Bayesian machine learning

Bayesian nonparametrics

Julyan Arbel

Statify team, Inria Grenoble Rhône-Alpes & Univ. Grenoble-Alpes, France
✉ julyan.arbel@inria.fr 🖱 www.julyanarbel.com

http://github.com/rbardenet/bml-course

Outline

Outline

Outline

Outline

Outline

Parametric versus nonparametric

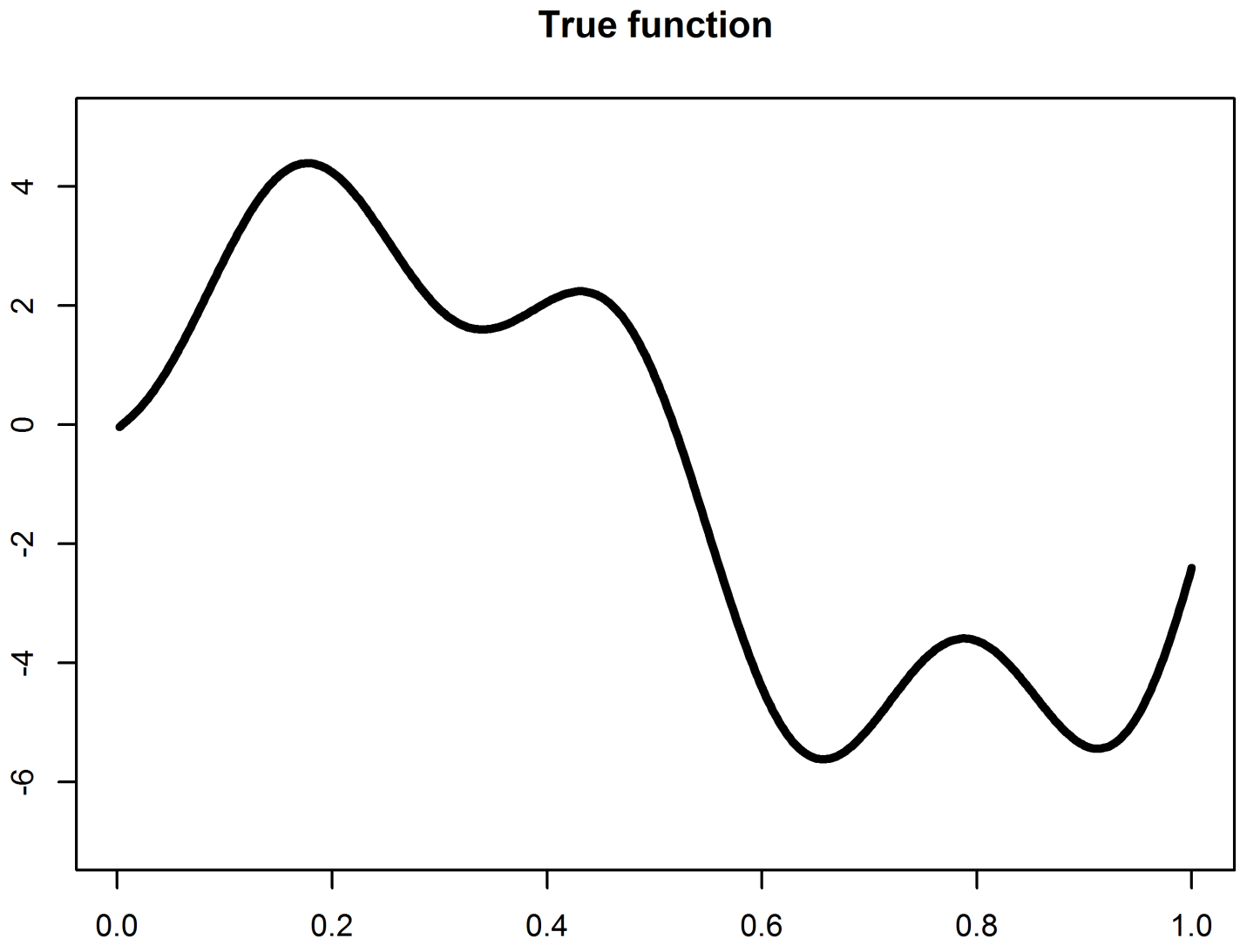
Parametric models

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

