

This can be seen as a BNP mixture with weights

$$\mathsf{w}_i = \sum_{N=i}^{\infty} \frac{q_N}{N}$$

q_N a prob. mass function on $\mathbb{N}_+$

- Weights are ordered!

For example if q_N is a $\mathsf{Neg-Bin}(r, \lambda)$ we get

$$\mathsf{w}_i = \frac{1}{i}\binom{i+r-2}{r-1}\lambda^r(1-\lambda)^{i-1}{}_2\mathsf{F}_1(1, i+r-1; i+1; \lambda)$$

which for $r = 2$ we recover the geometric case

$$\mathsf{w}_i = \lambda(1-\lambda)^{i-1}$$

Dependent processes

What happens with a different type of dependence?

Namely, we have observations typically capture with models such as:

- $X_{n+1} = \phi X_n + \varepsilon_t$
- $dX_t = a(X_t, \theta)dt + \sigma(X_t, \theta)dW_t$
- $X_i = f(\mathbf{Z}, \beta)$
- etc..

We still want to be nonparametric!

- Nonparametric dependent random measures, *i.e*

$$\{\mu_n\}_{n=0}^{\infty}, \quad \{\mu_t\}_{t\geq 0}, \quad \{\mu_z\}_{z\in Z}$$

Dependent processes

What happens with a different type of dependence?

Namely, we have observations typically capture with models such as:

- $X_{n+1} = \phi X_n + \varepsilon_t$
- $dX_t = a(X_t, \theta)dt + \sigma(X_t, \theta)dW_t$
- $X_i = f(\mathbf{Z}, \beta)$
- etc..

We still want to be nonparametric!

- Nonparametric dependent random measures, *i.e*

$$\{\mu_n\}_{n=0}^{\infty}, \quad \{\mu_t\}_{t\geq 0}, \quad \{\mu_z\}_{z\in Z}$$

Dependent processes

What happens with a different type of dependence?

Namely, we have observations typically capture with models such as:

- $X_{n+1} = \phi X_n + \varepsilon_t$
- $dX_t = a(X_t, \theta)dt + \sigma(X_t, \theta)dW_t$
- $X_i = f(\mathbf{Z}, \beta)$
- etc..

We still want to be nonparametric!

- Nonparametric dependent random measures, *i.e*

$$\{\mu_n\}_{n=0}^{\infty}, \quad \{\mu_t\}_{t\geq 0}, \quad \{\mu_z\}_{z\in Z}$$

Dependent processes

What happens with a different type of dependence?

Namely, we have observations typically capture with models such as:

- $X_{n+1} = \phi X_n + \varepsilon_t$
- $dX_t = a(X_t, \theta)dt + \sigma(X_t, \theta)dW_t$
- $X_i = f(\mathbf{Z}, \beta)$
- etc..

We still want to be nonparametric!

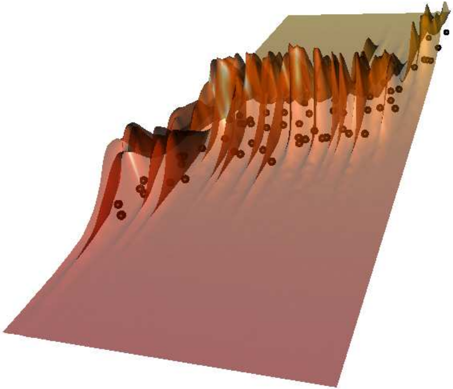
- Nonparametric dependent random measures, *i.e*

$$\{\mu_n\}_{n=0}^{\infty}, \quad \{\mu_t\}_{t\geq 0}, \quad \{\mu_z\}_{z\in Z}$$

Covariate dependent

- Introduce dependence through $\{\lambda_z\}_{z\in\mathcal{Z}}$

$$\lambda_z = \frac{e^{\xi(z)}}{1+e^{\xi(z)}}, \qquad \{\xi(z)\} \sim \mathrm{GP}(\mu,\sigma)$$



$$\eta_z := \int y\, f_z(y) \mathrm{d}y$$

$$f_z(y) = \sum_{l\geq 1} \lambda_z(1-\lambda_z)^{l-1} f(y \mid \theta_l)$$

Let's look at a continuous time dependent NP process.

$$\mu(t) = \sum_{i \geq 0} w_i(t)\, \delta_{x_i(t)}$$

where, for each $i \geq 0$, $\{w_i(t)\}_{t\geq 0}$, $\{x_i(t)\}_{t\geq 0}$ are certain *ad hoc* stochastic processes.

