An Introduction to Non-parametric Bayesian Statistics

Reza Solgi

reza.solgi@usi.ch

May 2, 2012

Introduction

- In a parametric model, we fix the complexity of the model ($\equiv \#$ parameters) *a priori*, so the data itself would not speak about the degree of model complexity needed to fit the real data adequately.

Introduction

- In a parametric model, we fix the complexity of the model ($\equiv \#$ parameters) *a priori*, so the data itself would not speak about the degree of model complexity needed to fit the real data adequately.
- Bayesian non-parametric models provide a flexible modeling setting in which the model itself can infer about the necessary amount of flexibility for the data in hand.

Introduction

- In a parametric model, we fix the complexity of the model ($\equiv \#$ parameters) *a priori*, so the data itself would not speak about the degree of model complexity needed to fit the real data adequately.
- Bayesian non-parametric models provide a flexible modeling setting in which the model itself can infer about the necessary amount of flexibility for the data in hand.
- In fact a non-parametric model can be thought of the limit of a parametric model with a finite number of parameters, when we take the number of parameters to infinity.

Introduction

- In a parametric model, we fix the complexity of the model ($\equiv \#$ parameters) *a priori*, so the data itself would not speak about the degree of model complexity needed to fit the real data adequately.
- Bayesian non-parametric models provide a flexible modeling setting in which the model itself can infer about the necessary amount of flexibility for the data in hand.
- In fact a non-parametric model can be thought of the limit of a parametric model with a finite number of parameters, when we take the number of parameters to infinity.
- A suitable non-parametric model should have a large support; Ideally the support of the prior should be dense in the set of all plausible models. Moreover the hyperparameters of the model should be interpretable; so we can assign reasonable values to them, and last but not least, the model should be computationally tractable.

Example 1: Density Estimation

Consider a univariate density estimation problem:

$$x_i \overset{ind.}{\sim} F$$

where $x_i \in \mathbb{R}$, and $F \in \mathcal{F}$, where $\mathcal{F}$ is the family of all continuous distributions on $\mathbb{R}$.

Example 1: Density Estimation

Consider a univariate density estimation problem:

$$x_i \overset{ind.}{\sim} F$$

where $x_i \in \mathbb{R}$, and $F \in \mathcal{F}$, where $\mathcal{F}$ is the family of all continuous distributions on $\mathbb{R}$.

In a **Parametric Modelling Setting**:
We should first restrict the family of distribution to a family of parametrized distribution, say Normal distributions:

$$\begin{aligned} \mathcal{F}_{\text{param}} &:= \{\text{all Normal distributions with mean } \mu \in \mathbb{R}, \text{ and variance } \sigma^2 \in \mathbb{R}_+\} \\ &\subset \mathcal{F} \end{aligned}$$

and we put some priors on μ and σ^2.

Example 1: Density Estimation

Consider a univariate density estimation problem:

$$x_i \overset{ind.}{\sim} F$$

where $x_i \in \mathbb{R}$, and $F \in \mathcal{F}$, where $\mathcal{F}$ is the family of all continuous distributions on $\mathbb{R}$.

In a **Parametric Modelling Setting**:
We should first restrict the family of distribution to a family of parametrized distribution, say Normal distributions:

$$\begin{aligned} \mathcal{F}_{\text{param}} &:= \{\text{all Normal distributions with mean } \mu \in \mathbb{R}, \text{ and variance } \sigma^2 \in \mathbb{R}_+\} \\ &\subset \mathcal{F} \end{aligned}$$

and we put some priors on μ and σ^2.

So a priori we decide about the number of parameters and impose crucial restrictions on the model (in this case: symmetry, unimodality, light-tails, ...).

Example 1: Density Estimation

Consider a univariate density estimation problem:

$$x_i \overset{ind.}{\sim} F$$

where $x_i \in \mathbb{R}$, and $F \in \mathcal{F}$, where $\mathcal{F}$ is the family of all continuous distributions on $\mathbb{R}$.

In a **Parametric Modelling Setting**:
We should first restrict the family of distribution to a family of parametrized distribution, say Normal distributions:

$$\begin{aligned} \mathcal{F}_{\text{param}} &:= \{\text{all Normal distributions with mean } \mu \in \mathbb{R}, \text{ and variance } \sigma^2 \in \mathbb{R}_+\} \\ &\subset \mathcal{F} \end{aligned}$$

and we put some priors on μ and σ^2.

So a priori we decide about the number of parameters and impose crucial restrictions on the model (in this case: symmetry, unimodality, light-tails, ...).

Note that support of $\mathcal{F}_{\text{param}}$ is dramatically restricted w.r.t. the support of $\mathcal{F}$. In fact $\mathcal{F}_{\text{param}}$ is not dense in $\mathcal{F}$, in the sense that for any $F \in \mathcal{F}$ we can not find a member of $\mathcal{F}_{\text{param}}$ sufficiently close to F.

Example 1: Density Estimation

Consider a univariate density estimation problem:

$$x_i \overset{ind.}{\sim} F$$

where $x_i \in \mathbb{R}$, and $F \in \mathcal{F}$, where $\mathcal{F}$ is the family of all continuous distributions on $\mathbb{R}$.

In a **Parametric Modelling Setting**:
We should first restrict the family of distribution to a family of parametrized distribution, say Normal distributions:

$$\begin{aligned} \mathcal{F}_{\text{param}} &:= \{\text{all Normal distributions with mean } \mu \in \mathbb{R}, \text{ and variance } \sigma^2 \in \mathbb{R}_+\} \\ &\subset \mathcal{F} \end{aligned}$$

and we put some priors on μ and σ^2.

So a priori we decide about the number of parameters and impose crucial restrictions on the model (in this case: symmetry, unimodality, light-tails, ...).

Note that support of $\mathcal{F}_{\text{param}}$ is dramatically restricted w.r.t. the support of $\mathcal{F}$. In fact $\mathcal{F}_{\text{param}}$ is not dense in $\mathcal{F}$, in the sense that for any $F \in \mathcal{F}$ we can not find a member of $\mathcal{F}_{\text{param}}$ sufficiently close to F.

$\mathcal{F}$ is an infinite dimensional space, while $\mathcal{F}_{\text{param}}$ is just a two dimensional subset of it.

Example 1: Density Estimation

A parametric model is not able to model all features of the data.

Example 1: Density Estimation

A parametric model is not able to model all features of the data.

Blind Men and The Elephant

Example 1: Density Estimation

In a non-parametric Bayesian approach to this problem, we do not assume any specific type for the unknown distribution F: We treat F as the unknown parameter of the model.

Example 1: Density Estimation

In a non-parametric Bayesian approach to this problem, we do not assume any specific type for the unknown distribution F: We treat F as the unknown parameter of the model.

Therefore a non-parametric Bayesian model is not a model without any parameter! In fact the term non-parametric Bayesian modelling is an oxymoron and a misnomer: A non-parametric Bayesian model has infinitely many parameters.

Example 1: Density Estimation

In a non-parametric Bayesian approach to this problem, we do not assume any specific type for the unknown distribution F: We treat F as the unknown parameter of the model.

Therefore a non-parametric Bayesian model is not a model without any parameter! In fact the term non-parametric Bayesian modelling is an oxymoron and a misnomer: A non-parametric Bayesian model has infinitely many parameters.

Bayesian non-parametric models are models on function spaces.

Example 1: Density Estimation

In a non-parametric Bayesian approach to this problem, we do not assume any specific type for the unknown distribution F: We treat F as the unknown parameter of the model.

Therefore a non-parametric Bayesian model is not a model without any parameter! In fact the term non-parametric Bayesian modelling is an oxymoron and a misnomer: A non-parametric Bayesian model has infinitely many parameters.

Bayesian non-parametric models are models on function spaces.

In the density estimation example F can be seen as the parameter of the model: F is an infinite dimensional object. So roughly speaking we can say our non-parametric model has infinitely many parameters.

Example 1: Density Estimation

In a non-parametric Bayesian approach to this problem, we do not assume any specific type for the unknown distribution F: We treat F as the unknown parameter of the model.

Therefore a non-parametric Bayesian model is not a model without any parameter! In fact the term non-parametric Bayesian modelling is an oxymoron and a misnomer: A non-parametric Bayesian model has infinitely many parameters.

Bayesian non-parametric models are models on function spaces.

In the density estimation example F can be seen as the parameter of the model: F is an infinite dimensional object. So roughly speaking we can say our non-parametric model has infinitely many parameters.

In a Bayesian model, each parameter is equipped with a prior distribution. In this example the model's parameter is the unknown distribution F, and therefore we need to assign a prior distribution to the distributions...

Example 1: Density Estimation

In a non-parametric Bayesian approach to this problem, we do not assume any specific type for the unknown distribution F: We treat F as the unknown parameter of the model.

Therefore a non-parametric Bayesian model is not a model without any parameter! In fact the term non-parametric Bayesian modelling is an oxymoron and a misnomer: A non-parametric Bayesian model has infinitely many parameters.

Bayesian non-parametric models are models on function spaces.

In the density estimation example F can be seen as the parameter of the model: F is an infinite dimensional object. So roughly speaking we can say our non-parametric model has infinitely many parameters.

In a Bayesian model, each parameter is equipped with a prior distribution. In this example the model's parameter is the unknown distribution F, and therefore we need to assign a prior distribution to the distributions...

A Distribution on Distributions?

Example 1: Density Estimation

In a non-parametric Bayesian approach to this problem, we do not assume any specific type for the unknown distribution F: We treat F as the unknown parameter of the model.

Therefore a non-parametric Bayesian model is not a model without any parameter! In fact the term non-parametric Bayesian modelling is an oxymoron and a misnomer: A non-parametric Bayesian model has infinitely many parameters.

Bayesian non-parametric models are models on function spaces.

In the density estimation example F can be seen as the parameter of the model: F is an infinite dimensional object. So roughly speaking we can say our non-parametric model has infinitely many parameters.

In a Bayesian model, each parameter is equipped with a prior distribution. In this example the model's parameter is the unknown distribution F, and therefore we need to assign a prior distribution to the distributions...

A Distribution on Distributions?

This is where the Dirichlet Process comes into play!

Example 2: Nonlinear Regression

Consider the nonlinear regression problem:

$$y_i = f(x_i) + \varepsilon_i$$

Example 2: Nonlinear Regression

Consider the nonlinear regression problem:

$$y_i = f(x_i) + \varepsilon_i$$

where f is a non-linear function, and ε_i is i.i.d. from $\mathcal{N}(0, \sigma_\varepsilon^2)$. For simplicity we assume both x and y belong to $\mathbb{R}$, and therefore:

$$f : \mathbb{R} \to \mathbb{R}$$

Example 2: Nonlinear Regression

Consider the nonlinear regression problem:

$$y_i \quad = \quad f(x_i) + \varepsilon_i$$

where f is a non-linear function, and ε_i is i.i.d. from $\mathcal{N}(0, \sigma_\varepsilon^2)$. For simplicity we assume both x and y belong to $\mathbb{R}$, and therefore:

$$f \quad : \quad \mathbb{R} \rightarrow \mathbb{R}$$

In a parametric approach to this problem, we should parametrize this unknown function using a vector of parameters θ:

$$f \quad \equiv \quad f(\theta)$$

For instance, we can use a linear/polynomial function, in which θ is the vector of coefficients of the linear/polynomial function.

Example 2: Nonlinear Regression

Then we put a prior on the vector of parameters:

$$\theta \quad \sim \quad p(\theta)$$

Example 2: Nonlinear Regression

Then we put a prior on the vector of parameters:

$$\theta \quad \sim \quad p(\theta)$$

and find its posterior:

$$p(\theta|\mathcal{D}) \quad \propto \quad p(\theta)\ p(\mathcal{D}|\theta)$$

where $\mathcal{D} = (\mathbf{x}, \mathbf{y})$ is the observed data.

Example 2: Nonlinear Regression

Then we put a prior on the vector of parameters:

$$\theta \quad \sim \quad p(\theta)$$

and find its posterior:

$$p(\theta|\mathcal{D}) \quad \propto \quad p(\theta)\ p(\mathcal{D}|\theta)$$

where $\mathcal{D} = (\mathbf{x}, \mathbf{y})$ is the observed data.

How we can decide about the parametric form of f? Is a linear form adequate? or we need a higher degree polynomial? How can we avoid under-fitting/over-fitting? How can we let the data itself "*speak*" about the unknown function f?

Example 2: Nonlinear Regression

Then we put a prior on the vector of parameters:

$$\theta \sim p(\theta)$$

and find its posterior:

$$p(\theta|\mathcal{D}) \propto p(\theta)\ p(\mathcal{D}|\theta)$$

where $\mathcal{D} = (\mathbf{x}, \mathbf{y})$ is the observed data.

How we can decide about the parametric form of f? Is a linear form adequate? or we need a higher degree polynomial? How can we avoid under-fitting/over-fitting? How can we let the data itself "*speak*" about the unknown function f?

In a Bayesian non-parametric approach to this problem, we assume the function f is the unknown parameter of the model. Therefore we need to specify a prior on it...

Example 2: Nonlinear Regression

Then we put a prior on the vector of parameters:

$$\theta \quad \sim \quad p(\theta)$$

and find its posterior:

$$p(\theta|\mathcal{D}) \quad \propto \quad p(\theta)\ p(\mathcal{D}|\theta)$$

where $\mathcal{D} = (\mathbf{x}, \mathbf{y})$ is the observed data.

How we can decide about the parametric form of f? Is a linear form adequate? or we need a higher degree polynomial? How can we avoid under-fitting/over-fitting? How can we let the data itself "*speak*" about the unknown function f?

In a Bayesian non-parametric approach to this problem, we assume the function f is the unknown parameter of the model. Therefore we need to specify a prior on it...

A Distribution on Functions?

Example 2: Nonlinear Regression

Then we put a prior on the vector of parameters:

$$\theta \quad \sim \quad p(\theta)$$

and find its posterior:

$$p(\theta|\mathcal{D}) \quad \propto \quad p(\theta)\ p(\mathcal{D}|\theta)$$

where $\mathcal{D} = (\mathbf{x}, \mathbf{y})$ is the observed data.

How we can decide about the parametric form of f? Is a linear form adequate? or we need a higher degree polynomial? How can we avoid under-fitting/over-fitting? How can we let the data itself "*speak*" about the unknown function f?