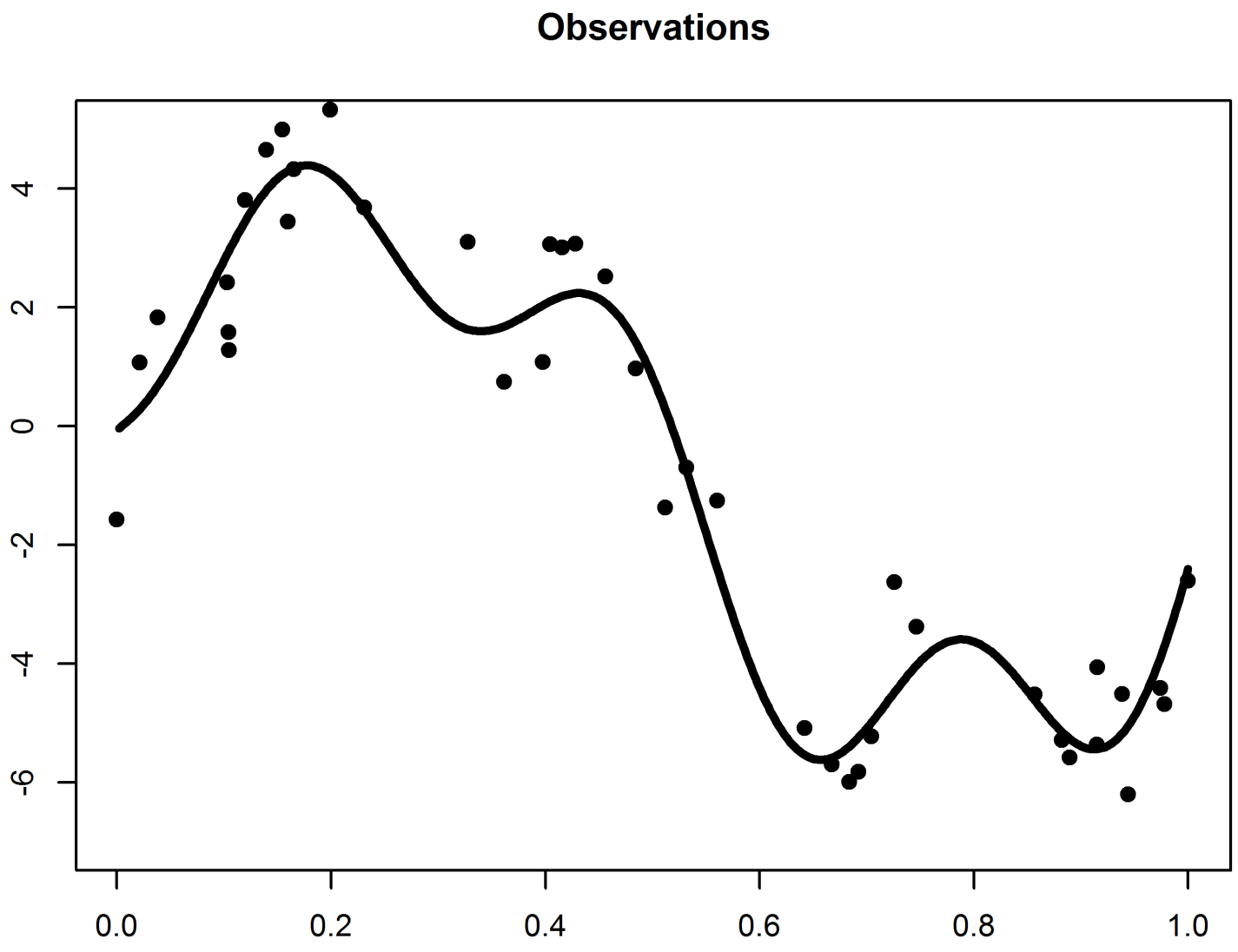
Nonparametric models

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

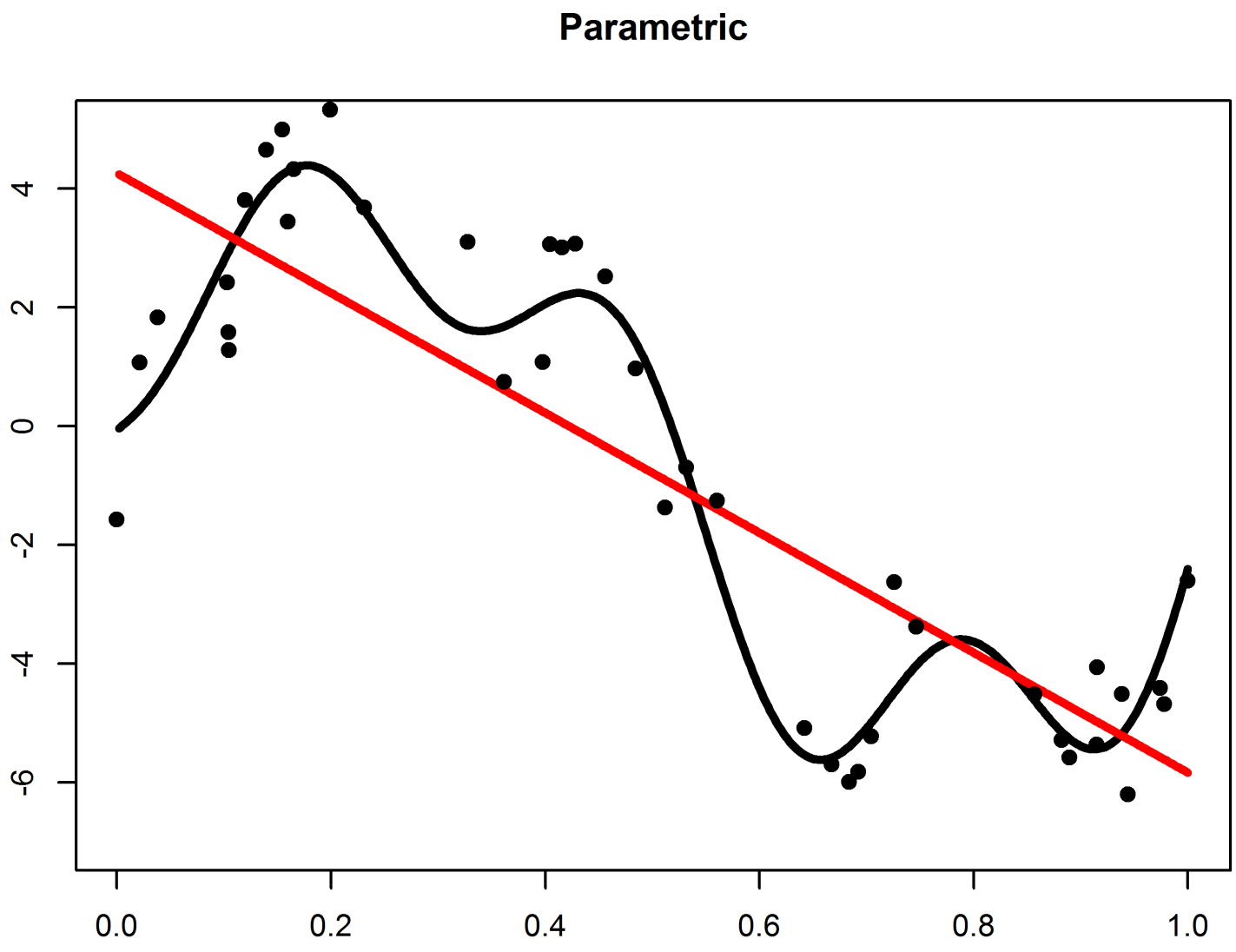
Underlying function