- In general we might think $\mu(t)$ inherits some of the continuity and stability properties of the processes $\{w_i(t)\}$ and $\{x_i(t)\}$

Geometric stick-breaking process

Definition

Let $\{\mu(t), t \geq 0\}$ a stochastic process with values on $\mathcal{P}_{\mathbb{X}}$ defined on $(\Omega, \mathscr{F}, \mathbb{P})$ such that for each $t \geq 0$

$$\mu(t) = \lambda_t \sum_{i \geq 0} (1 - \lambda_t)^{i-1} \, \delta_{x_i}$$

where ν_0 is an non-atomic distribution on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and $\{\lambda_t\}_{t \geq 0}$ is a diffusion process with paths in $\mathcal{C}_{[0,1]}([0, \infty))$ and infinitesimal generator

$$\mathcal{A} = \left[\frac{c}{a+b-1}(a - (a+b)\lambda)\right] \frac{\mathrm{d}}{\mathrm{d}\lambda} + \frac{c}{a+b-1}\lambda(1-\lambda)\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}$$

with domain $\mathscr{D}(\mathcal{A}) = \mathcal{C}^2([0,1])$. We name $\{\mu(t), t \geq 0\}$ the Geometric Stick Breaking process with parameters (a, b, c, ν_0) denoted by $\mathrm{GSB}(a, b, c, \nu_0)$

Geometric stick-breaking process

- $\{\lambda_t\}_{t\geq 0}$ is a diffusion process with the following features:
  - Stationary with invariant distribution $\mathsf{Be}(a, b)$
  - Reversible
  - When $c := (a + b - 1)/2 \Rightarrow \{\lambda_t\}_{t\geq 0}$ Wright-Fisher model

Which of these properties are inherited by $\mu_t \sim \text{GSBP}(a, b, c, \nu_0)$?

- Let $\mathscr{P}_g(\mathbb{X}) \subset \mathcal{P}_{\mathbb{X}}$ the set of purely atomic probability measures on $\mathbb{X}$

Propiedades GSB(a, b, c, ν_0)

Proposition

Let $\{\mu_t\}_{t\geq 0}$ a GSB(a, b, c, ν_0) process. Then, $\{\mu_t\}_{t\geq 0}$ has an infinitesimal generator given by

$$\mathcal{B}\varphi_m(\mu) = \left(\frac{a}{2}(1-\lambda) - \frac{b}{2}\lambda\right) \sum_{i_1,\dots,i_m\geq 1} f(x_{i_1},\dots,x_{i_m}) \frac{\partial}{\partial\lambda} h(\lambda; m, i_1, \dots, i_m)$$
$$+ \frac{1}{2}\lambda(1-\lambda) \sum_{i_1,\dots,i_m\geq 1} f(x_{i_1},\dots,x_{i_m}) \frac{\partial^2}{\partial\lambda^2} h(\lambda; m, i_1, \dots, i_m)$$

with domain

$$\mathscr{D}(\mathcal{B}) = \left\{\varphi \in C(\mathscr{P}_g(\mathbb{X})) :\ \varphi = \varphi_m(\mu) = \langle f, \mu^m\rangle,\ f \in C(\mathbb{X}^m),\ m \in \mathbb{N}\right\}$$

and where

$$h(\lambda; m, i_1, \dots, i_m) = \lambda_t^m (1-\lambda_t)^{\sum_{j=1}^m i_j - m}.$$

Properties of $\mathrm{GSB}(a, b, c, \nu_0)$

Proposition

Let $\mathbb{X}$ be a Polish space, $\{\mu_t\}_{t\geq 0}$ a $\mathrm{GSB}(a, b, c, \nu_0)$ process on $\mathscr{P}_g(\mathbb{X})$. Hence $\{\mu_t\}_{t\geq 0}$ is a Feller process with trajectories on $\mathcal{C}_{\mathscr{P}_g(\mathbb{X})}([0, \infty))$.

Proposition

Let $\mathbb{X}$ be a Polish space, $\{\mu_t\}_{t\geq 0}$ a $\mathrm{GSB}(a, b, c, \nu_0)$ process on $\mathscr{P}_g(\mathbb{X})$. Hence $\{\mu_t\}_{t\geq 0}$ is reversible and strictly stationary.

Summing up, $\{\mu_t\}_{t\geq 0}$ is a diffusion process with values in the space of purely atomic probability measures, with continuous trajectories, stationary and reversible!