In a Bayesian non-parametric approach to this problem, we assume the function f is the unknown parameter of the model. Therefore we need to specify a prior on it...

A Distribution on Functions?

This is where the Gaussian Process comes into play!

Tutorial's Outline

In this tutorial the following subjects will be covered:

- **Gaussian Process**
- **Dirichlet Process**
- **Dirichlet Process Mixture**

Remembering Some Definitions...

Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (where Ω is the sample space, $\mathcal{F}$ is a σ-algebra, and $\mathbb{P}$ is a probability measure).

Remembering Some Definitions...

Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (where Ω is the sample space, $\mathcal{F}$ is a σ-algebra, and $\mathbb{P}$ is a probability measure).

Random Variable

Assume $(E, \mathcal{E})$ is a measurable space, and X is an E-valued function defined on Ω:

$$X : \Omega \to E$$

We call X a random variable on $(\Omega, \mathcal{F})$, if:

$$\sigma(X) := \{X^{-1}(A);\ A \in \mathcal{E}\} \quad \subset \quad \mathcal{F}$$

Remembering Some Definitions...

Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (where Ω is the sample space, $\mathcal{F}$ is a σ-algebra, and $\mathbb{P}$ is a probability measure).

Random Variable

Assume $(E, \mathcal{E})$ is a measurable space, and X is an E-valued function defined on Ω:

$$X : \Omega \to E$$

We call X a random variable on $(\Omega, \mathcal{F})$, if:

$$\sigma(X) := \{X^{-1}(A);\ A \in \mathcal{E}\} \quad \subset \quad \mathcal{F}$$

Note that by this definition, for any $A \in \mathcal{E}$, the following probability:

$$\mathbb{P}(X \in A) := \mathbb{P}(\{\omega; X(\omega) \in A\})$$

is well-defined, because $\{\omega; X(\omega) \in A\} = X^{-1}(A)$ lies in the σ-algebra of our assumed probability space $\mathcal{F}$.

Remembering Some Definitions...

Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (where Ω is the sample space, $\mathcal{F}$ is a σ-algebra, and $\mathbb{P}$ is a probability measure).

Random Variable

Assume $(E, \mathcal{E})$ is a measurable space, and X is an E-valued function defined on Ω:

$$X : \Omega \to E$$

We call X a random variable on $(\Omega, \mathcal{F})$, if:

$$\sigma(X) := \{X^{-1}(A);\ A \in \mathcal{E}\} \quad \subset \quad \mathcal{F}$$

Note that by this definition, for any $A \in \mathcal{E}$, the following probability:

$$\mathbb{P}(X \in A) := \mathbb{P}(\{\omega; X(\omega) \in A\})$$

is well-defined, because $\{\omega; X(\omega) \in A\} = X^{-1}(A)$ lies in the σ-algebra of our assumed probability space $\mathcal{F}$.

If we assume $E = \mathbb{R}$, we can use the Borel set of $\mathbb{R}$ in the place of $\mathcal{E}$, and the resulting random variable is called a real-valued random variable.

Remembering Some Definitions...

Stochastic Process

A stochastic Process $X = (X_s)_{s \in S}$ is a collection of random variables X_s indexed by a variable $s \in S$.

Remembering Some Definitions...

Stochastic Process

A stochastic Process $X = (X_s)_{s \in S}$ is a collection of random variables X_s indexed by a variable $s \in S$.

The index set S could be finite, countable, or uncountable.

Remembering Some Definitions...

Stochastic Process

A stochastic Process $X = (X_s)_{s \in S}$ is a collection of random variables X_s indexed by a variable $s \in S$.

The index set S could be finite, countable, or uncountable.

Examples of S: $\{1, 2...., N\}$, $\mathbb{N}$, $[0, T]$, $\mathbb{R}$.

Remembering Some Definitions...

Stochastic Process

A stochastic Process $X = (X_s)_{s \in S}$ is a collection of random variables X_s indexed by a variable $s \in S$.

The index set S could be finite, countable, or uncountable.

Examples of S: $\{1, 2...., N\}$, $\mathbb{N}$, $[0, T]$, $\mathbb{R}$.

FDD of a Stochastic Process

Assume $X = (X_s)_{s \in S}$ is a stochastic process, and t_1, ..., $t_n \in S$ for some n. The probability distribution function of the vector $(X_{t_1}, ..., X_{t_n})$ is called the finite dimensional distribution (FDD) of the stochastic process $(X_s)_{s \in S}$ for $\{t_1, ..., t_n\}$

Kolmogorov Existence Theorem

Consistent Collection of Measures

Assume $(E, \mathcal{E})$ is a measurable space, and let S be a set, t_1, ..., $t_n \in S$, and assume $\mu_{t_1,\ldots,t_n}$ is a measure on $(E^n, \mathcal{E}^n)$. This collection of measures is called consistent iff for all natural n, for all t_1, ..., t_n, t_{n+1}, every permutation π of $\{1, ..., n\}$, and for all A_1, ..., $A_n \in \mathcal{E}$, we have:

$$\begin{aligned} \mu_{t_1,\ldots,t_n}(A_1 \times \cdots \times A_n) &= \mu_{\pi_1,\ldots,\pi_n}(A_{\pi_1} \times \cdots \times A_{\pi_n}) \\ \mu_{t_1,\ldots,t_n,t_{n+1}}(A_1 \times \cdots \times A_n \times E) &= \mu_{t_1,\ldots,t_n}(A_1 \times \cdots \times A_n) \end{aligned}$$

Kolmogorov Existence Theorem

Consistent Collection of Measures

Assume $(E, \mathcal{E})$ is a measurable space, and let S be a set, t_1, ..., $t_n \in S$, and assume $\mu_{t_1,\ldots,t_n}$ is a measure on $(E^n, \mathcal{E}^n)$. This collection of measures is called consistent iff for all natural n, for all t_1, ..., t_n, t_{n+1}, every permutation π of $\{1, ..., n\}$, and for all A_1, ..., $A_n \in \mathcal{E}$, we have:

$$\begin{aligned} \mu_{t_1,\ldots,t_n}(A_1 \times \cdots \times A_n) &= \mu_{\pi_1,\ldots,\pi_n}(A_{\pi_1} \times \cdots \times A_{\pi_n}) \\ \mu_{t_1,\ldots,t_n,t_{n+1}}(A_1 \times \cdots \times A_n \times E) &= \mu_{t_1,\ldots,t_n}(A_1 \times \cdots \times A_n) \end{aligned}$$

Kolmogorov Existence Theorem

Consider the measurable space $(E, \mathcal{E})$, and the set S, and let $\mu_{t_1,\ldots,t_n}$ is a measure defined for all t_1, ..., $t_n \in S$ and $n \in \mathbb{N}$. If the collection of measures $\mu_{t_1,\ldots,t_n}$ is a consistent collection of measures, then there **exist** a **unique** stochastic process $(X_s)_{s \in S}$ which $\mu_{t_1,\ldots,t_n}$ is its FDD.

Kolmogorov Existence Theorem

Consistent Collection of Measures

Assume $(E, \mathcal{E})$ is a measurable space, and let S be a set, t_1, ..., $t_n \in S$, and assume $\mu_{t_1,\ldots,t_n}$ is a measure on $(E^n, \mathcal{E}^n)$. This collection of measures is called consistent iff for all natural n, for all t_1, ..., t_n, t_{n+1}, every permutation π of $\{1, \ldots, n\}$, and for all A_1, ..., $A_n \in \mathcal{E}$, we have:

$$\begin{aligned} \mu_{t_1,\ldots,t_n}(A_1 \times \cdots \times A_n) &= \mu_{\pi_1,\ldots,\pi_n}(A_{\pi_1} \times \cdots \times A_{\pi_n}) \\ \mu_{t_1,\ldots,t_n,t_{n+1}}(A_1 \times \cdots \times A_n \times E) &= \mu_{t_1,\ldots,t_n}(A_1 \times \cdots \times A_n) \end{aligned}$$

Kolmogorov Existence Theorem

Consider the measurable space $(E, \mathcal{E})$, and the set S, and let $\mu_{t_1,\ldots,t_n}$ is a measure defined for all t_1, ..., $t_n \in S$ and $n \in \mathbb{N}$. If the collection of measures $\mu_{t_1,\ldots,t_n}$ is a consistent collection of measures, then there **exist** a **unique** stochastic process $(X_s)_{s \in S}$ which $\mu_{t_1,\ldots,t_n}$ is its FDD.

As a consequence of this theorem, we can uniquely define a stochastic process by defining its FDDs (as long as the FDD collection is consistent).

Gaussian Process (GP) - Definition

Gaussian Process

The process $X = (X_s)_{s\in S}$ is a Gaussian Process with mean function $\mu(s)$ and the covariance function $K(s_1, s_2)$,

$$(X_s)_{s\in S} \quad \sim \quad \mathrm{GP}(\mu, K)$$

iff, for all $n \in \mathbb{N}$, and for all s_1, ..., $s_n \in S$:

$$\begin{bmatrix} X_{s_1} \\ \vdots \\ X_{s_n} \end{bmatrix} \quad \sim \quad \mathcal{N}_n\left(\begin{bmatrix} \mu(s_1) \\ \vdots \\ \mu(s_n) \end{bmatrix}, \begin{bmatrix} K(s_1, s_1) & \cdots & K(s_1, s_n) \\ \vdots & \ddots & \vdots \\ K(s_n, s_1) & \cdots & K(s_n, s_n) \end{bmatrix} \right)$$

Gaussian Process (GP) - Definition

Gaussian Process

The process $X = (X_s)_{s \in S}$ is a Gaussian Process with mean function $\mu(s)$ and the covariance function $K(s_1, s_2)$,

$$(X_s)_{s \in S} \quad \sim \quad \mathrm{GP}(\mu, K)$$

iff, for all $n \in \mathbb{N}$, and for all s_1, ..., $s_n \in S$:

$$\begin{bmatrix} X_{s_1} \\ \vdots \\ X_{s_n} \end{bmatrix} \quad \sim \quad \mathcal{N}_n \left(\begin{bmatrix} \mu(s_1) \\ \vdots \\ \mu(s_n) \end{bmatrix}, \begin{bmatrix} K(s_1, s_1) & \cdots & K(s_1, s_n) \\ \vdots & \ddots & \vdots \\ K(s_n, s_1) & \cdots & K(s_n, s_n) \end{bmatrix} \right)$$

It is straightforward to show the FDD defined in the definition above is a consistent collection of distributions, and therefore due to the Kolmogorov Existence Theorem, the GP defined above exists and is uniquely characterized by the mean function $\mu(s)$ and the covariance function $K(s, s')$.

Gaussian Process (GP) - Definition

Gaussian Process

The process $X = (X_s)_{s\in S}$ is a Gaussian Process with mean function $\mu(s)$ and the covariance function $K(s_1, s_2)$,

$$(X_s)_{s\in S} \quad \sim \quad \mathrm{GP}(\mu, K)$$

iff, for all $n \in \mathbb{N}$, and for all s_1, ..., $s_n \in S$:

$$\begin{bmatrix} X_{s_1} \\ \vdots \\ X_{s_n} \end{bmatrix} \quad \sim \quad \mathcal{N}_n\left(\begin{bmatrix} \mu(s_1) \\ \vdots \\ \mu(s_n) \end{bmatrix}, \begin{bmatrix} K(s_1, s_1) & \cdots & K(s_1, s_n) \\ \vdots & \ddots & \vdots \\ K(s_n, s_1) & \cdots & K(s_n, s_n) \end{bmatrix}\right)$$

It is straightforward to show the FDD defined in the definition above is a consistent collection of distributions, and therefore due to the Kolmogorov Existence Theorem, the GP defined above exists and is uniquely characterized by the mean function $\mu(s)$ and the covariance function $K(s, s')$.

Note that the mean function μ is a rather arbitrary function, while the covariance function K should be defined in such way that ensures the semi-positiveness of the covariance matrix of FDD.

Gaussian Process (GP) - Example

Example - Wiener process (Brownian Motion)
Assume $S = [0, T]$, and:

$$\begin{aligned} \mu(s) &= 0 \\ K(s_1, s_2) &= s_1 \wedge s_2 \end{aligned}$$

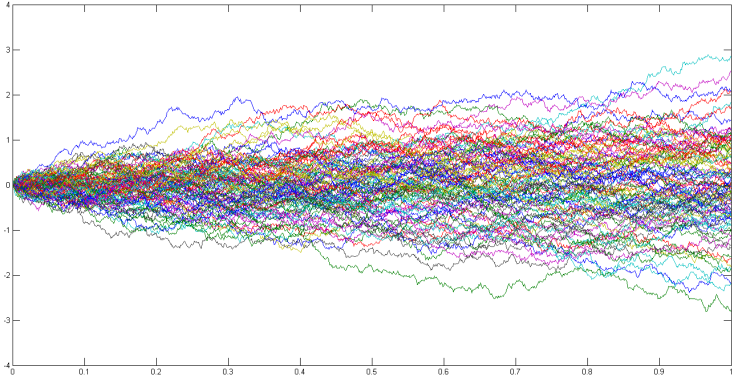
$(X_s)_{s \in [0,T]} \sim \mathrm{GP}(\mu, K)$ is called the Wiener process (or Brownian Motion).

Gaussian Process (GP) - Example

Example - Wiener process (Brownian Motion)
Assume $S = [0, T]$, and:

$$\begin{aligned} \mu(s) &= 0 \\ K(s_1, s_2) &= s_1 \wedge s_2 \end{aligned}$$

$(X_s)_{s \in [0,T]} \sim \mathrm{GP}(\mu, K)$ is called the Wiener process (or Brownian Motion).