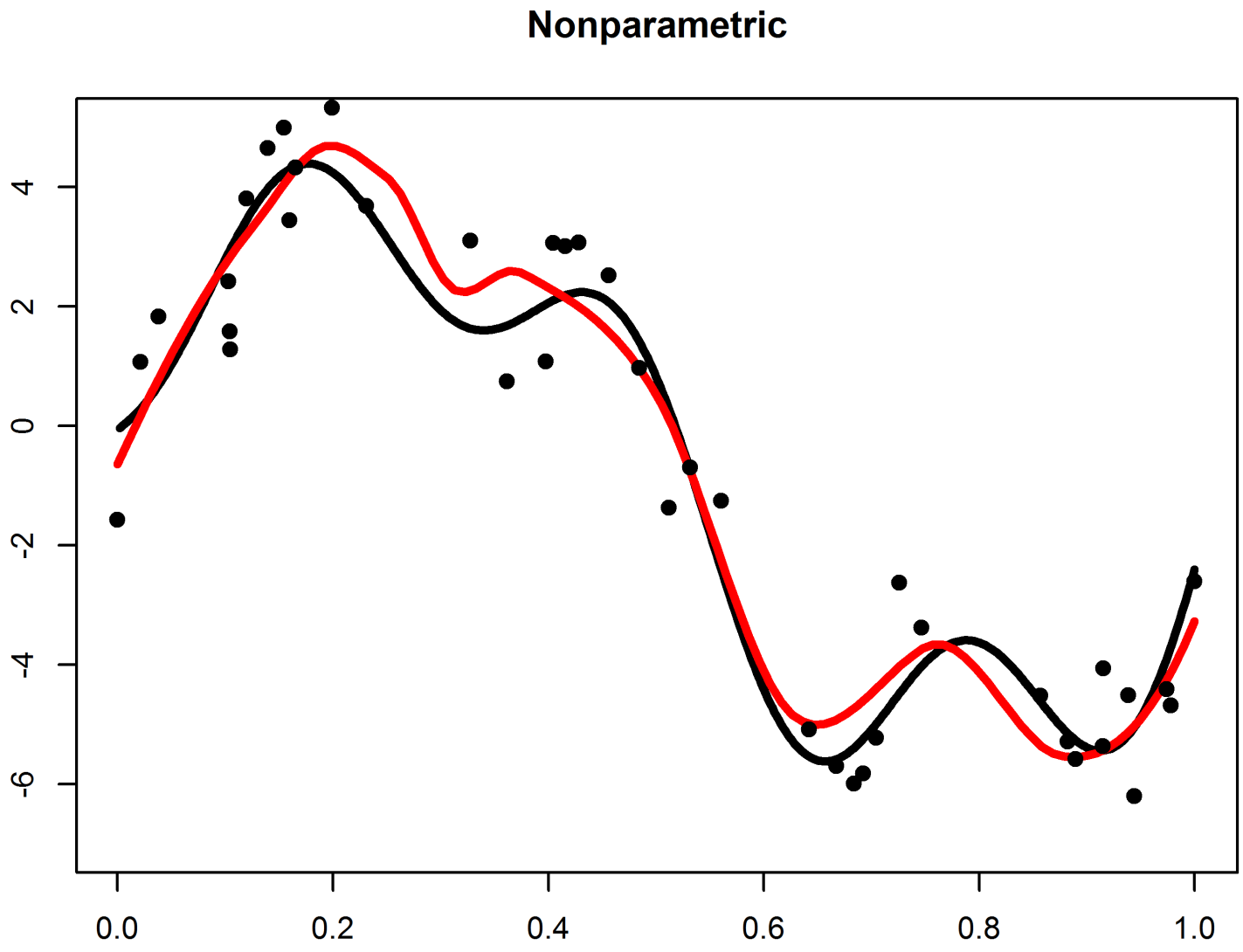
Data



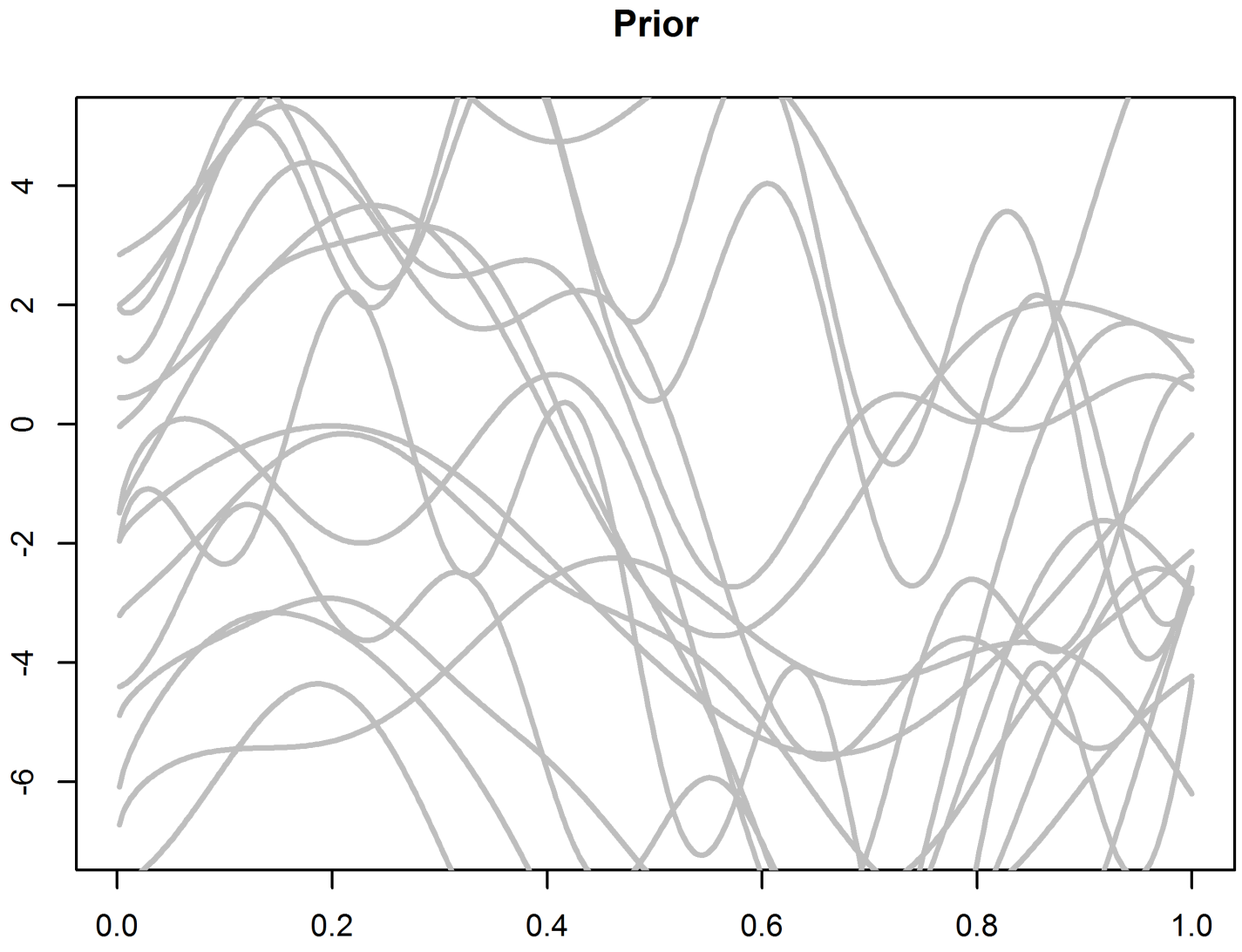
Parametric fitting



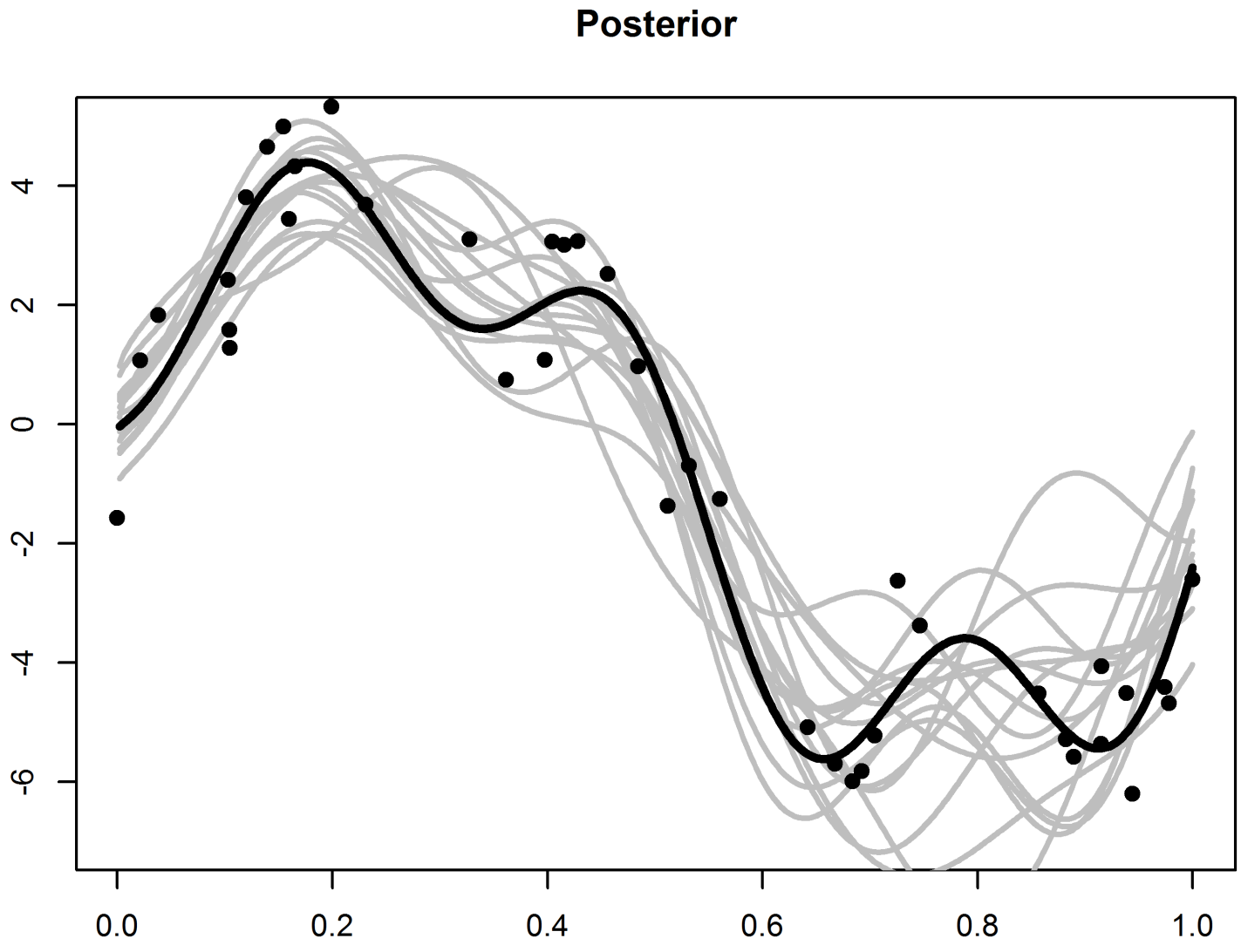
Nonparametric fitting



Prior



Posterior