Mixtures of GSB(a, b, c, ν_0) process

- If we require that the process takes values on $\mathscr{P}_c(\mathbb{X}) \subset \mathcal{P}_{\mathbb{X}}$ (all continuous prob. measures), we consider

$$f_t(y) = \int_{\mathbb{X}} f(y \mid z)\mu_t(\mathrm{d}z) = \sum_{l \geq 1} \lambda_t(1-\lambda_t)^{l-1} f(y \mid \theta_l)$$

where $f(\cdot \mid \theta)$ is a well defined Lebesgue density and $\theta_l \overset{\text{iid}}{\sim} \nu_0$, ν_0 non-atomic.

Estimation for single trajectory data

Sup. we observe only one trayectory $\{y_{t_i}\}_{i=1}^n$ and we use the mixture model. In hierarchical notation

$$\begin{aligned} y_i \mid t_i, x_i &\sim f(\cdot \mid x_i) \\ \{x_i\} &\sim \mu_t \\ \mu_t &\sim \mathrm{GSB}(a, b, c, \nu_0). \end{aligned} \tag{1}$$

where $x_i := x_{t_i}$.

- We will estimate this model through a Gibbs sampler algorithm

Diffusion part $\{\lambda_t\}$

The transition density for $\{\lambda_t\}$ can be expressed as

$$p(\lambda_t \mid \lambda_0) = \sum_{m=0}^{\infty} \mathsf{q}_t(m)\, D(\lambda_t | m, \lambda_0)$$

where

$$\mathsf{q}_t(m) = \frac{(a+b)_m\, e^{-m\,c\,t}}{m!} (1 - e^{-c\,t})^{a+b}$$

and

$$D(\lambda_t | m, \lambda_0) = \sum_{k=0}^{m} \mathsf{Be}(\lambda_t | a+k, b+m-k)\, \mathsf{Bin}(k | m, \lambda_0).$$

(M. and Walker, 2009)

Diffusion part $\{\lambda_t\}$

- Sup. we have observations (t_i, s_i), where

$$s_i \mid \lambda_i \sim \text{Geom}(\lambda_i)$$
$$(\lambda_1, \ldots, \lambda_n) \sim \text{WF}(a, b, c)$$

With the *fidis* for $\{\lambda_t\}$ given by

$$p(\lambda_1, \ldots, \lambda_n) = p(\lambda_0) \prod_{i=1}^{n} p(\lambda_i \mid \lambda_{i-1}), \qquad \text{where } \lambda_i := \lambda_{t_i}$$

and $p(\lambda_0) = \mathsf{Be}(\lambda_0 \mid a, b)$

$p(\lambda_i \mid \lambda_{i-1})$ has an infinite summation $\Rightarrow$ slice it!

$$p(\lambda_t \mid \lambda_0) = \sum_{m=0}^{\infty} \frac{g(m)}{g(m)} \mathsf{q}_t(m)\, D(\lambda_t | m, \lambda_0)$$

where g is a decreasing func. with known inverse, *e.g.* $g(m) = e^{-m}$

Diffusion part $\{\lambda_t\}$

- Augment the transition density via the latent variables $(u_i, d_i, k_i)_{i=1}^n$

$$p(\lambda_i, u_i, k_i, d_i \mid \lambda_{i-1}) =$$

$$\mathbf{1}(u_i < g(d_i))\,\frac{\mathsf{q}_i(d_i)}{g(d_i)}\,\mathsf{Be}(\lambda_i|a+k_i, b+d_i-k_i)\,\mathsf{Bin}(k_i|d_i,\lambda_{i-1})$$

Hence, the likelihood for the "complete data" is

$$l(a,b,c) = \mathrm{Beta}(\lambda_0|a,b)\prod_{i=1}^n p(\lambda_i,u_i,k_i,d_i|\lambda_{i-1})\,\lambda_i(1-\lambda_i)^{s_i-1}$$

If we assume priors for $a, b, c \overset{\text{iid}}{\sim} \mathrm{Exp}(1)$ then the posterior distributions $\pi(a \mid b, c, \ldots) \propto l(a,b,c)e^{-a}$, etc. are log-concave, *e.g.*

$$\log\pi(c|a,b,\ldots) = \sum_{i=1}^n \left\{(a+b)\log(1-e^{-c\,\tau_i}) - d_i\,c\,\tau_i\right\} - c + C,$$