Gaussian Process (GP) - Example

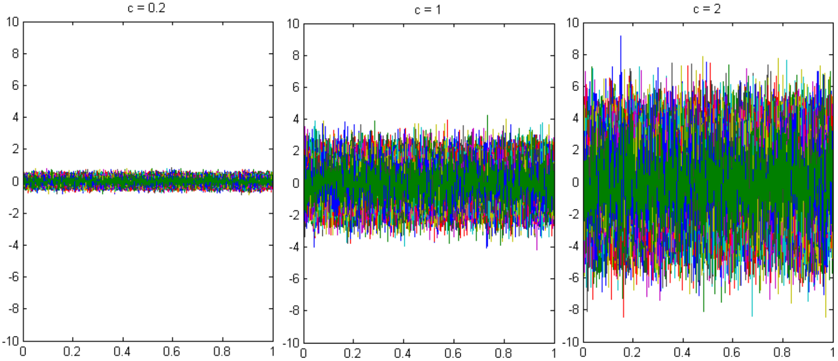
Example:

$$\begin{aligned} \mu(s) &= 0 \\ K(s_1, s_2) &= c^2 \, \delta_{\{0\}}(s_1 - s_2) \end{aligned}$$

Gaussian Process (GP) - Example

Example:

$$
\begin{aligned}
\mu(s) &= 0 \\
K(s_1, s_2) &= c^2 \; \delta_{\{0\}}(s_1 - s_2)
\end{aligned}
$$



Gaussian Process (GP) - Example

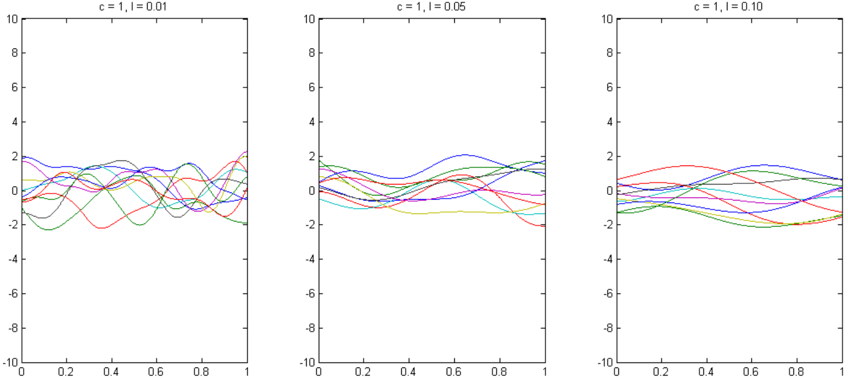
Example:

$$
\begin{aligned}
\mu(s) &= 0 \\
K(s_1, s_2) &= c^2 \ \exp\left[-\frac{1}{2l^2}(s_1 - s_2)^2\right]
\end{aligned}
$$

Gaussian Process (GP) - Example

Example:

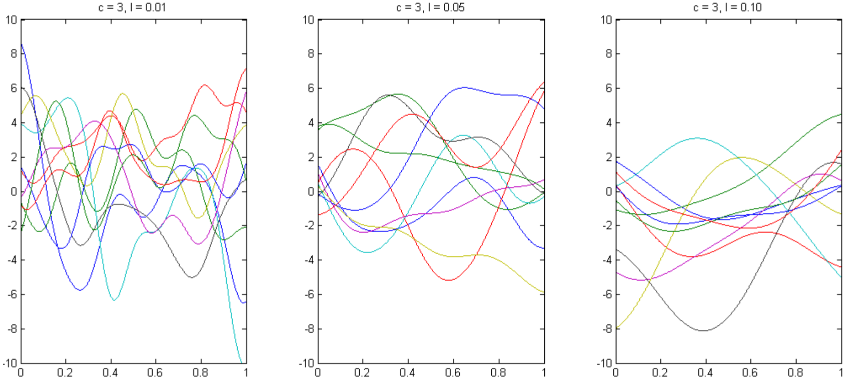
$$\begin{aligned} \mu(s) &= 0 \\ K(s_1, s_2) &= c^2 \exp\left[-\frac{1}{2l^2}(s_1 - s_2)^2\right] \end{aligned}$$



Gaussian Process (GP) - Example

Example:

$$\mu(s) = 0$$
$$K(s_1, s_2) = c^2 \exp\left[-\frac{1}{2l^2}(s_1 - s_2)^2\right]$$



Semi-parametric Regression with GP

Semi-parametric Regression with GP

$$\begin{aligned} y &= f(x) + \varepsilon \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2) \\ f &\sim \mathrm{GP}(\mu, K) \end{aligned}$$

Semi-parametric Regression with GP

Semi-parametric Regression with GP

$$
\begin{aligned}
y &= f(x) + \varepsilon \\
\varepsilon &\sim \mathcal{N}(0, \sigma^2) \\
f &\sim \mathrm{GP}(\mu, K)
\end{aligned}
$$

Assume the observed data is (x_i, y_i), for $i = 1, ..., N$, and define $\mathbf{x} = (x_1, ..., x_N)$, $\mathbf{y} = (y_1, ..., y_N)$.

Semi-parametric Regression with GP

Semi-parametric Regression with GP

$$\begin{aligned} y &= f(x) + \varepsilon \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2) \\ f &\sim \mathrm{GP}(\mu, K) \end{aligned}$$

Assume the observed data is (x_i, y_i), for $i = 1, ..., N$, and define $\mathbf{x} = (x_1, ..., x_N)$, $\mathbf{y} = (y_1, ..., y_N)$.

Moreover define $\mathbf{f} = (f(x_1), ..., f(x_N))$.

Semi-parametric Regression with GP

Semi-parametric Regression with GP

$$
\begin{aligned}
y &= f(x) + \varepsilon \\
\varepsilon &\sim \mathcal{N}(0, \sigma^2) \\
f &\sim \mathrm{GP}(\mu, K)
\end{aligned}
$$

Assume the observed data is (x_i, y_i), for $i = 1, \ldots, N$, and define $\mathbf{x} = (x_1, \ldots, x_N)$, $\mathbf{y} = (y_1, \ldots, y_N)$.

Moreover define $\mathbf{f} = (f(x_1), \ldots, f(x_N))$.

From the definition of the GP:

$$
\mathbf{f} \sim \mathcal{N}(\mathbf{m}, \mathbf{S})
$$

where

$$
\mathbf{m} = (\mu(x_1), \ldots, \mu(x_N)), \qquad \text{and} \qquad \mathbf{S} = \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_N) \\ \vdots & \ddots & \vdots \\ K(x_N, x_1) & \cdots & K(x_N, x_N) \end{bmatrix}
$$

Semi-parametric Regression with GP

From the definition of the model:

$$\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2\, \mathbf{I}_N)$$

Semi-parametric Regression with GP

From the definition of the model:

$$\mathbf{y}|\mathbf{f} \quad \sim \quad \mathcal{N}(\mathbf{f}, \sigma^2 \ \mathbf{I}_N)$$

Therefore the posterior distribution of $\mathbf{f}$ is:

$$\begin{aligned} p(\mathbf{f}|\mathbf{y}) \quad &\propto \quad p(\mathbf{f}) \ p(\mathbf{y}|\mathbf{f}) \\ &= \quad \varphi_N(\mathbf{f}; \mathbf{m}, \mathbf{S}) \ \varphi_N(\mathbf{y}; \mathbf{f}, \sigma^2 \mathbf{I}_N) \end{aligned}$$

Semi-parametric Regression with GP

From the definition of the model:

$$\mathbf{y}|\mathbf{f} \quad \sim \quad \mathcal{N}(\mathbf{f}, \sigma^2 \ \mathbf{I}_N)$$

Therefore the posterior distribution of $\mathbf{f}$ is:

$$\begin{aligned} p(\mathbf{f}|\mathbf{y}) \quad &\propto \quad p(\mathbf{f}) \ p(\mathbf{y}|\mathbf{f}) \\ &= \quad \varphi_N(\mathbf{f}; \mathbf{m}, \mathbf{S}) \ \varphi_N(\mathbf{y}; \mathbf{f}, \sigma^2 \mathbf{I}_N) \end{aligned}$$

and from this it follows that:

$$\mathbf{f}|\mathbf{y} \quad \sim \quad \mathcal{N}\left(\left[\sigma^{-2}\mathbf{I}_N + \mathbf{S}^{-1}\right]^{-1} \left[\sigma^{-2}\mathbf{y} + \mathbf{S}^{-1}\mathbf{m}\right], \left[\sigma^{-2}\mathbf{I}_N + \mathbf{S}^{-1}\right]^{-1} \right)$$

Semi-parametric Regression with GP

if we assume $\mathbf{x}^* = (x_1^*, ..., x_{N^*}^*)$, where $x_i \neq x_j^*$, and $\mathbf{f}^* = (f_1^*, ..., f_{N^*}^*)$, then:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m} \\ \mathbf{m}^* \end{bmatrix}, \begin{bmatrix} \mathbf{S} + \sigma^2 \mathbf{I}_N & \mathbf{S}_{xx^*} \\ \mathbf{S}_{x^*x} & \mathbf{S}^* \end{bmatrix} \right)$$

Semi-parametric Regression with GP

if we assume $\mathbf{x}^* = (x_1^*, ..., x_{N^*}^*)$, where $x_i \neq x_j^*$, and $\mathbf{f}^* = (f_1^*, ..., f_{N^*}^*)$, then:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m} \\ \mathbf{m}^* \end{bmatrix}, \begin{bmatrix} \mathbf{S} + \sigma^2 \mathbf{I}_N & \mathbf{S}_{xx^*} \\ \mathbf{S}_{x^*x} & \mathbf{S}^* \end{bmatrix} \right)$$

Therefore the posterior distribution of $\mathbf{f}^*$ is:

$$\mathbf{f}^* | \mathbf{y} \sim \mathcal{N}\left(\mathbf{m}^* + S_{x^*x} \left[\mathbf{S} + \sigma^2 \mathbf{I}_N \right]^{-1} (\mathbf{y} - \mathbf{m}), \mathbf{S}^* - \mathbf{S}_{x^*x} \left[\mathbf{S} + \sigma^2 \mathbf{I}_N \right]^{-1} \mathbf{S}_{xx^*} \right)$$

Semi-parametric Regression with GP

if we assume $\mathbf{x}^* = (x_1^*, ..., x_{N^*}^*)$, where $x_i \neq x_j^*$, and $\mathbf{f}^* = (f_1^*, ..., f_{N^*}^*)$, then:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m} \\ \mathbf{m}^* \end{bmatrix}, \begin{bmatrix} \mathbf{S} + \sigma^2 \mathbf{I}_N & \mathbf{S}_{xx^*} \\ \mathbf{S}_{x^*x} & \mathbf{S}^* \end{bmatrix} \right)$$

Therefore the posterior distribution of $\mathbf{f}^*$ is:

$$\mathbf{f}^* | \mathbf{y} \sim \mathcal{N}\left(\mathbf{m}^* + S_{x^*x} \left[\mathbf{S} + \sigma^2 \mathbf{I}_N \right]^{-1} (\mathbf{y} - \mathbf{m}), \mathbf{S}^* - \mathbf{S}_{x^*x} \left[\mathbf{S} + \sigma^2 \mathbf{I}_N \right]^{-1} \mathbf{S}_{xx^*} \right)$$

Remember if

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \right)$$

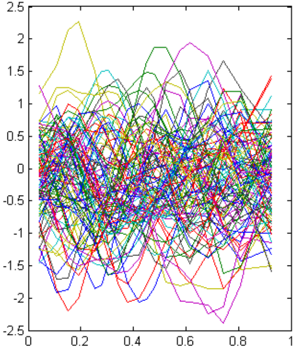
then

$$\mathbf{x}_2 | \mathbf{x}_1 \sim \mathcal{N}\left(\mathbf{m}_2 + \mathbf{S}_{21} \mathbf{S}_{11}^{-1} (\mathbf{x}_1 - \mathbf{m}_1), \mathbf{S}_{22} - \mathbf{S}_{21} \mathbf{S}_{11}^{-1} \mathbf{S}_{12} \right)$$

Semi-parametric Regression with GP: Example

prior:

$$
\begin{aligned}
\mu &= -0.16 \\
K &= 0.5 \exp\left[-\frac{(s_1 - s_2)^2}{0.2}\right]
\end{aligned}
$$

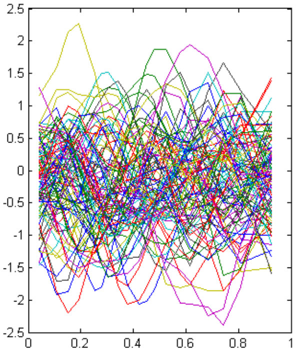


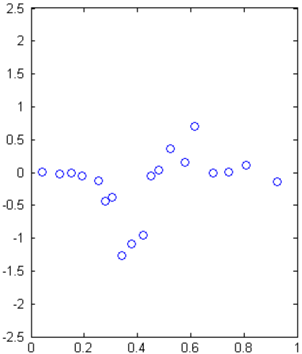
Semi-parametric Regression with GP: Example

prior:

$$\mu = -0.16$$
$$K = 0.5 \exp\left[-\frac{(s_1 - s_2)^2}{0.2}\right]$$

Data





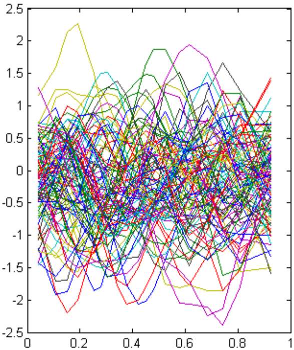
Semi-parametric Regression with GP: Example

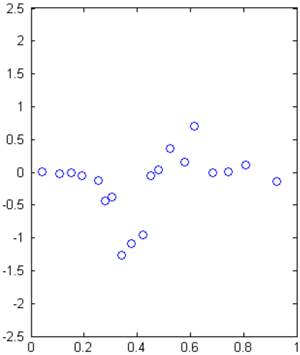
prior:

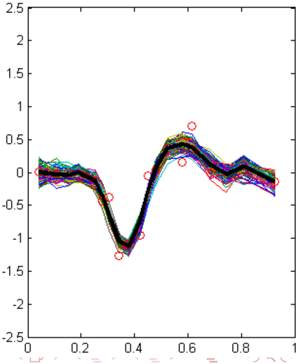
$$\mu = -0.16$$
$$K = 0.5 \exp\left[-\frac{(s_1 - s_2)^2}{0.2}\right]$$

Data

Posterior







Semi-parametric Regression with GP

Note that the value of hyperparameters (μ and K parameters) changes the posterior distribution. In the absence of prior information (on the smoothness of f and the noise level), we should put non-informative priors on them. In this case the analytical computation of posterior is not possible and we should employ the an MCMC simulation to sample from the posterior. Conditioned on the hyperparameters $\mathbf{f}$ is Gaussian, however usually the full conditional of parameters are not standard distributions, and we should use a (possibly adaptive) Metropolis-Hastings step to sample from them.

