Advanced Bayesian Learning

Lecture 3 - Dirichlet Process Priors

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

Topic overview

- Reminder: **Multinomial data** - **Dirichlet prior**
- **Bayesian histograms**
- The **Dirichlet process**
- Beyond DP: **Pitman-Yor** and **Probit stick-breaking**
- **Dirichlet process mixtures**
- **MCMC** for Dirichlet process mixtures
- **Dependent Dirichlet Process** constructions

The Dirichlet distribution

- $\theta \sim \mathrm{Dirichlet}(a_1, ..., a_k)$ with density

$$p(\theta_1, \theta_2, ..., \theta_k) \propto \prod_{j=1}^{k} \theta_j^{a_j - 1}.$$

- Define $\alpha = \sum_{j=1}^{k} a_j$ and $\pi_0 = a/\alpha$.
- **Expected value** and **variance** of $\mathrm{Dirichlet}(a_1, ..., a_k)$

$$\mathrm{E}(\theta_j) = \frac{a_j}{\alpha} = \pi_{0j} \qquad \mathrm{V}(\theta_j) = \frac{\mathrm{E}(\theta_j)\left[1 - \mathrm{E}(\theta_j)\right]}{1 + \alpha}$$

- Note that α is a **precision** parameter (large α, low variance).

Conjugate analysis for multinomial data

- **Data**: $y = (n_1, ..., n_k)$, where $n_j = \#$ items in category j.
- **Prior**

$$\theta \sim \mathrm{Dirichlet}(a_1, ..., a_k)$$

- **Likelihood**

$$p(n_1, n_2, ..., n_k | \theta_1, \theta_2, ..., \theta_k) \propto \prod_{j=1}^{k} \theta_j^{n_j}$$

- **Posterior**

$$\theta | n_1, ..., n_k \sim \mathrm{Dirichlet}(n_1 + a_1, ..., n_k + a_k)$$

- **Posterior mean**

$$E(\theta_j | n_1, ..., n_k) = \frac{n_j + a_j}{n + \alpha}$$

Bayesian histograms

- **Partition** the data space $\xi_0 < \xi_1 < ... < \xi_k$ in k **bins** B_h.
- Probability model for **histograms**

$$f(y) = \sum_{h=1}^{k} 1_{\xi_{h-1} < y \leq \xi_h} \frac{\pi_h}{(\xi_h - \xi_{h-1})}$$

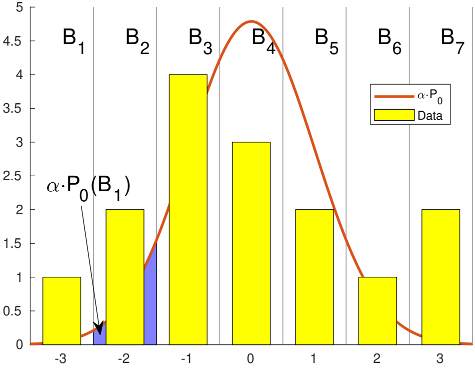
- $n_h =$ number of obs in B_h: $\xi_{h-1} < y \leq \xi_h$. Multinomial.
- **Prior** on $\pi = (\pi_1, ..., \pi_k)$

$$\pi \sim Dirichlet(a_1, ..., a_k)$$

- **Posterior**

$$\pi | n_1, ..., n_k \sim \mathrm{Dirichlet}(n_1 + a_1, ..., n_k + a_k)$$

Illustration of Bayesian histograms



Bayesian histograms

- Posterior

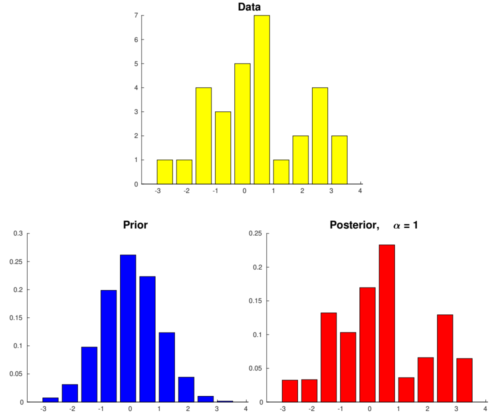
$$\pi | n_1, ..., n_k \sim \text{Dirichlet}(n_1 + a_1, ..., n_k + a_k)$$