Condicionales completas

$$\pi(k_i|\ldots) \propto \binom{d_i}{k_i} \frac{\mathbf{1}(k_i \in \{0,1,\ldots,d_i\})}{\Gamma(a+k_i)\Gamma(b+d_i-k_i)} \left\{ \frac{\lambda_i \lambda_{i-1}}{(1-\lambda_i)(1-\lambda_{i-1})} \right\}^{k_i}$$

easy to sample as it takes a finite number of values

$$\pi(u_i \mid \ldots) = \mathrm{U}_{[0,g(d_i)]}(u_i)$$

$$\pi(d_i|\ldots) \propto \frac{\Gamma(a+d+d_i)^2 e^{d_i[1-c\tau_i]}\, \mathbf{1}(k_i \le d_i \le -\log u_i)}{\Gamma(b+d_i-k_i)\Gamma(d_i-k_i+1)\{(1-\lambda_{i-1})(1-\lambda_i)\}^{-d_i}}$$

Also finite due to the u_i's

Complete conditionals

The complete conditionals for λ_i, $i \neq 0, n$, are given by

$$\pi(\lambda_i | \ldots) = \mathrm{Beta}(1 + a + k_i + k_{i+1}, s_i - 1 + b + d_i + d_{i+1} - k_i - k_{i+1}),$$

and

$$\pi(\lambda_0 | \ldots) = \mathrm{Beta}(a + k_1, b + d_1 - k_1)$$

and

$$\pi(\lambda_n | \ldots) = \mathrm{Beta}(1 + a + k_n, s_n - 1 + b + d_n - k_n).$$

This procedure via the latent variables could also be useful to estimate other diffusion processes

Gibbs sampler

For the remaining part of the model we use a similar idea "slice"

- That is, we "augment" the model

$$y_i|t_i,\lambda_i,\theta \sim \sum_{l=1}^{\infty} \lambda_i(1-\lambda_i)^{l-1} f(y_i|\theta_l),$$

with two random variables (s_i, v_i) and $\{\psi_l\}$ (a seq. of decreasing numbers s.t. $\{l : \psi_l > v\}$ is a known set), i.e.

$$y_i, v_i, s_i|\lambda_i,\theta \sim \psi_{s_i}^{-1}\mathbf{1}(v_i < \psi_{s_i})\,\lambda_i(1-\lambda_i)^{s_i-1}\,f(y_i|\theta_{s_i}).$$

In this way

$$\begin{aligned}
\pi(s_i|\ldots) &\propto \psi_{s_i}^{-1}\lambda_i(1-\lambda_i)^{s_i-1}\,f(y_i|\theta_{s_i})\,\mathbf{1}(s_i \in \{l : \psi_l > v_i\}) \\
\pi(v_i|\ldots) &= \mathrm{U}_{(0,\psi_{s_i})}(v_i) \\
\pi(\theta_l|\ldots) &\propto \prod_{s_i=l} f(y_i|\theta_l)\,g_0(\theta_l) \qquad \text{for } l = 1,\ldots,\max_i\{l : \psi_l > v_i\}
\end{aligned}$$

Gibbs sampler

Summarizing, we need

- $\pi(a \mid b, c, \ldots)$, $\pi(b \mid a, c, \ldots)$ and $\pi(c \mid a, b, \ldots)$ (via ARS)
- $\pi(k_i| \ldots)$, $\pi(u_i| \ldots)$ y $\pi(d_i| \ldots)$ (via Inverse CDF)
- $\pi(\lambda_i| \ldots)$ (Beta's)
- $\pi(s_i| \ldots)$ and $\pi(v_i| \ldots)$ (via Inverse CDF)
- $\pi(\theta_i| \ldots)$ (if f y g_0 are conjugated ✓, otherwise via ARMS, M-H, etc)

A bit long, but only a very simple Gibbs sampler

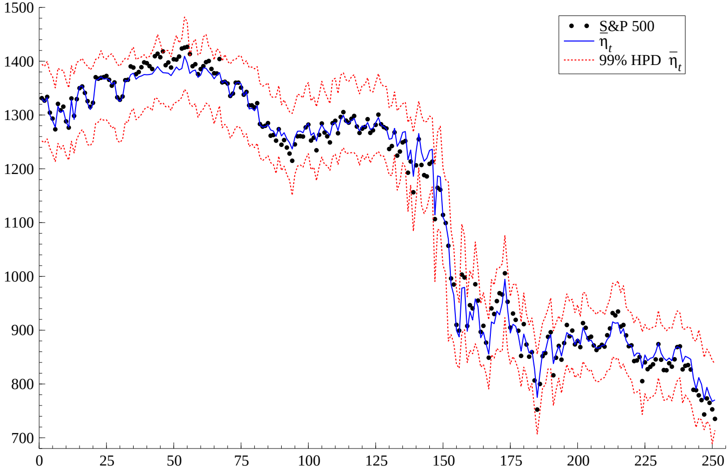


Figura: MC estimator for $\bar{\eta}_t$ (solid) and corresponding 99 % highest posterior density intervals (dotted) for the S&P 500 data set (dots). The estimates are based on 10000 iterations of the Gibbs sampler algorithm after 2000 iterations of burn in.