Semi-parametric Regression with GP

Note that the value of hyperparameters (μ and K parameters) changes the posterior distribution. In the absence of prior information (on the smoothness of f and the noise level), we should put non-informative priors on them. In this case the analytical computation of posterior is not possible and we should employ the an MCMC simulation to sample from the posterior. Conditioned on the hyperparameters $\mathbf{f}$ is Gaussian, however usually the full conditional of parameters are not standard distributions, and we should use a (possibly adaptive) Metropolis-Hastings step to sample from them.

Introduction

Binomial-Beta Conjugacy

Consider the Binomial r.v.:

$$
\begin{aligned}
S = \mathrm{Supp}(X) &= \{\mathbf{x} = (x_1, x_2);\; x_1, x_2 \in \{0, 1, ..., n\}, x_1 + x_2 = n\} \\
f(\mathbf{x}) &= \frac{n!}{x_1! x_2!} p_1^{x_1} p_2^{x_2} \; \mathbb{I}_S(\mathbf{x})
\end{aligned}
$$

with the parameter $p_1 \in [0, 1]$, and $p_2 = 1 - p_1$. The likelihood of this model is:

$$
L(p_1 | \mathbf{x}) \quad \propto \quad p_1^{x_1} p_2^{x_2}
$$

Binomial-Beta Conjugacy

Consider the Binomial r.v.:

$$\begin{aligned} S = \mathrm{Supp}(X) &= \{\mathbf{x} = (x_1, x_2);\ x_1, x_2 \in \{0, 1, ..., n\}, x_1 + x_2 = n\} \\ f(\mathbf{x}) &= \frac{n!}{x_1! x_2!} p_1^{x_1} p_2^{x_2}\ \mathbb{I}_S(\mathbf{x}) \end{aligned}$$

with the parameter $p_1 \in [0, 1]$, and $p_2 = 1 - p_1$. The likelihood of this model is:

$$L(p_1|\mathbf{x}) \quad \propto \quad p_1^{x_1} p_2^{x_2}$$

Therefore the conjugate prior has the following kernel:

$$p(p_1) \quad \propto \quad p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1}$$

That is the Kernel of the Beta distribution:

$$f(p_1) \quad = \quad \frac{1}{\mathrm{Beta}(\alpha_1, \alpha_2)}\ p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1}$$

Binomial-Beta Conjugacy

Binomial-Beta Conjugacy

Model:

$$X \sim \text{Binomial}(p)$$

Prior:

$$p \sim \text{Beta}(\alpha_1, \alpha_2)$$

Posterior:

$$p|X \sim \text{Beta}(\alpha_1 + x_1, \alpha_2 + x_2)$$

Multinomial-Dirichlet Conjugacy

Consider the Multinomial r.v.:

$$\begin{aligned} S = \mathrm{Supp}(X) &= \{\mathbf{x} = (x_1, \ldots, x_K);\ x_1, \ldots, x_K \in \{0, 1, \ldots, n\}, x_1 + \cdots + x_K = n\} \\ f(\mathbf{x}) &= \frac{n!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K}\ \mathbb{I}_S(\mathbf{x}) \end{aligned}$$

with the parameter $\mathbf{p} = (p_1, \ldots, p_{K-1}) \in (K-1)$-simplex[1], and $p_K = 1 - \sum_{j=1}^{K-1} p_j$.
The likelihood of this model is:

$$L(\mathbf{p}|\mathbf{x}) \quad \propto \quad p_1^{x_1} \cdots p_K^{x_K}$$

[1] The d-simplex is $\Delta^d = \{(x_1, \ldots, x_d); \forall j\ x_j \geq 0, \quad \text{and} \quad \sum_{j=1}^{d} x_j \leq 1\}$.

Multinomial-Dirichlet Conjugacy

Consider the Multinomial r.v.:

$$
\begin{aligned}
S = \mathrm{Supp}(X) &= \{\mathbf{x} = (x_1, \ldots, x_K);\; x_1, \ldots, x_K \in \{0, 1, \ldots, n\}, x_1 + \cdots + x_K = n\} \\
f(\mathbf{x}) &= \frac{n!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K} \; \mathbb{I}_S(\mathbf{x})
\end{aligned}
$$

with the parameter $\mathbf{p} = (p_1, \ldots, p_{K-1}) \in (K-1)$-simplex[1], and $p_K = 1 - \sum_{j=1}^{K-1} p_j$. The likelihood of this model is:

$$L(\mathbf{p}|\mathbf{x}) \quad \propto \quad p_1^{x_1} \cdots p_K^{x_K}$$

Therefore the conjugate prior has the following kernel:

$$p(\mathbf{p}) \quad \propto \quad p_1^{\alpha_1 - 1} \cdots p_2^{\alpha_2 - 1}$$

That is the Kernel of the **Dirichlet distribution**:

$$f(\mathbf{p}) \quad = \quad \frac{\prod_{j=1}^{K} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{K} \alpha_j\right)} \; p_1^{\alpha_1 - 1} \cdots p_K^{\alpha_K - 1}$$

[1] The d-simplex is $\Delta^d = \{(x_1, \ldots, x_d); \forall j \; x_j \geq 0, \;$ and $\; \sum_{j=1}^{d} x_j \leq 1\}$.

Dirichlet Distribution

Assume G_i are independent Gamma-distributed r.v.s:

$$G_i \quad \sim \quad \text{Gamma}(\alpha_i, 1) \qquad 1 \leq i \leq K$$

and define:

$$P_i \quad = \quad \frac{G_i}{\sum_{j=1}^{K} G_j} \qquad 1 \leq i \leq K$$

Dirichlet Distribution

Assume G_i are independent Gamma-distributed r.v.s:

$$G_i \quad \sim \quad \text{Gamma}(\alpha_i, 1) \qquad 1 \leq i \leq K$$

and define:

$$P_i \quad = \quad \frac{G_i}{\sum_{j=1}^{K} G_j} \qquad 1 \leq i \leq K$$

It can be shown that the PDF of $(P_1, ..., P_{K-1})$ is:

$$f(\mathbf{p}) \quad = \quad \frac{\prod_{j=1}^{K} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{K} \alpha_j\right)} \; p_1^{\alpha_1 - 1} \cdots p_K^{\alpha_K - 1}$$

where $\mathbf{p} = (p_1, ..., p_{K-1})$ and $p_K = 1 - \sum_{j=1}^{K-1} p_j$.

Dirichlet Distribution - Some Properties

$$\text{Dirichlet}(\alpha_1, \alpha_2) \quad \equiv \quad \text{Beta}(\alpha_1, \alpha_2)$$

Dirichlet Distribution - Some Properties

$$\text{Dirichlet}(\alpha_1, \alpha_2) \equiv \text{Beta}(\alpha_1, \alpha_2)$$

$$\text{Dirichlet}(1, ..., 1) \equiv \text{Unif}(\Delta^{K-1})$$

Dirichlet Distribution - Some Properties

$$\text{Dirichlet}(\alpha_1, \alpha_2) \equiv \text{Beta}(\alpha_1, \alpha_2)$$

$$\text{Dirichlet}(1, ..., 1) \equiv \text{Unif}(\Delta^{K-1})$$

Moreover:

$$E\left(\begin{bmatrix} P_1 \\ \vdots \\ P_K \end{bmatrix}\right) = \begin{bmatrix} \frac{\alpha_1}{\sum_{j=1}^K \alpha_j} \\ \vdots \\ \frac{\alpha_K}{\sum_{j=1}^K \alpha_j} \end{bmatrix}$$

Dirichlet Distribution - Some Properties

$$\text{Dirichlet}(\alpha_1, \alpha_2) \equiv \text{Beta}(\alpha_1, \alpha_2)$$

$$\text{Dirichlet}(1, ..., 1) \equiv \text{Unif}(\Delta^{K-1})$$

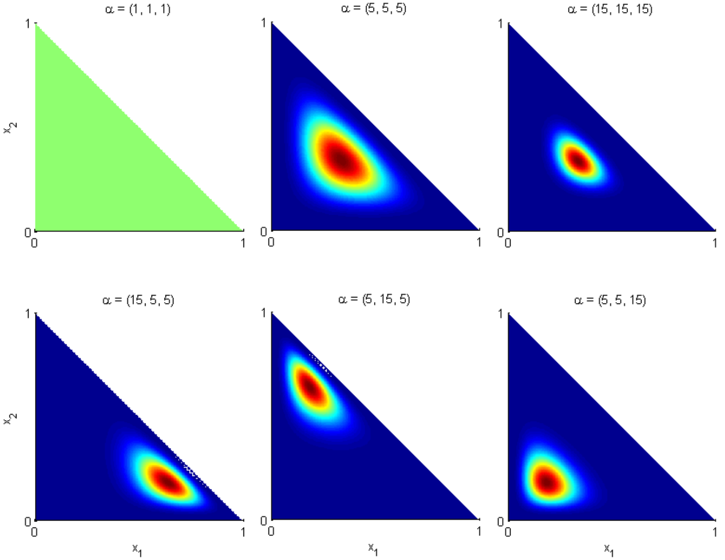
Moreover:

$$E\left(\begin{bmatrix} P_1 \\ \vdots \\ P_K \end{bmatrix}\right) = \begin{bmatrix} \frac{\alpha_1}{\sum_{j=1}^K \alpha_j} \\ \vdots \\ \frac{\alpha_K}{\sum_{j=1}^K \alpha_j} \end{bmatrix}$$

and the mode of the distribution is:

$$\begin{bmatrix} \frac{\alpha_1 - 1}{\left(\sum_{j=1}^K \alpha_j\right) - K} \\ \vdots \\ \frac{\alpha_{K-1} - 1}{\left(\sum_{j=1}^K \alpha_j\right) - K} \end{bmatrix} \qquad \forall j \; \alpha_j > 1$$

$(x_1, x_2) \sim \text{Dirichlet}(\alpha)$



Dirichlet Distribution - Some Properties

The marginals of Dirichlet distribution are beta:

$$P_i \quad \sim \quad \mathrm{Beta}(\alpha_i, \sum_{j \neq i} \alpha_j)$$

Dirichlet Distribution - Some Properties

The marginals of Dirichlet distribution are beta:

$$P_i \quad \sim \quad \text{Beta}(\alpha_i, \sum_{j \neq i} \alpha_j)$$

and

$$\left(P_1, \ldots, P_s, \sum_{j=s+1}^{K} P_j\right) \quad \sim \quad \text{Dirichlet}(\alpha_1, \ldots, \alpha_s, \sum_{j=s+1}^{K} \alpha_j)$$

Dirichlet Distribution - Some Properties

The marginals of Dirichlet distribution are beta:

$$P_i \sim \text{Beta}(\alpha_i, \sum_{j \neq i} \alpha_j)$$

and

$$\left(P_1, ..., P_s, \sum_{j=s+1}^{K} P_j\right) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_s, \sum_{j=s+1}^{K} \alpha_j)$$

In general, we can aggregate any subset of Dirichlet variables yields to obtain a new random variable, and this new random variable is Dirichlet distributed, with corresponding aggregation of the parameters (remember that $\text{Gamma}(\alpha_1, 1) + \text{Gamma}(\alpha_2, 1) \stackrel{L}{\equiv} \text{Gamma}(\alpha_1 + \alpha_2, 1)$).

Dirichlet Distribution - Aggregation Property

Assume:

$$(P_1, ..., P_d) \quad \sim \quad \text{Dirichlet}(\alpha_1, ..., \alpha_d)$$

and let $\{A_1, ..., A_K\}$ be a partition of $\{1, ..., d\}$. Then

$$(\sum_{i \in A_1} P_i, ..., \sum_{i \in A_K} P_i) \quad \sim \quad \text{Dirichlet}(\sum_{i \in A_1} \alpha_i, ..., \sum_{i \in A_K} \alpha_i)$$

This is called the “Aggregation Property” of Dirichlet distribution.

Dirichlet Distribution - Random Number Generation

The first way to generate random numbers from the Dirichlet distribution is to rely on its definition based on normaliziation of independent Gamma random variables.

Dirichlet Distribution - Random Number Generation

The first way to generate random numbers from the Dirichlet distribution is to rely on its definition based on normaliziation of independent Gamma random variables.

The second approach is the stick-breaking method. First we should generate v_1, ..., v_{d-1}:

$$v_i \quad \sim \quad \text{Beta}(\alpha_i, \sum_{j=i+1}^{d} \alpha_j)$$

and set $v_d = 1$, and then build the p_is:

$$p_i \quad = \quad \begin{cases} v_1 & i = 1 \\ v_i \prod_{j=1}^{i-1}(1 - v_j) & 1 < i \leq d \end{cases}$$

It can be shown that:

$$(p_1, ..., p_d) \quad \sim \quad \text{Dirichlet}(\alpha_i, ..., \alpha_d)$$

Dirichlet Distribution - Pólya Urn Scheme

Consider the Dirichlet random variable

$$(P_1, ..., P_d) \quad \sim \quad \text{Dirichlet}(\alpha_1, ..., \alpha_d)$$

We put α_i balls of color i for $i = 1, ..., d$ in an empty urn. The initial fraction of the balls of different color in the urn are:

$$(f_1^0, ..., f_d^0) \quad = \quad (\alpha_1/\alpha, ..., \alpha_d/\alpha)$$

where $\alpha = \sum_{i=1}^{d} = \alpha_i$.