- Specify $a_1, .., a_k$ through $\pi_0 = (\pi_{01}, ..., \pi_{0k})$ and $\alpha = \sum_{j=1}^k a_j$.
- Specify π_0 from a base distribution P_0. For the hth bin:

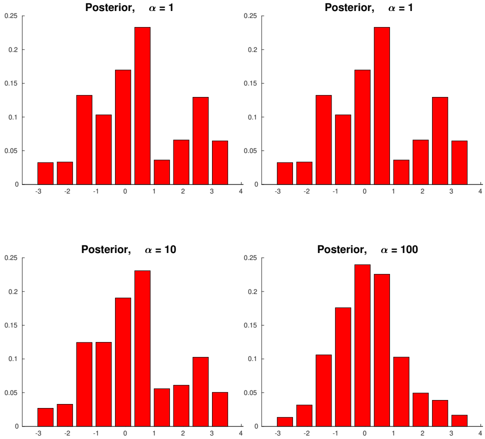
$$\pi_{0h} = P_0(B_h) = \Pr(\xi_{h-1} < y \leq \xi_h).$$

- Properties of Dirichlet prior:
  - **easy computations** ☺
  - **easy to specify hyperparameter** π_0 and α. ☺
  - **no smoothness**: adjacent bin are negatively correlated. ☹
  - **sensitive** to the choice of **bins**. ☹

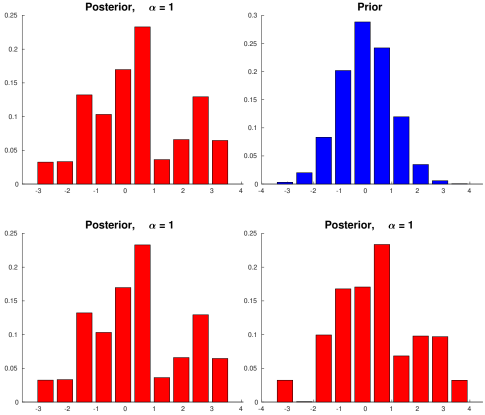
Bayesian histogram example



Larger α give higher weight to prior



Histograms are sensitive to the choice of bins



The Dirichlet process

- Let $B_1, B_2, ..., B_k$ be a partition of the outcome space Ω.
- $P(B_1), ..., P(B_k)$ denotes the distribution over the partition.
- Dirichlet distribution is a **distribution over distributions**:

$$(P(B_1), ..., P(B_k)) \sim \mathrm{Dirichlet}(\alpha P_0(B_1), ..., \alpha P_0(B_k))$$

  where P_0 is a fixed probability measure (e.g. $N(0, 1)$).

- Dirichlet is **closed under summation or splitting** of bins.
  $\Rightarrow$ consistent definition of a **stochastic process**. c.f. GPs.

- A random probability measure P follows a **Dirichlet process** $P \sim \mathrm{DP}(\alpha \cdot P_0)$ with **base measure** P_0 iff

$$(P(B_1), ..., P(B_k)) \sim \mathrm{Dirichlet}(\alpha P_0(B_1), ..., \alpha P_0(B_k))$$

  for any finite measureable partition $B_1, ..., B_k$.

The Dirichlet process - properties

- If $P \sim \mathrm{DP}(\alpha P_0)$ then

$$P(B) \sim \mathrm{Beta}\left[\alpha P_0(B), \alpha\left(1 - P_0(B)\right)\right], \text{ for any } B \in \mathcal{B}$$

$$E\left[P(B)\right] = P_0(B)$$

$$\mathrm{Var}\left[P(B)\right] = P_0(B)\left[1 - P_0(B)\right]/(1+\alpha)$$

- **Model**

$$y_i | P \overset{iid}{\sim} P \text{ , for } i = 1, \ldots, n$$

- **Prior**

$$P \sim \mathrm{DP}(\alpha P_0)$$

- **Posterior** for a **finite partition**, $P(B_1), \ldots, P(B_k)|\mathbf{y}$ is

$$\mathrm{Dirichlet}\left(\alpha P_0(B_1) + \sum_{i=1}^{n} 1_{y_i \in B_1}, \ldots, \alpha P_0(B_k) + \sum_{i=1}^{n} 1_{y_i \in B_k}\right)$$

The Dirichlet process - properties

- **Posterior** for the unknown probability distribution P

$$P|y_1, ..., y_n \sim \mathrm{DP}\left(\alpha P_0 + \sum_{i=1}^{n} \delta_{y_i}\right)$$