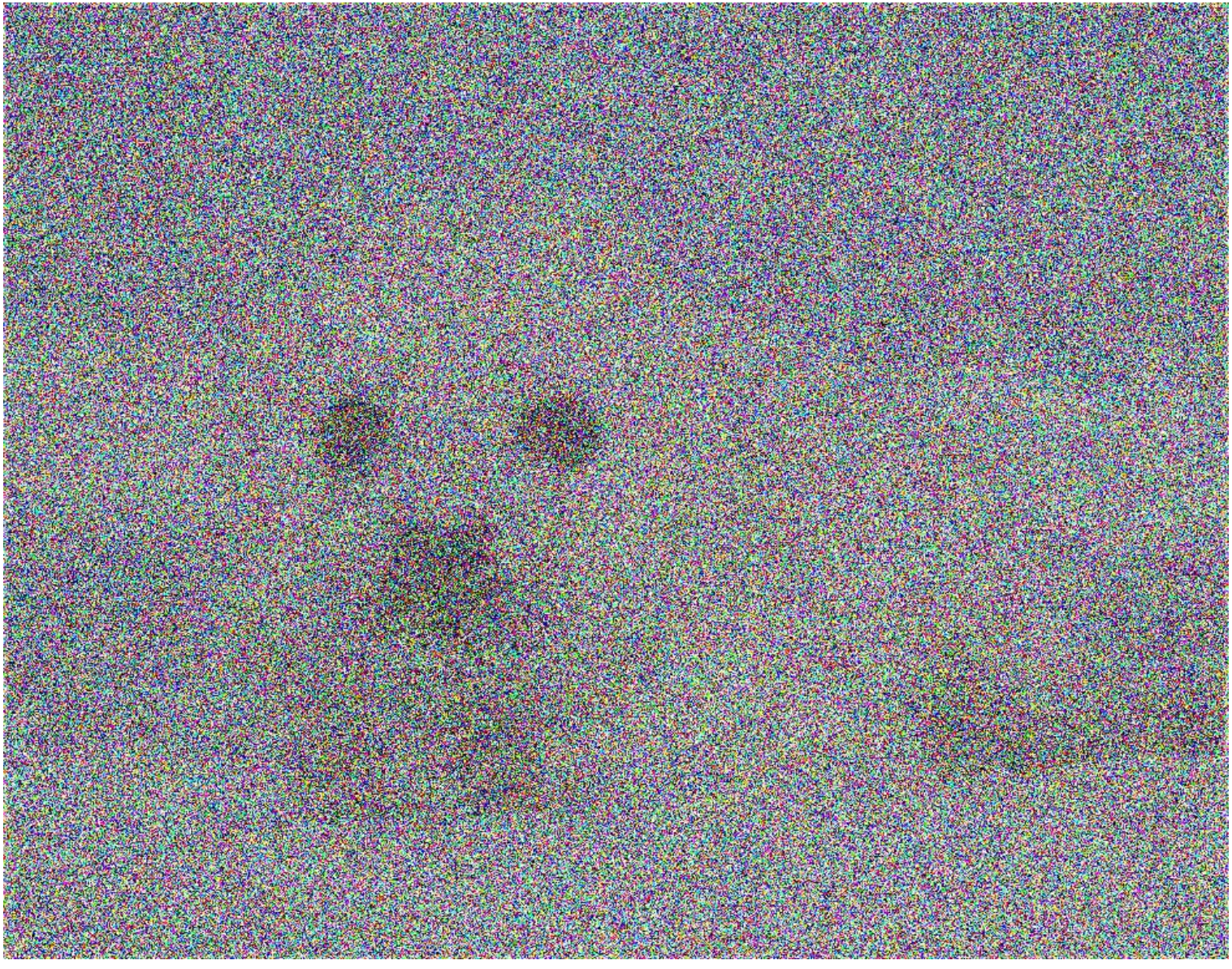


Parametric versus nonparametric

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

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

Noisy picture