Dirichlet Distribution - Pólya Urn Scheme

Consider the Dirichlet random variable

$$(P_1, ..., P_d) \quad \sim \quad \text{Dirichlet}(\alpha_1, ..., \alpha_d)$$

We put α_i balls of color i for $i = 1, ..., d$ in an empty urn. The initial fraction of the balls of different color in the urn are:

$$(f_1^0, ..., f_d^0) \quad = \quad (\alpha_1/\alpha, ..., \alpha_d/\alpha)$$

where $\alpha = \sum_{i=1}^{d} = \alpha_i$. Then in the first step we draw a ball randomly from this urn, record it color X_1, and then return it back into the urn with another ball of the same color. After the first step the fraction of the balls of different color in the urn are:

$$(f_1^{(1)}, ..., f_d^{(1)}) \quad = \quad ((\alpha_1 + \delta_1(X_1))/(\alpha + 1), ..., (\alpha_d + \delta_d(X_1))/(\alpha + 1))$$

Dirichlet Distribution - Pólya Urn Scheme

Consider the Dirichlet random variable

$$(P_1, ..., P_d) \quad \sim \quad \text{Dirichlet}(\alpha_1, ..., \alpha_d)$$

We put α_i balls of color i for $i = 1, ..., d$ in an empty urn. The initial fraction of the balls of different color in the urn are:

$$(f_1^0, ..., f_d^0) \quad = \quad (\alpha_1/\alpha, ..., \alpha_d/\alpha)$$

where $\alpha = \sum_{i=1}^{d} = \alpha_i$. Then in the first step we draw a ball randomly from this urn, record it color X_1, and then return it back into the urn with another ball of the same color. After the first step the fraction of the balls of different color in the urn are:

$$(f_1^{(1)}, ..., f_d^{(1)}) \quad = \quad ((\alpha_1 + \delta_1(X_1))/(\alpha + 1), ..., (\alpha_d + \delta_d(X_1))/(\alpha + 1))$$

If we continue with this sampling scheme, the asymptotic fraction of the balls of different colors in the urn will be a random draw from Dirichlet$(\alpha_1, ..., \alpha_d)$:

$$(f_1^{(n)}, ..., f_d^{(n)}) \quad \underset{n \to +\infty}{\overset{L}{\to}} \quad \text{Dirichlet}(\alpha_1, ..., \alpha_d)$$

RPD

Consider the measurable space $(E, \mathcal{E})$, and assume $\mathcal{F}$ is the space of probability measures on this space.

RPD

Consider the measurable space $(E, \mathcal{E})$, and assume $\mathcal{F}$ is the space of probability measures on this space.

The simplest Case: $E = \{0, 1\}$
On this space the family of probability distributions can be formulated by a single parameter $p \in [0, 1]$:

$$F(x) \quad = \quad (1-p)\ \mathbf{1}_{[0,+\infty)}(x) + p\ \mathbf{1}_{[1,+\infty)}(x)$$

Therefore by putting a prior on p we can define a prior on the probability distributions on this space.

RPD

Consider the measurable space $(E, \mathcal{E})$, and assume $\mathcal{F}$ is the space of probability measures on this space.

The simplest Case: $E = \{0, 1\}$
On this space the family of probability distributions can be formulated by a single parameter $p \in [0, 1]$:

$$F(x) \quad = \quad (1 - p)\ \mathbf{1}_{[0,+\infty)}(x) + p\ \mathbf{1}_{[1,+\infty)}(x)$$

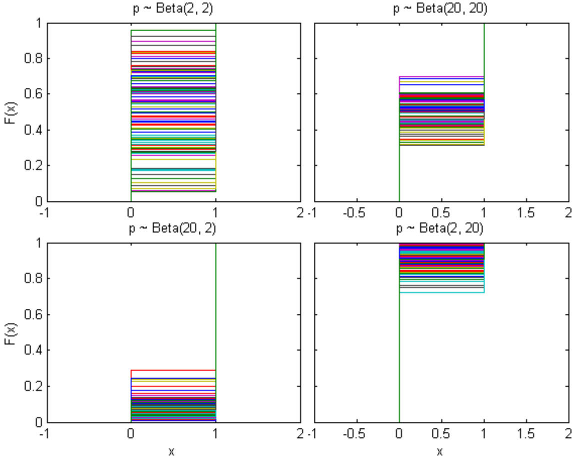
Therefore by putting a prior on p we can define a prior on the probability distributions on this space.

A natural choice is the beta distribution:

$$p \quad \sim \quad \text{Beta}(\alpha_1, \alpha_2)$$

and this completes the definition of a random measure on $(E, \mathcal{E})$.

$F(x) = (1-p)\mathbf{1}_{[0,+\infty)}(x) + p\mathbf{1}_{[1,+\infty)}(x)$



RPD

The same idea can be extended to a finite space.

RPD

The same idea can be extended to a finite space.

Example: $E = \{1, ..., K\}$
On this space the family of probability distributions can be formulated by the vector of parameters $\mathbf{p} \in \Delta^{K-1}$:

$$F(x) = p_1 \, \mathbf{1}_{[1,+\infty)}(x) + \cdots + p_{K-1} \, \mathbf{1}_{[K-1,+\infty)}(x) + p_K \, \mathbf{1}_{[K,+\infty)}(x)$$

where $p_K = 1 - \sum_{j=1}^{K-1} p_j$. Therefore by putting a prior on p we can define a prior on the probability distributions on this space.

RPD

The same idea can be extended to a finite space.

Example: $E = \{1, ..., K\}$
On this space the family of probability distributions can be formulated by the vector of parameters $\mathbf{p} \in \Delta^{K-1}$:

$$F(x) \quad = \quad p_1 \ \mathbf{1}_{[1,+\infty)}(x) + \cdots + p_{K-1} \ \mathbf{1}_{[K-1,+\infty)}(x) + p_K \ \mathbf{1}_{[K,+\infty)}(x)$$

where $p_K = 1 - \sum_{j=1}^{K-1} p_j$. Therefore by putting a prior on p we can define a prior on the probability distributions on this space.

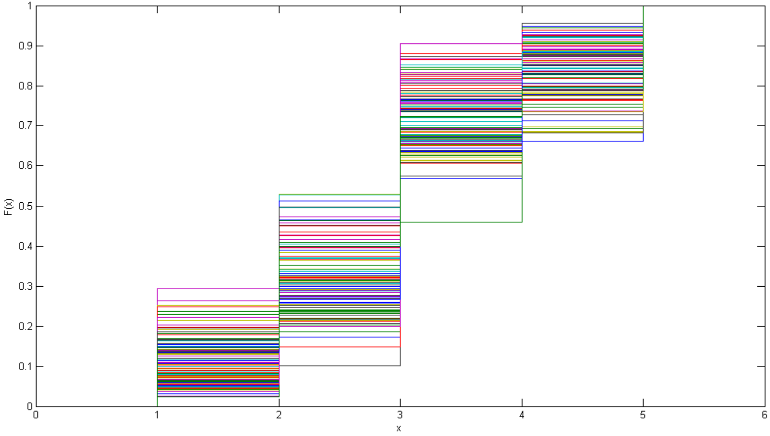
A natural choice is the Dirichlet distribution:

$$\mathbf{p} \quad \sim \quad \text{Dirichlet}(\alpha_1, ..., \alpha_K)$$

and this completes the definition of a random measure on $(E, \mathcal{E})$.

$F(x) = \sum_{j=1}^{K} p_j \mathbf{1}_{[j,+\infty)}(x)$

$p \sim \text{Dirichlet}(3, 5, 10, 3, 4)$



Dirichlet Process - Definition

Dirichlet process generalizes this idea and defines a prior on the distributions on the infinite measurable space $(E, \mathcal{E})$.

Dirichlet Process - Definition

Dirichlet process generalizes this idea and defines a prior on the distributions on the infinite measurable space $(E, \mathcal{E})$.

Dirichlet Process

Assume G_0 is a probability measure on $(E, \mathcal{E})$, and let c be a positive real number. A random measure F is called a Dirichlet Process (DP),

$$F \quad \sim \quad \mathrm{DP}(c, G_0)$$

iff for any finite measurable partition of E,

$$\forall k \in \mathbb{N}, \; E_1 \cup \cdots \cup E_k = E, \quad E_i \cap E_j = \emptyset, \quad E_i \in \mathcal{E}$$

we have:

$$(F(E_1), ..., F(E_k)) \quad \sim \quad \mathrm{Dirichlet}(c \;\, G_0(E_1), ..., c \;\, G_0(E_k))$$

G_0 and c are called the centering (or base) measure, and the concentration parameter, respectively.

DP: Parameter Interpretability

The existence and uniqueness of DP is ensured by the Kolmogorov Existence Theorem[2].

[2] T. S. Ferguson, A Bayesian analysis of some nonparametric problems, Annals of Statistics, 1(2), pp. 209–230, 1973

DP: Parameter Interpretability

The existence and uniqueness of DP is ensured by the Kolmogorov Existence Theorem[2].

Consider the measurable set $A \in \mathcal{E}$. Then (A, A^c) is a finite measurable partition of E, and by Ferguson's definition of DP we have:

$$(F(A), F(A^c)) \quad \sim \quad \text{Dirichlet}(c \; G_0(A), c \; G_0(A^c))$$

therefore:

$$\mathbb{E}\left[F(A)\right] \quad = \quad \frac{c \; G_0(A)}{c \; G_0(A) + c \; G_0(A^c)} = G_0(A)$$

$$\mathbb{V}\left[F(A)\right] \quad = \quad \frac{c^2 \; G_0(A) \; G_0(A^c)}{\left[c \; G_0(A) + c \; G_0(A^c)\right]^2 \left[c \; G_0(A) + c \; G_0(A^c) + 1\right]} = \frac{G_0(A)\left[1 - G_0(A)\right]}{c + 1}$$

[2] T. S. Ferguson, A Bayesian analysis of some nonparametric problems, Annals of Statistics, 1(2), pp. 209–230, 1973

DP: Parameter Interpretability

The existence and uniqueness of DP is ensured by the Kolmogorov Existence Theorem[2].

Consider the measurable set $A \in \mathcal{E}$. Then (A, A^c) is a finite measurable partition of E, and by Ferguson's definition of DP we have:

$$(F(A), F(A^c)) \quad \sim \quad \text{Dirichlet}(c\ G_0(A), c\ G_0(A^c))$$

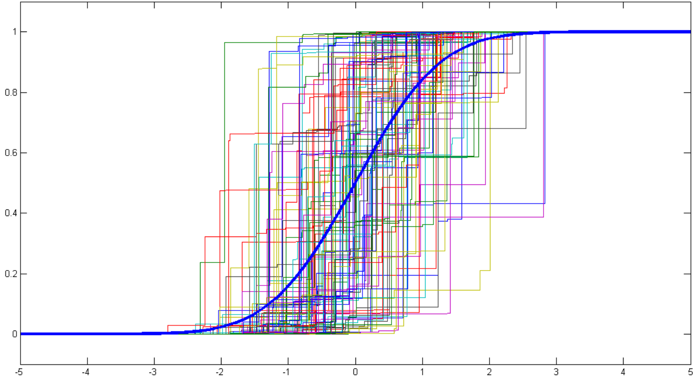
therefore:

$$\begin{aligned}
\mathbb{E}\left[F(A)\right] &= \frac{c\ G_0(A)}{c\ G_0(A) + c\ G_0(A^c)} = G_0(A) \\
\mathbb{V}\left[F(A)\right] &= \frac{c^2\ G_0(A)\ G_0(A^c)}{\left[c\ G_0(A) + c\ G_0(A^c)\right]^2 \left[c\ G_0(A) + c\ G_0(A^c) + 1\right]} = \frac{G_0(A)\left[1 - G_0(A)\right]}{c + 1}
\end{aligned}$$

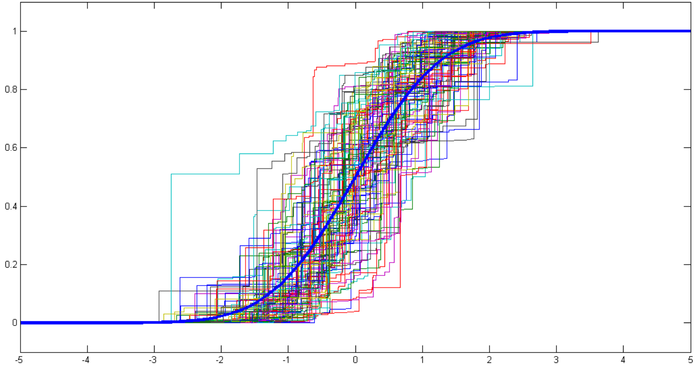
Therefore It can be seen that draws from DP(c, G_0) are concentrated around the centering measure G_0, with a dispersion that is inversely proportional to the concentration parameter c.

[2] T. S. Ferguson, A Bayesian analysis of some nonparametric problems, Annals of Statistics, 1(2), pp. 209–230, 1973

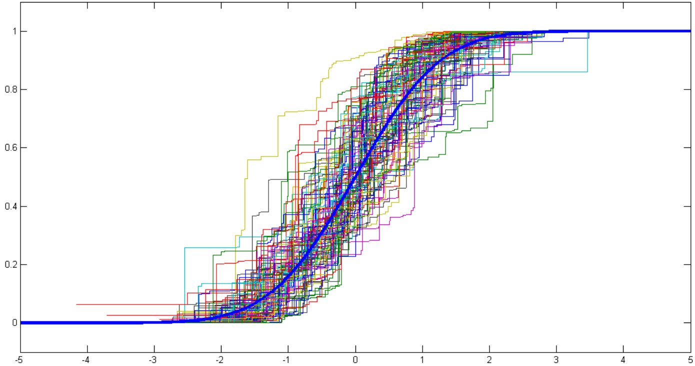
$G \sim \mathrm{DP}(G_0, 1)$, with $G_0 \equiv \mathcal{N}(0, 1)$



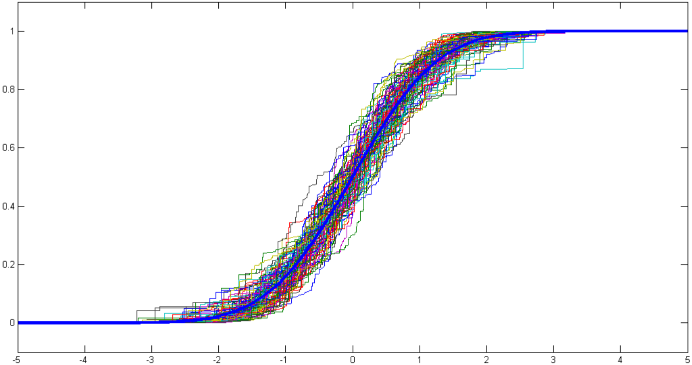
$G \sim \mathrm{DP}(G_0, 5)$, with $G_0 \equiv \mathcal{N}(0, 1)$



$G \sim \mathrm{DP}(G_0, 10)$, with $G_0 \equiv \mathcal{N}(0,1)$



$G \sim \mathrm{DP}(G_0, 50)$, with $G_0 \equiv \mathcal{N}(0, 1)$



DP: Conjugacy

Assume:

$$F \quad \sim \quad \mathrm{DP}(c, G_0)$$

and let θ_1 be a draw from F:

$$\theta_1 | F \quad \sim \quad F$$

DP: Conjugacy

Assume:

$$F \sim \mathrm{DP}(c, G_0)$$

and let θ_1 be a draw from F:

$$\theta_1 | F \sim F$$

It can be shown that the posterior of F is also a DP:

$$F | \theta_1 \sim \mathrm{DP}\left(c+1, \frac{c}{c+1} G_0 + \frac{1}{c+1} \delta_{\theta_1}\right)$$

Where δ_θ is a point mass centered at θ.