- Since

$$P(B) \sim \mathrm{Beta}\left(\alpha P_0(B) + \sum_{i=1}^{n} 1_{y_i \in B}, \alpha(1 - P_0(B)) + \sum_{i=1}^{n} 1_{y_i \in B^c}\right)$$

so

$$E\left(P(B)|y_1, ..., y_n\right) = \left(\frac{\alpha}{\alpha + n}\right) P_0(B) + \left(\frac{n}{\alpha + n}\right) \sum_{i=1}^{n} \frac{1}{n} \delta_{y_i}(B)$$

Estimating a distribution function with a DP prior

- If $B = (-\infty, y]$ then

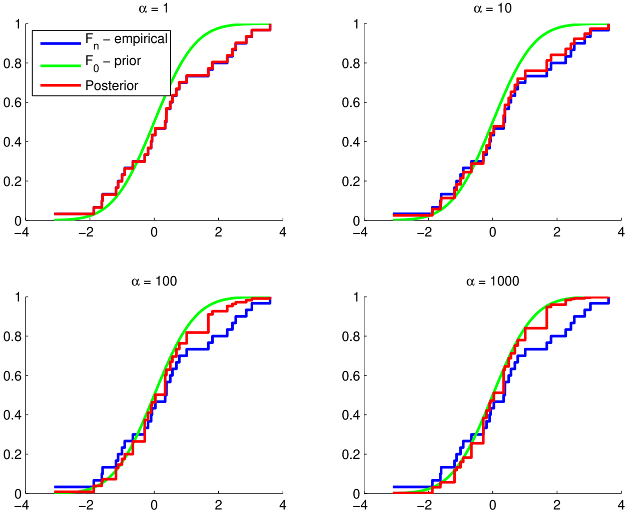
$$E\left(F(y)|y_1, ..., y_n\right) = \left(\frac{\alpha}{\alpha + n}\right) F_0(y) + \left(\frac{n}{\alpha + n}\right) F_n(y)$$

  - $F(y)$ is the unknown d.f.
  - $F_0(y)$ is the d.f. from P_0
  - $F_n(y) = \frac{1}{n} \sum 1_{y_i \leq y}$ is the empirical d.f.

- $F(\cdot)$ is **discrete with probability one** in the DP posterior.
- **Realisations from a DP are discrete with probability one**.

  - Clearly a bad property for continuous data ...
  - But very useful for clustering (mixture models).

Estimating a distribution function with a DP prior



Stick-breaking characterization of the DP

- $P \sim DP(\alpha P_0) \equiv$ infinite mixture of point masses

$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_i}$$

$$\pi_h = V_h \prod_{\ell < h} (1 - V_\ell)$$

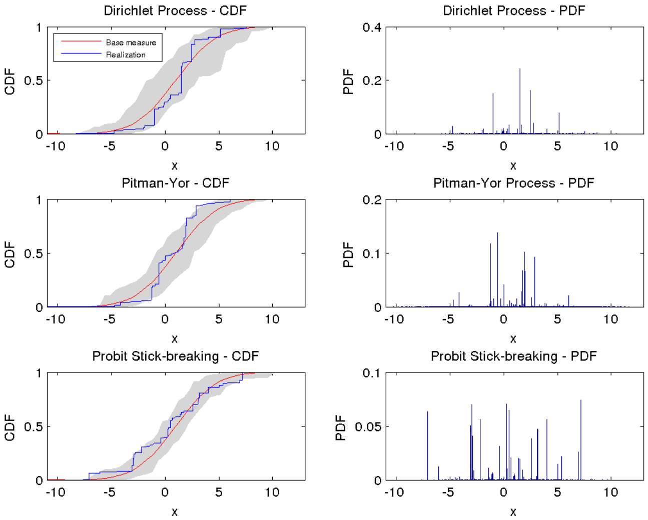
$$V_h \overset{iid}{\sim} Beta(1, \alpha)$$