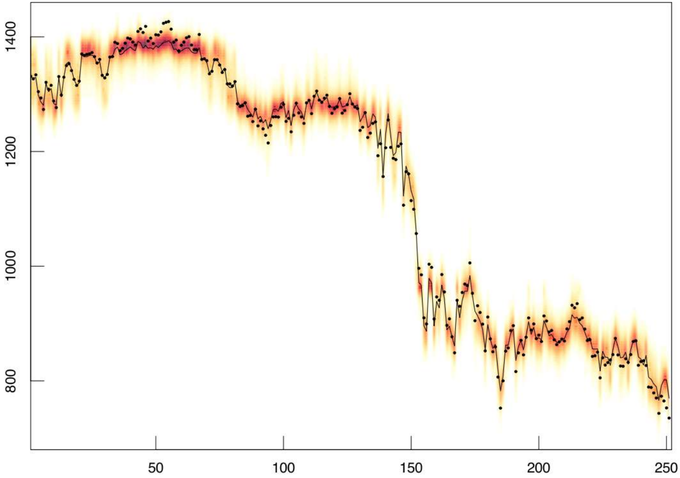


Figura: MCMC density estimator for the random density process, $\hat{f}_t$, (heat contour), mean of mean functional $\bar{\eta}_t$ (solid) for the S&P 500 data set (dots). The estimates are based on 10000 effective

EPPF

$$\Pi_k^n(n_1, n_2, \ldots, n_k) = \left(\frac{\lambda}{1-\lambda}\right)^n \sum_{(*)_k} (1-\lambda)^{\sum_{l=1}^k n_l j_l}$$

Then, one can obtains results such as

$$\mathsf{E}[K_n] = \sum_{r=1}^{n} (-1)^{r-1} \binom{n}{r} \frac{\lambda^r}{1-(1-\lambda)^r}$$

when k is large

$$\Pi_k^n(n_1, n_2, \ldots, n_k) \approx \left(\frac{\lambda}{1-\lambda}\right)^n (1-\lambda)^{n_{(1)}+2n_{(2)}+\cdots+kn_{(k)}}$$

(M. and Walker, 2012)

Thanks !

References

ANTONIAK, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, **2**, 1152–1174.

ESCOBAR, M.D. AND WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Stat. Assoc.*, **90**, 577–588.

EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, **3**, 87-112.

FENG, S. (2010). *The Poisson-Dirichlet Distribution and Related Topics: Models and Asymptotic Behaviors.* Springer.

FERGUSON, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

FUENTES-GARCÍA, R., MENA, R. H. AND WALKER, S. G. (2009). A nonparametric dependent process for Bayesian regression. *Statistics and Probability Letters*. **79**, 1112–1119.

FUENTES-GARCÍA, R., MENA, R. H. AND WALKER, S. G. (2010). A new Bayesian nonparametric mixture model. *Communications in Statistics-Simulation and Computation*. **39**, 669–682.

FUENTES-GARCÍA, R., MENA, R. H. AND WALKER, S. G. (2010). A probability for classification based on the mixture of Dirichlet process model. *Journal of Classification*. **In press**.

ISHWARAN, H. AND JAMES, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Stat. Assoc.*, **96**, 161–173.

MENA, R.H. AND WALKER, S.G. (2009). On a construction of Markov models in continuous time. *Metron*, **67**, 303-323.

MENA, R.H. AND WALKER, S.G. (2012). An EPPF from independent sequences of geometric random variables. *Statistics and Probability Letters*. To appear.

MENA, R.H., RUGGIERO, M. AND WALKER, S.G. (2011). Geometric stick-breaking processes for continuous-time Bayesian nonparametric modeling. *Journal of Statistical Planning and Inference*, **141**, 3217-3230.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica*, **4**, 639–650.

WALKER, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics*, **36**, 45–54.