DP: Conjugacy

Assume:

$$F \sim \mathrm{DP}(c, G_0)$$

and let θ_1 be a draw from F:

$$\theta_1 | F \sim F$$

It can be shown that the posterior of F is also a DP:

$$F|\theta_1 \sim \mathrm{DP}\left(c+1, \frac{c}{c+1} G_0 + \frac{1}{c+1}\delta_{\theta_1}\right)$$

Where δ_θ is a point mass centered at θ.

By mathematical induction, this result can be extended to an iid sample of size N:

$$\theta_i | F \overset{iid}{\sim} F \qquad i = 1, 2, ..., N$$

Then

$$F|\theta_1, ..., \theta_N \sim \mathrm{DP}\left(c+N, \frac{c}{c+N} G_0 + \frac{1}{c+N}\sum_{j=1}^{N}\delta_{\theta_j}\right)$$

Stick-Breaking Representation

The Ferguson's definition of DP is not a constructive definition. The "Steak-Breaking Representation" (SB) of DP provides a constructive definition[3].

[3] J. Sethuraman, A constructive definition of Dirichlet priors, Statistica Sinica, 4, pp. 639-650, 1994

Stick-Breaking Representation

The Ferguson's definition of DP is not a constructive definition. The "Steak-Breaking Representation" (SB) of DP provides a constructive definition[3].

SB Representation of DP

Assume $v_i \overset{iid}{\sim} \text{Beta}(1, c)$, and let $w = (w_1, w_2, \dots)$ be an infinite sequence of weights, obtained by the following SB process:

$$w_k = \begin{cases} v_1 & k = 1 \\ v_k \prod_{j=1}^{k-1}(1 - v_j) & k > 1 \end{cases}$$

Moreover assume $\theta_i \overset{iid}{\sim} G_0$ (and independent from v_is), and consider the following discrete random measure:

$$F(\theta) = \sum_{j=1}^{+\infty} w_j \delta_{\theta_j}(\theta)$$

This guarantees that $F \sim \text{DP}(c, G_0)$. Conversely samples from a DP are a.s. discrete measures and have a SB representation.

[3] J. Sethuraman, A constructive definition of Dirichlet priors, Statistica Sinica, 4, pp. 639-650, 1994

Stick-Breaking Representation: Mixture Weights

The mixture weights $(w_1, w_2, \ldots)$ distribute a "stick" of length 1 between the countably infinite number of mixture components.

Stick-Breaking Representation: Mixture Weights

The mixture weights $(w_1, w_2, \ldots)$ distribute a "stick" of length 1 between the countably infinite number of mixture components.

The first component's weight is equal to $w_1 := \beta_1$. The remaining part of the stick, with the length $1 - \beta_1$, will be distributed among the rest of the weights.

Stick-Breaking Representation: Mixture Weights

The mixture weights $(w_1, w_2, \ldots)$ distribute a "stick" of length 1 between the countably infinite number of mixture components.

The first component's weight is equal to $w_1 := \beta_1$. The remaining part of the stick, with the length $1 - \beta_1$, will be distributed among the rest of the weights.

The second component's weight is equal to proportion β_2 of the remaining part of the stick: $w_2 := \beta_2(1 - \beta_1)$. The remaining part of the stick, with the length $(1 - \beta_1)(1 - \beta_2)$, will be distributed among the rest of the weights.

Stick-Breaking Representation: Mixture Weights

The mixture weights $(w_1, w_2, \ldots)$ distribute a "stick" of length 1 between the countably infinite number of mixture components.

The first component's weight is equal to $w_1 := \beta_1$. The remaining part of the stick, with the length $1 - \beta_1$, will be distributed among the rest of the weights.

The second component's weight is equal to proportion β_2 of the remaining part of the stick: $w_2 := \beta_2(1 - \beta_1)$. The remaining part of the stick, with the length $(1 - \beta_1)(1 - \beta_2)$, will be distributed among the rest of the weights.

The third component's weight is equal to proportion β_3 of the remaining part of the stick: $w_3 := \beta_3(1 - \beta_1)(1 - \beta_2)$, and so on...

Stick-Breaking Representation: Mixture Weights

The mixture weights $(w_1, w_2, \ldots)$ distribute a "stick" of length 1 between the countably infinite number of mixture components.

The first component's weight is equal to $w_1 := \beta_1$. The remaining part of the stick, with the length $1 - \beta_1$, will be distributed among the rest of the weights.

The second component's weight is equal to proportion β_2 of the remaining part of the stick: $w_2 := \beta_2(1 - \beta_1)$. The remaining part of the stick, with the length $(1 - \beta_1)(1 - \beta_2)$, will be distributed among the rest of the weights.

The third component's weight is equal to proportion β_3 of the remaining part of the stick: $w_3 := \beta_3(1 - \beta_1)(1 - \beta_2)$, and so on...

In the literature it is common to denote this sequence of weights by

$$\mathbf{w} \sim \mathrm{GEM}(\alpha)$$

where GEM stands for Griffiths, Engen, and McCloskey.

Stick-Breaking Representation: Mixture Weights

The mixture weights definition implies:

$$\mathbb{E}(w_k) \quad = \quad \frac{1}{1+c}\left(\frac{c}{1+c}\right)^{k-1}$$

That means the expected value of the weights decays exponentially to zero.

Stick-Breaking Representation: Mixture Weights

The mixture weights definition implies:

$$\mathbb{E}(w_k) = \frac{1}{1+c}\left(\frac{c}{1+c}\right)^{k-1}$$

That means the expected value of the weights decays exponentially to zero.

For small values of c, almost all the probability mass will be assigned to the first few components of the mixture. When $c \to +\infty$, the random probability measure F goes to the centering measure G_0.

DP: Predictive Distribution

Assume $G \sim \mathrm{DP}(c, G_0)$, and $\theta_i | G \overset{iid}{\sim} G$, for $i = 1, ..., N$. Since G is almost surely discrete, with positive probability there are ties among θ_is. Let $\{\theta_1^*, ..., \theta_{N^*}^*\}$ be the set of distinct values of θ_is:

$$\forall i \in \{1, ..., N\}, \quad \exists! j \in \{1, ..., N^*\}, \quad \theta_i = \theta_j^*$$

[4] D. Blackwell and J. B. MacQueen, Ferguson distributions via Pólya urn schemes, Annals of Statistics, 1(2), pp. 353-355, 1973.

DP: Predictive Distribution

Assume $G \sim \mathrm{DP}(c, G_0)$, and $\theta_i | G \overset{iid}{\sim} G$, for $i = 1, ..., N$. Since G is almost surely discrete, with positive probability there are ties among θ_is. Let $\{\theta_1^*, ..., \theta_{N^*}^*\}$ be the set of distinct values of θ_is:

$$\forall i \in \{1, ..., N\}, \quad \exists ! j \in \{1, ..., N^*\}, \quad \theta_i = \theta_j^*$$

It can be shown that the predictive distribution of θ_{N+1} is[4]:

$$\theta_{N+1} | \theta_1, ..., \theta_N \quad \sim \quad \frac{c}{c+N} G_0 + \frac{1}{c+N} \sum_{j=1}^{N^*} n_j \delta_{\theta_j^*}$$

where n_j is the number of previous observations of θ_j^*.

[4] D. Blackwell and J. B. MacQueen, Ferguson distributions via Pólya urn schemes, Annals of Statistics, 1(2), pp. 353-355, 1973.

DP: Predictive Distribution

Assume $G \sim \mathrm{DP}(c, G_0)$, and $\theta_i | G \overset{iid}{\sim} G$, for $i = 1, \ldots, N$. Since G is almost surely discrete, with positive probability there are ties among θ_is. Let $\{\theta_1^*, \ldots, \theta_{N^*}^*\}$ be the set of distinct values of θ_is:

$$\forall i \in \{1, \ldots, N\}, \quad \exists! j \in \{1, \ldots, N^*\}, \quad \theta_i = \theta_j^*$$

It can be shown that the predictive distribution of θ_{N+1} is[4]:

$$\theta_{N+1} | \theta_1, \ldots, \theta_N \quad \sim \quad \frac{c}{c+N} G_0 + \frac{1}{c+N} \sum_{j=1}^{N^*} n_j \delta_{\theta_j^*}$$

where n_j is the number of previous observations of θ_j^*.

In other words, the new realization from G, is either a new realization from the base measure G_0 with probability $\frac{c}{c+N}$, or it is equal to one of the previously observed realizations with probability $\frac{n_j}{c+N}$.

[4] D. Blackwell and J. B. MacQueen, Ferguson distributions via Pólya urn schemes, Annals of Statistics, 1(2), pp. 353-355, 1973.

DP's Predictive Distribution As a Pólya Urn

We can interpret the predictive distribution,

$$\theta_{N+1}|\theta_1, \ldots, \theta_N \quad \sim \quad \frac{c}{c+N} G_0 + \frac{1}{c+N} \sum_{j=1}^{N^*} n_j \delta_{\theta_j^*}$$

as a Pólya urn model. We should interpret each θ as a "color", so G_0 and G are distributions over colors.

DP's Predictive Distribution As a Pólya Urn

We can interpret the predictive distribution,

$$\theta_{N+1}|\theta_1, \ldots, \theta_N \quad \sim \quad \frac{c}{c+N} G_0 + \frac{1}{c+N} \sum_{j=1}^{N^*} n_j \delta_{\theta_j^*}$$

as a Pólya urn model. We should interpret each θ as a "color", so G_0 and G are distributions over colors. At the first iteration, we draw a color from the base distribution G_0 and put a ball of this color in the empty urn.

DP's Predictive Distribution As a Pólya Urn

We can interpret the predictive distribution,

$$\theta_{N+1}|\theta_1, \ldots, \theta_N \quad \sim \quad \frac{c}{c+N} G_0 + \frac{1}{c+N} \sum_{j=1}^{N^*} n_j \delta_{\theta_j^*}$$

as a Pólya urn model. We should interpret each θ as a "color", so G_0 and G are distributions over colors. At the first iteration, we draw a color from the base distribution G_0 and put a ball of this color in the empty urn. In the $k+1$-th step (note that at this step there are k balls in the urn), with probability $\frac{c}{c+N}$ we draw a new color from G_0 and put a new ball with this color in the urn, or with probability $\frac{N}{c+N}$ we draw a ball from the urn, and return it to the urn along with a new ball of the same color.

DP's Predictive Distribution As a Pólya Urn

We can interpret the predictive distribution,

$$\theta_{N+1}|\theta_1, ..., \theta_N \quad \sim \quad \frac{c}{c+N} G_0 + \frac{1}{c+N} \sum_{j=1}^{N^*} n_j \delta_{\theta_j^*}$$

as a Pólya urn model. We should interpret each θ as a "color", so G_0 and G are distributions over colors. At the first iteration, we draw a color from the base distribution G_0 and put a ball of this color in the empty urn. In the $k+1$-th step (note that at this step there are k balls in the urn), with probability $\frac{c}{c+N}$ we draw a new color from G_0 and put a new ball with this color in the urn, or with probability $\frac{N}{c+N}$ we draw a ball from the urn, and return it to the urn along with a new ball of the same color.

Asymptotically, the distribution of the balls' colors in the urn is a random draw from $\text{DP}(c, G_0)$.

DP's Predictive Distribution As a Pólya Urn

We can interpret the predictive distribution,

$$\theta_{N+1}|\theta_1, ..., \theta_N \quad \sim \quad \frac{c}{c+N} G_0 + \frac{1}{c+N} \sum_{j=1}^{N^*} n_j \delta_{\theta_j^*}$$

as a Pólya urn model. We should interpret each θ as a "color", so G_0 and G are distributions over colors. At the first iteration, we draw a color from the base distribution G_0 and put a ball of this color in the empty urn. In the $k+1$-th step (note that at this step there are k balls in the urn), with probability $\frac{c}{c+N}$ we draw a new color from G_0 and put a new ball with this color in the urn, or with probability $\frac{N}{c+N}$ we draw a ball from the urn, and return it to the urn along with a new ball of the same color.

Asymptotically, the distribution of the balls' colors in the urn is a random draw from $\mathrm{DP}(c, G_0)$.

Note that the obtained asymptotic distribution is discrete, and for any finite N, it provides a clustering of N observations in $K \leq N$ clusters (K is the number of distinct colors).

DP's Predictive Distribution As a Pólya Urn

We can interpret the predictive distribution,

$$\theta_{N+1}|\theta_1, \ldots, \theta_N \quad \sim \quad \frac{c}{c+N} G_0 + \frac{1}{c+N} \sum_{j=1}^{N^*} n_j \delta_{\theta_j^*}$$

as a Pólya urn model. We should interpret each θ as a "color", so G_0 and G are distributions over colors. At the first iteration, we draw a color from the base distribution G_0 and put a ball of this color in the empty urn. In the $k+1$-th step (note that at this step there are k balls in the urn), with probability $\frac{c}{c+N}$ we draw a new color from G_0 and put a new ball with this color in the urn, or with probability $\frac{N}{c+N}$ we draw a ball from the urn, and return it to the urn along with a new ball of the same color.

Asymptotically, the distribution of the balls' colors in the urn is a random draw from $\mathrm{DP}(c, G_0)$.

Note that the obtained asymptotic distribution is discrete, and for any finite N, it provides a clustering of N observations in $K \leq N$ clusters (K is the number of distinct colors).

Moreover note that K is not fixed a priori and increases with N.

Chinese Restaurant Process

Assume $G \sim \mathrm{DP}(c, G_0)$, and for $i = 1, \ldots, n$, let $\theta_i | G \sim G$. As it has been explained earlier with positive probability there are repeated values among θ_is. Therefore DP induces a partitioning among the n observations.

[5] J. Pitman, Combinatorial stochastic processes, Technical Report 621, U. C. Berkeley Department of Statistics, August 2002.

Chinese Restaurant Process

Assume $G \sim \text{DP}(c, G_0)$, and for $i = 1, \ldots, n$, let $\theta_i | G \sim G$. As it has been explained earlier with positive probability there are repeated values among θ_is. Therefore DP induces a partitioning among the n observations.
The Chinese Restaurant Process (CRP)[5] explains this partitioning of natural numbers.