$$\theta_h \overset{iid}{\sim} P_0$$

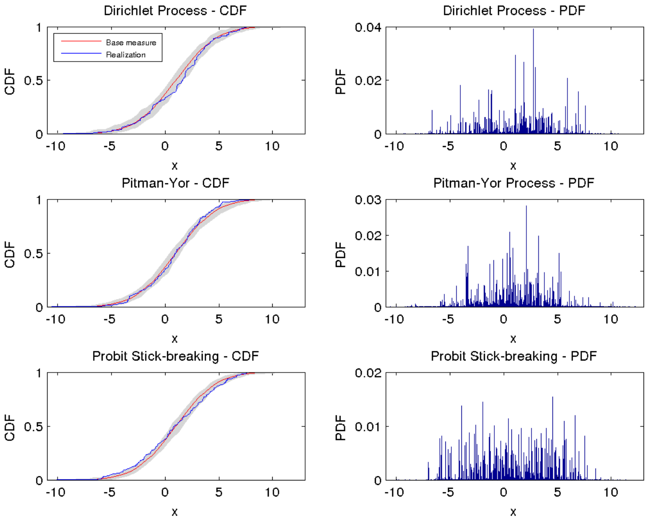
- Alternative notation for $P \sim DP(\alpha P_0)$:

$$\pi = (\pi_1, \pi_2, ...) \sim \text{Stick}(\alpha) \text{ and } \theta_h \sim P_0$$

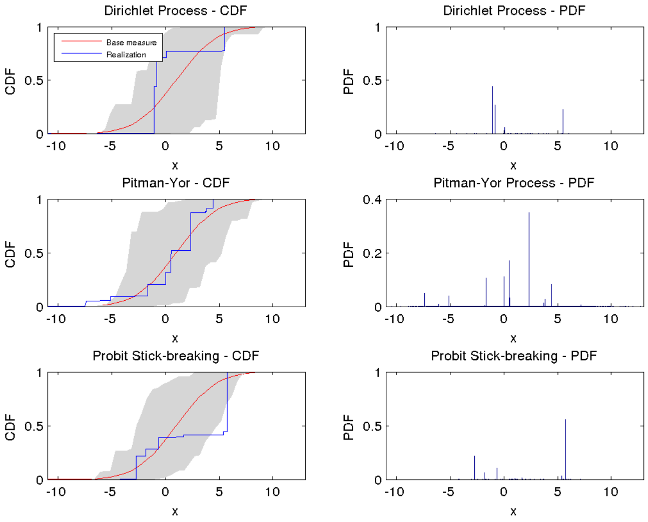
Simulating stick-breaking $\alpha = 10, P_0 = N(1, 3^2)$



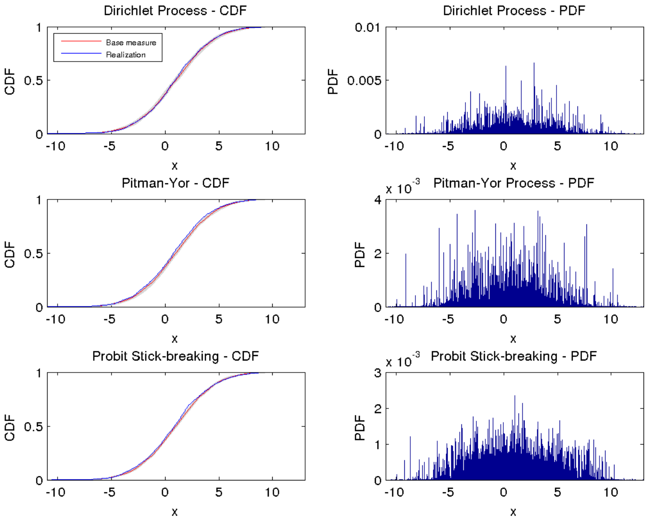
Simulating stick-breaking $\alpha = 100,\ P_0 = N(1, 3^2)$



Simulating stick-breaking $\alpha = 1, P_0 = N(1, 3^2)$



Simulating stick-breaking $\alpha = 1000,\ P_0 = N(1, 3^2)$



Beyond DP - Pitman-Yor and Probit sticks

- **Pitman-Yor process** with parameters P_0, $0 \leq a < 1$ and $b > -a$:

$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_i} \quad \theta_h \overset{iid}{\sim} P_0$$

$$\pi_h = V_h \prod_{\ell < h} (1 - V_\ell)$$

$$V_h \overset{iid}{\sim} Beta(1 - a, b + ha)$$

- **Probit stick-breaking** with parameters μ and σ:

$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_i} \quad \theta_h \overset{iid}{\sim} P_0$$

$$\pi_h = V_h \prod_{\ell < h} (1 - V_\ell)$$

$$V_h = \Phi(x_h), \ \ \text{where } x_h \overset{iid}{\sim} N(\mu, \sigma^2)$$