[5] J. Pitman, Combinatorial stochastic processes, Technical Report 621, U. C. Berkeley Department of Statistics, August 2002.

Chinese Restaurant Process

Assume $G \sim \mathrm{DP}(c, G_0)$, and for $i = 1, \ldots, n$, let $\theta_i | G \sim G$. As it has been explained earlier with positive probability there are repeated values among θ_is. Therefore DP induces a partitioning among the n observations.
The Chinese Restaurant Process (CRP)[5] explains this partitioning of natural numbers.

Chinese Restaurant Process

Set $n = 0$ and $K = 0$.

- Customer 1 enters the restaurant and sits at the 1st table, and orders the dish $\theta_1 \sim G_0$.
  $n = 1,\ K = 1,\ n_1 = 1$.

[5] J. Pitman, Combinatorial stochastic processes, Technical Report 621, U. C. Berkeley Department of Statistics, August 2002.

Chinese Restaurant Process

Assume $G \sim \mathrm{DP}(c, G_0)$, and for $i = 1, ..., n$, let $\theta_i | G \sim G$. As it has been explained earlier with positive probability there are repeated values among θ_is. Therefore DP induces a partitioning among the n observations.
The Chinese Restaurant Process (CRP)[5] explains this partitioning of natural numbers.

Chinese Restaurant Process

Set $n = 0$ and $K = 0$.

- Customer 1 enters the restaurant and sits at the 1st table, and orders the dish $\theta_1 \sim G_0$.
  $n = 1,\ K = 1,\ n_1 = 1$.
- For $i = 2, 3, ..$:
  Customer i,
  - with probability $\frac{c}{c+n}$ sits at a new table (the $K+1$-st table), and orders a new dish $\theta_{K+1} \sim G_0$.
    $n \leftarrow n+1,\ K \leftarrow K+1,\ n_{K+1} = 1$.

[5] J. Pitman, Combinatorial stochastic processes, Technical Report 621, U. C. Berkeley Department of Statistics, August 2002.

Chinese Restaurant Process

Assume $G \sim \mathrm{DP}(c, G_0)$, and for $i = 1, ..., n$, let $\theta_i | G \sim G$. As it has been explained earlier with positive probability there are repeated values among θ_is. Therefore DP induces a partitioning among the n observations.
The Chinese Restaurant Process (CRP)[5] explains this partitioning of natural numbers.

Chinese Restaurant Process

Set $n = 0$ and $K = 0$.

- Customer 1 enters the restaurant and sits at the 1st table, and orders the dish $\theta_1 \sim G_0$.
  $n = 1,\ K = 1,\ n_1 = 1$.
- For $i = 2, 3, ..$:

  Customer i,
  - with probability $\frac{c}{c+n}$ sits at a new table (the $K+1$-st table), and orders a new dish $\theta_{K+1} \sim G_0$.
    $n \leftarrow n+1,\ K \leftarrow K+1,\ n_{K+1} = 1$.
  - chooses one of the K already occupied tables with probability $\frac{n_k}{c+n}$, and orders the dish chosen by customers on this table θ_k.
    $n \leftarrow n+1,\ n_k \leftarrow n_k + 1$.

[5] J. Pitman, Combinatorial stochastic processes, Technical Report 621, U. C. Berkeley Department of Statistics, August 2002.

Number of Clusters (Tables)

Obviously the number of occupied tables (number of data clusters) increases with the number of customers (observations). In other words the model is able to self-adapt to more complex models when the new data arrives (because it is a non-parametric model!), and we do not need to set the model complexity a priori.

Number of Clusters (Tables)

Obviously the number of occupied tables (number of data clusters) increases with the number of customers (observations). In other words the model is able to self-adapt to more complex models when the new data arrives (because it is a non-parametric model!), and we do not need to set the model complexity a priori.

Let U_i be a binary random variable, indicating if the i-th customer occupies a new table ($U_i = 1$), or sits at one of the previous tables ($U_i = 0$).

Number of Clusters (Tables)

Obviously the number of occupied tables (number of data clusters) increases with the number of customers (observations). In other words the model is able to self-adapt to more complex models when the new data arrives (because it is a non-parametric model!), and we do not need to set the model complexity a priori.

Let U_i be a binary random variable, indicating if the i-th customer occupies a new table ($U_i = 1$), or sits at one of the previous tables ($U_i = 0$). Obviously the number of clusters is:

$$K = \sum_{i=1}^{N} U_i$$

Number of Clusters (Tables)

Obviously the number of occupied tables (number of data clusters) increases with the number of customers (observations). In other words the model is able to self-adapt to more complex models when the new data arrives (because it is a non-parametric model!), and we do not need to set the model complexity a priori.

Let U_i be a binary random variable, indicating if the i-th customer occupies a new table ($U_i = 1$), or sits at one of the previous tables ($U_i = 0$). Obviously the number of clusters is:

$$K = \sum_{i=1}^{N} U_i$$

and

$$\mathbb{E}(K) = \mathbb{E}\left(\sum_{i=1}^{N} U_i\right) = \sum_{i=1}^{N} \mathbb{E}(U_i) = \sum_{i=1}^{N} \frac{c}{c+i-1} = k(N, c)$$

Number of Clusters (Tables)

Obviously the number of occupied tables (number of data clusters) increases with the number of customers (observations). In other words the model is able to self-adapt to more complex models when the new data arrives (because it is a non-parametric model!), and we do not need to set the model complexity a priori.

Let U_i be a binary random variable, indicating if the i-th customer occupies a new table ($U_i = 1$), or sits at one of the previous tables ($U_i = 0$). Obviously the number of clusters is:

$$K = \sum_{i=1}^{N} U_i$$

and

$$\mathbb{E}(K) = \mathbb{E}\left(\sum_{i=1}^{N} U_i\right) = \sum_{i=1}^{N} \mathbb{E}(U_i) = \sum_{i=1}^{N} \frac{c}{c+i-1} = k(N,c)$$

It can be shown that as $N \to +\infty$, we have:

$$\mathbb{E}(K) = k(N,c) \quad \in \quad \mathcal{O}(\ln N)$$

that means the number of clusters is proportional to the logarithm of the observations.

SBP: Beyond DP

As we have seen earlier the DP has a SB representation:

$$F(\theta) \quad = \quad \sum_{j=1}^{+\infty} w_j \delta_{\theta_j}(\theta)$$

where $\theta_i \overset{iid}{\sim} G_0$, $v_i \overset{iid}{\sim} \text{Beta}(1, c)$, $w_1 = v_1$ and $w_k = v_k \prod_{j=1}^{k-1}(1 - v_j)$ for $k > 1$.

[6] Ishwaran, H., James, L. F., 2001. Gibbs sampling methods for stick-breaking priors, Journal of the American Statistical Association 96, 161173.

SBP: Beyond DP

As we have seen earlier the DP has a SB representation:

$$F(\theta) \quad = \quad \sum_{j=1}^{+\infty} w_j \delta_{\theta_j}(\theta)$$

where $\theta_i \overset{iid}{\sim} G_0$, $v_i \overset{iid}{\sim} \text{Beta}(1, c)$, $w_1 = v_1$ and $w_k = v_k \prod_{j=1}^{k-1}(1 - v_j)$ for $k > 1$.

Ishwaran and James [6] have generalized this process to $v_i \overset{iid}{\sim} \text{Beta}(a_i, b_i)$, and have shown that the process is well-defined ($\sum_{i=1}^{+\infty} w_i = 1$ a.s.) if:

$$\sum_{i=1}^{+\infty} \ln(1 + a_i/b_i) \quad = \quad +\infty$$

[6] Ishwaran, H., James, L. F., 2001. Gibbs sampling methods for stick-breaking priors, Journal of the American Statistical Association 96, 161173.

SBP: Beyond DP

As we have seen earlier the DP has a SB representation:

$$F(\theta) = \sum_{j=1}^{+\infty} w_j \delta_{\theta_j}(\theta)$$

where $\theta_i \overset{iid}{\sim} G_0$, $v_i \overset{iid}{\sim} \text{Beta}(1, c)$, $w_1 = v_1$ and $w_k = v_k \prod_{j=1}^{k-1}(1 - v_j)$ for $k > 1$.

Ishwaran and James [6] have generalized this process to $v_i \overset{iid}{\sim} \text{Beta}(a_i, b_i)$, and have shown that the process is well-defined ($\sum_{i=1}^{+\infty} w_i = 1$ a.s.) if:

$$\sum_{i=1}^{+\infty} \ln(1 + a_i/b_i) = +\infty$$

Note that DP is a special case with $a_i = 1$ and $b_i = c$.

[6] Ishwaran, H., James, L. F., 2001. Gibbs sampling methods for stick-breaking priors, Journal of the American Statistical Association 96, 161173.

SBP: Beyond DP

As we have seen earlier the DP has a SB representation:

$$F(\theta) = \sum_{j=1}^{+\infty} w_j \delta_{\theta_j}(\theta)$$

where $\theta_i \overset{iid}{\sim} G_0$, $v_i \overset{iid}{\sim} \text{Beta}(1, c)$, $w_1 = v_1$ and $w_k = v_k \prod_{j=1}^{k-1}(1 - v_j)$ for $k > 1$.

Ishwaran and James [6] have generalized this process to $v_i \overset{iid}{\sim} \text{Beta}(a_i, b_i)$, and have shown that the process is well-defined ($\sum_{i=1}^{+\infty} w_i = 1$ a.s.) if:

$$\sum_{i=1}^{+\infty} \ln(1 + a_i/b_i) = +\infty$$

Note that DP is a special case with $a_i = 1$ and $b_i = c$.

Another special case is $a_i = 1 - a$ and $b_i = c + ai$, that is called PoissonDirichlet or Pitman-Yor process.

[6] Ishwaran, H., James, L. F., 2001. Gibbs sampling methods for stick-breaking priors, Journal of the American Statistical Association 96, 161173.

Finite Mixture Models: A Simple Example

Consider a mixture of two Normal distributions with unknown mixing proportion p_1, and known variances:

$$Y_i | p_1, \mu_1, \mu_2 \quad \sim \quad p_1 \, \mathcal{N}(\mu_1, 1) + (1 - p_1) \, \mathcal{N}(\mu_2, 1)$$

where the unknown parameters are p_1, μ_1 and μ_2. Moreover assume y_1, ..., y_N are the observed data.

Finite Mixture Models: A Simple Example

Consider a mixture of two Normal distributions with unknown mixing proportion p_1, and known variances:

$$Y_i | p_1, \mu_1, \mu_2 \quad \sim \quad p_1 \; \mathcal{N}(\mu_1, 1) + (1 - p_1) \; \mathcal{N}(\mu_2, 1)$$

where the unknown parameters are p_1, μ_1 and μ_2. Moreover assume y_1, ..., y_N are the observed data.

We may put a Beta prior on p_1, and (independent) Normal priors on μ_1 and μ_2,

$$p(p_1, \mu_1, \mu_2) \quad = \quad p(p_1) \; p(\mu_1) \; p(\mu_2)$$

and sample the posterior

$$p(p_1, \mu_1, \mu_2 | \mathbf{y}) \quad \propto \quad p(p_1, \mu_1, \mu_2) p(\mathbf{y} | p_1, \mu_1, \mu_2)$$

by MCMC.

Finite Mixture Models: A Simple Example

The model can be re-formulated by introducing indicator variables d_1, ..., d_N that are $\{1,2\}$-valued random variables:

$$\begin{aligned} Y_i|\mu_1,\mu_2,d_i &\sim \mathcal{N}(\mu_{d_i},1) \\ d_i &\sim \text{Discrete}(\{1,2\},(p_1,1-p_1)) \end{aligned}$$

Finite Mixture Models: A Simple Example

The model can be re-formulated by introducing indicator variables d_1, ..., d_N that are $\{1, 2\}$-valued random variables:

$$\begin{aligned} Y_i | \mu_1, \mu_2, d_i &\sim \mathcal{N}(\mu_{d_i}, 1) \\ d_i &\sim \text{Discrete}(\{1, 2\}, (p_1, 1 - p_1)) \end{aligned}$$

We can also generalize the model to a mixture of K components:

$$\begin{aligned} Y_i | \theta_1, ..., \theta_K, d_i &\sim H(\theta_{d_i}) \\ d_i &\sim \text{Discrete}(\{1, ..., K\}, (p_1, ..., p_K)) \end{aligned}$$

where $\sum_{i=1}^K p_i = 1$, and $H(\theta)$ is a distribution parameterized by the vector of parameters θ.

Finite Mixture Models: A Simple Example

The model can be re-formulated by introducing indicator variables d_1, ..., d_N that are $\{1, 2\}$-valued random variables:

$$\begin{aligned} Y_i | \mu_1, \mu_2, d_i &\sim \mathcal{N}(\mu_{d_i}, 1) \\ d_i &\sim \text{Discrete}(\{1, 2\}, (p_1, 1 - p_1)) \end{aligned}$$

We can also generalize the model to a mixture of K components:

$$\begin{aligned} Y_i | \theta_1, ..., \theta_K, d_i &\sim H(\theta_{d_i}) \\ d_i &\sim \text{Discrete}(\{1, ..., K\}, (p_1, ..., p_K)) \end{aligned}$$

where $\sum_{i=1}^{K} p_i = 1$, and $H(\theta)$ is a distribution parameterized by the vector of parameters θ.

Obviously a natural choice for the prior of $(p_1, ..., p_K)$ is the Dirichlet distribution.

Finite Mixture Models

Good News:

- Any distribution can be approximated by a location-scale mixture of Normals with enough number of components. This is true for both univariate and multivariate distributions.
- Univariate unimodel distributions can be approximated by a mixture of uniforms with enough number of components.
- We can use mixtures for density estimation, clustering, classification, ...

Finite Mixture Models

Good News:

- Any distribution can be approximated by a location-scale mixture of Normals with enough number of components. This is true for both univariate and multivariate distributions.
- Univariate unimodel distributions can be approximated by a mixture of uniforms with enough number of components.
- We can use mixtures for density estimation, clustering, classification, ...

Bad News:
We do not know how many components is "enough"!

- More than enough: Risk of overfitting!
  Good fitting in terms of the likelihood, but poor generalization.
- Less than enough: The model would not be complex enough to exploit the patterns and information available in the data.

Finite Mixture Models

Good News:

- Any distribution can be approximated by a location-scale mixture of Normals with enough number of components. This is true for both univariate and multivariate distributions.
- Univariate unimodel distributions can be approximated by a mixture of uniforms with enough number of components.
- We can use mixtures for density estimation, clustering, classification, ...

Bad News:
We do not know how many components is "enough"!

- More than enough: Risk of overfitting!
  Good fitting in terms of the likelihood, but poor generalization.
- Less than enough: The model would not be complex enough to exploit the patterns and information available in the data.

Bayesian non-parametrics provides a solution to this problem: We put a prior on the model's complexity, and using the data do the inference on the model complexity too!

Finite Mixture Models: An Alternative Formulation

The finite mixture model

$$Y_i|\theta_1, ..., \theta_K, \mathbf{p} \quad \sim \quad \sum_{i=1}^{K} p_i \; H(\theta_i)$$

can be re-formulated in the following way:

$$\begin{aligned} Y_i|\theta_1, ..., \theta_K \quad &\sim \quad \int_{\Theta} H(\theta) \; dF(\theta) \\ F(\theta) \quad &= \quad \sum_{i=1}^{K} p_i \; \delta_{\theta_i}(\theta) \end{aligned}$$

Finite Mixture Models: An Alternative Formulation

The finite mixture model

$$Y_i|\theta_1, ..., \theta_K, \mathbf{p} \quad \sim \quad \sum_{i=1}^{K} p_i \; H(\theta_i)$$

can be re-formulated in the following way:

$$\begin{aligned} Y_i|\theta_1, ..., \theta_K \quad &\sim \quad \int_{\Theta} H(\theta) \; dF(\theta) \\ F(\theta) \quad &= \quad \sum_{i=1}^{K} p_i \; \delta_{\theta_i}(\theta) \end{aligned}$$

Note that the mixing distribution F is a random discrete measure with finite point masses.

Finite Mixture Models: An Alternative Formulation

The finite mixture model

$$Y_i|\theta_1, ..., \theta_K, \mathbf{p} \quad \sim \quad \sum_{i=1}^{K} p_i \; H(\theta_i)$$

can be re-formulated in the following way:

$$\begin{aligned} Y_i|\theta_1, ..., \theta_K \quad &\sim \quad \int_{\Theta} H(\theta) \; dF(\theta) \\ F(\theta) \quad &= \quad \sum_{i=1}^{K} p_i \; \delta_{\theta_i}(\theta) \end{aligned}$$

Note that the mixing distribution F is a random discrete measure with finite point masses. Moreover by putting priors on p_is and θ_is, we put a prior on this random measure.

Finite Mixture Models: An Alternative Formulation

The finite mixture model

$$Y_i|\theta_1, ..., \theta_K, \mathbf{p} \quad \sim \quad \sum_{i=1}^{K} p_i \; H(\theta_i)$$

can be re-formulated in the following way:

$$\begin{aligned} Y_i|\theta_1, ..., \theta_K \quad &\sim \quad \int_{\Theta} H(\theta) \; dF(\theta) \\ F(\theta) \quad &= \quad \sum_{i=1}^{K} p_i \; \delta_{\theta_i}(\theta) \end{aligned}$$

Note that the mixing distribution F is a random discrete measure with finite point masses. Moreover by putting priors on p_is and θ_is, we put a prior on this random measure.

This suggests that we may substitute our prior on the random measure F with a more general prior and build a richer family of mixture distributions...

Finite Mixture Models: An Alternative Formulation

The finite mixture model

$$Y_i|\theta_1, \ldots, \theta_K, \mathbf{p} \quad \sim \quad \sum_{i=1}^{K} p_i \; H(\theta_i)$$

can be re-formulated in the following way:

$$\begin{aligned} Y_i|\theta_1, \ldots, \theta_K \quad &\sim \quad \int_{\Theta} H(\theta) \; dF(\theta) \\ F(\theta) \quad &= \quad \sum_{i=1}^{K} p_i \; \delta_{\theta_i}(\theta) \end{aligned}$$

Note that the mixing distribution F is a random discrete measure with finite point masses. Moreover by putting priors on p_is and θ_is, we put a prior on this random measure.

This suggests that we may substitute our prior on the random measure F with a more general prior and build a richer family of mixture distributions...

Our natural candidate for the prior on F is Dirichlet Process!

Dirichlet Process Mixture (DPM)

As it has already been explained, the draws of from a DP are discrete distributions a.s. (remember the SB representation of the DP). Therefore they might be an inappropriate choice for modelling continuous distributions.

Dirichlet Process Mixture (DPM)

As it has already been explained, the draws of from a DP are discrete distributions a.s. (remember the SB representation of the DP). Therefore they might be an inappropriate choice for modelling continuous distributions.

Dirichlet process mixture model is a potential solution to this problem.

Dirichlet Process Mixture (DPM)

As it has already been explained, the draws of from a DP are discrete distributions a.s. (remember the SB representation of the DP). Therefore they might be an inappropriate choice for modelling continuous distributions.

Dirichlet process mixture model is a potential solution to this problem.

Dirichlet Process Mixture

Consider the following hierarchical model:

$$\begin{aligned} Y_i|\theta_i &\overset{iid}{\sim} H(\theta_i) \\ \theta_i|F &\sim F \end{aligned}$$

Using a DP as the prior of the unknown distribution F,

$$F \sim DP(c, F_0)$$

results a model that is called Dirichlet Process Mixture (DPM)

DPM: Limit of Finite Mixtures

Consider the finite mixture model with K components:

$$Y_i|\theta_1, ..., \theta_K, \mathbf{p} \quad \sim \quad \sum_{i=1}^{K} p_i \ H(\theta_i)$$

with the following priors:

$$\begin{aligned} \theta_i &\sim F_0 \\ \mathbf{p} &\sim \text{Dirichlet}(c/K, ..., c/K) \end{aligned}$$

[7] H. Ishwaran and M. Zarepour, Exact and approximate sumrepresentations for the Dirichlet process, Canadian Journal of Statistics, 30, pp. 269-283, 2002.

DPM: Limit of Finite Mixtures

Consider the finite mixture model with K components:

$$Y_i|\theta_1, ..., \theta_K, \mathbf{p} \quad \sim \quad \sum_{i=1}^{K} p_i \; H(\theta_i)$$

with the following priors:

$$\begin{aligned} \theta_i &\sim F_0 \\ \mathbf{p} &\sim \text{Dirichlet}(c/K, ..., c/K) \end{aligned}$$

It can be shown[7] that the limit of this finite mixture model when $K \to +\infty$, is the following DPM:

$$\begin{aligned} Y_i|\theta_i &\overset{iid}{\sim} H(\theta_i) \\ \theta_i|F &\sim F \\ F &\sim DP(c, F_0) \end{aligned}$$

[7] H. Ishwaran and M. Zarepour, Exact and approximate sumrepresentations for the Dirichlet process, Canadian Journal of Statistics, 30, pp. 269-283, 2002.

DPM - Example

Assume the mixture kernel is a univariate Normal distribution, with $\theta = (\mu, \sigma^2)$:

$$H(\theta) = \mathcal{N}(\mu, \sigma^2)$$

[8] M. D. Escobar, M. West, Bayesian density estimation and inference using mixtures, J. Am. Stat. Assoc., 90, pp. 577-588, 1995.

DPM - Example

Assume the mixture kernel is a univariate Normal distribution, with $\theta = (\mu, \sigma^2)$:

$$H(\theta) = \mathcal{N}(\mu, \sigma^2)$$

Therefore the base measure of the DP should be a distribution on $\mathbb{R} \times \mathbb{R}^+$. A convenient choice is the normal inverse Gamma:

$$F_0(\theta) = \text{N-IG}(\theta; \alpha, \beta, \lambda, \nu)$$

[8] M. D. Escobar, M. West, Bayesian density estimation and inference using mixtures, J. Am. Stat. Assoc., 90, pp. 577-588, 1995.

DPM - Example

Assume the mixture kernel is a univariate Normal distribution, with $\theta = (\mu, \sigma^2)$:

$$H(\theta) \quad = \quad \mathcal{N}(\mu, \sigma^2)$$

Therefore the base measure of the DP should be a distribution on $\mathbb{R} \times \mathbb{R}^+$. A convenient choice is the normal inverse Gamma:

$$F_0(\theta) \quad = \quad \text{N-IG}(\theta; \alpha, \beta, \lambda, \nu)$$

The inference in this model has been addressed in Escobar and West 1995[8]. It is important to mention that we may put priors on the hyperparameters of the DP (the parameters of the base measure α, β, λ and ν, and the concentration parameter c) and sample them as well.

[8] M. D. Escobar, M. West, Bayesian density estimation and inference using mixtures, J. Am. Stat. Assoc., 90, pp. 577-588, 1995.

DPM - Example

Assume the mixture kernel is a univariate Normal distribution, with $\theta = (\mu, \sigma^2)$:

$$H(\theta) = \mathcal{N}(\mu, \sigma^2)$$

Therefore the base measure of the DP should be a distribution on $\mathbb{R} \times \mathbb{R}^+$. A convenient choice is the normal inverse Gamma:

$$F_0(\theta) = \text{N-IG}(\theta; \alpha, \beta, \lambda, \nu)$$

The inference in this model has been addressed in Escobar and West 1995[8]. It is important to mention that we may put priors on the hyperparameters of the DP (the parameters of the base measure α, β, λ and ν, and the concentration parameter c) and sample them as well.

In particular the value of the concentration parameter c affects the posterior significantly, and in the absence of prior knowledge we should put a non-informative prior on it.

[8] M. D. Escobar, M. West, Bayesian density estimation and inference using mixtures, J. Am. Stat. Assoc., 90, pp. 577-588, 1995.

DPM - Example

In particular using results of Antoniak 1974[9], we know the PMF of the number of occupied components k:

$$P(k|c,N) = C_N(k)N!c^k \frac{\Gamma(c)}{\Gamma(c+N)}$$

and for large sample sizes, we have:

$$\mathbb{E}(k|c) \approx c \ln\left(\frac{c+N}{c}\right)$$

[9] C. E. Antoniak, Mixtures of Dirichlet Process with Applications to Nonparametric Problems, The Annals of Statistics, 2, pp. 1152–1174

DPM - Example

In particular using results of Antoniak 1974[9], we know the PMF of the number of occupied components k:

$$P(k|c,N) = C_N(k)N!c^k\frac{\Gamma(c)}{\Gamma(c+N)}$$

and for large sample sizes, we have:

$$\mathbb{E}(k|c) \approx c\ln\left(\frac{c+N}{c}\right)$$

Therefore with a fixed sample size, c controls the complexity (the number of occupied components) of the model.

[9] C. E. Antoniak, Mixtures of Dirichlet Process with Applications to Nonparametric Problems, The Annals of Statistics, 2, pp. 1152–1174

DPM - Example

In particular using results of Antoniak 1974[9], we know the PMF of the number of occupied components k:

$$P(k|c,N) = C_N(k)N!c^k \frac{\Gamma(c)}{\Gamma(c+N)}$$

and for large sample sizes, we have:

$$\mathbb{E}(k|c) \approx c \ln\left(\frac{c+N}{c}\right)$$

Therefore with a fixed sample size, c controls the complexity (the number of occupied components) of the model.

Therefore it is important to put a rather vague prior on c (a possible choice is a Gamma distribution), and sample it in a MCMC simulation.

[9] C. E. Antoniak, Mixtures of Dirichlet Process with Applications to Nonparametric Problems, The Annals of Statistics, 2, pp. 1152–1174

DPM: CRP Representation

By Introducing indicator variables d_1, ..., and d_N, we can represent the DPM in the following way:

$$
\begin{aligned}
Y_i | d_i, \theta_1, \theta_2, \ldots &\sim H(\theta_{d_i}) \\
\theta_i &\sim F_0 \\
\mathbf{d} &\sim \mathrm{CRP}(c)
\end{aligned}
$$

[10] R. M. Neal, Markov chain sampling methods for Dirichlet process mixture models, Journal of Computational and Graphical Statistics, 9(2), pp. 249-265, 2000.

DPM: CRP Representation

By Introducing indicator variables d_1, ..., and d_N, we can represent the DPM in the following way:

$$\begin{aligned} Y_i | d_i, \theta_1, \theta_2, \ldots &\sim H(\theta_{d_i}) \\ \theta_i &\sim F_0 \\ \mathbf{d} &\sim \mathrm{CRP}(c) \end{aligned}$$

Based on this representation, more efficient MCMC sampling schemes can be designed, for example see Neal 2000[10].

[10] R. M. Neal, Markov chain sampling methods for Dirichlet process mixture models, Journal of Computational and Graphical Statistics, 9(2), pp. 249-265, 2000.

DPM: SB Representation

Exploiting the stick-breaking representation of DP, we can re-formulate the DPM in the following way:

$$
\begin{aligned}
Y_i | \mathbf{w}, \theta_1, \theta_2, \ldots &\sim \sum_{i=1}^{+\infty} w_i \; H(\theta_i) \\
\theta_i &\sim F_0 \\
\mathbf{w} &\sim \mathrm{GEM}(c)
\end{aligned}
$$

[11] S. Walker, Sampling the Dirichlet mixture model with slices, Commun. Stat., Simul. Comput. 36, pp. 45-54, 2007.

[12] M. Kalli, J. Griffin, S. Walker, Slice Sampling Mixture Models, Statistics and Computing, 21, pp. 93-105, 2011.

DPM: SB Representation

Exploiting the stick-breaking representation of DP, we can re-formulate the DPM in the following way:

$$
\begin{aligned}
Y_i|\mathbf{w}, \theta_1, \theta_2, \ldots &\sim \sum_{i=1}^{+\infty} w_i \; H(\theta_i) \\
\theta_i &\sim F_0 \\
\mathbf{w} &\sim \mathrm{GEM}(c)
\end{aligned}
$$

Using this representation of the DPM, efficient and easy to implement slice sampler could be designed (for both conjugate and non-conjugate models). For example see Walker 2007[11] and Kalli, Griffin, and Walker 2011[12].

[11] S. Walker, Sampling the Dirichlet mixture model with slices, Commun. Stat., Simul. Comput. 36, pp. 45-54, 2007.

[12] M. Kalli, J. Griffin, S. Walker, Slice Sampling Mixture Models, Statistics and Computing, 21, pp. 93-105, 2011.