Parametric

Nonparametric

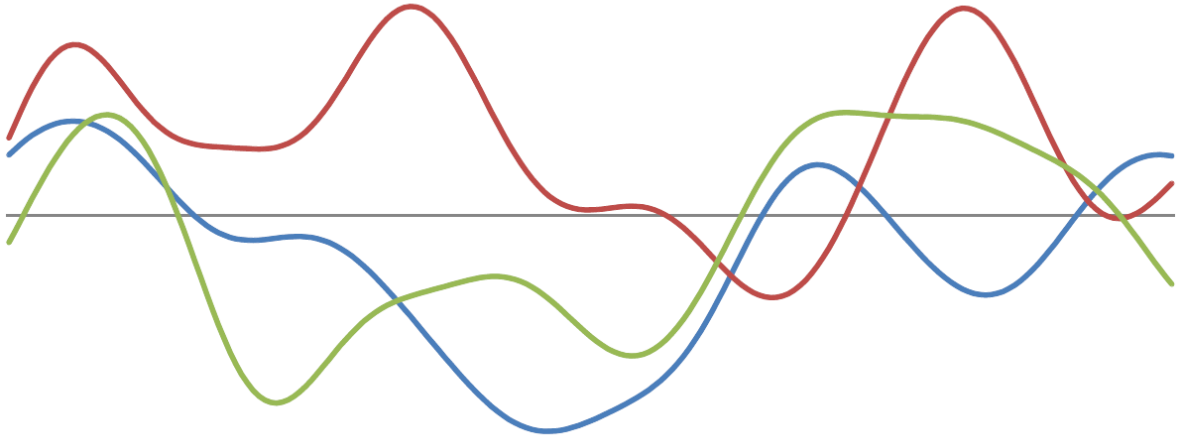
Bayesian nonparametric priors

Two main categories of priors depending on parameter spaces

Two main categories of priors depending on parameter spaces

Spaces of functions
random functions

- Continuous stochastic processes
  e.g. Gaussian processes
- Random basis expansions
- Random densities (expon.)



[Wikipedia]

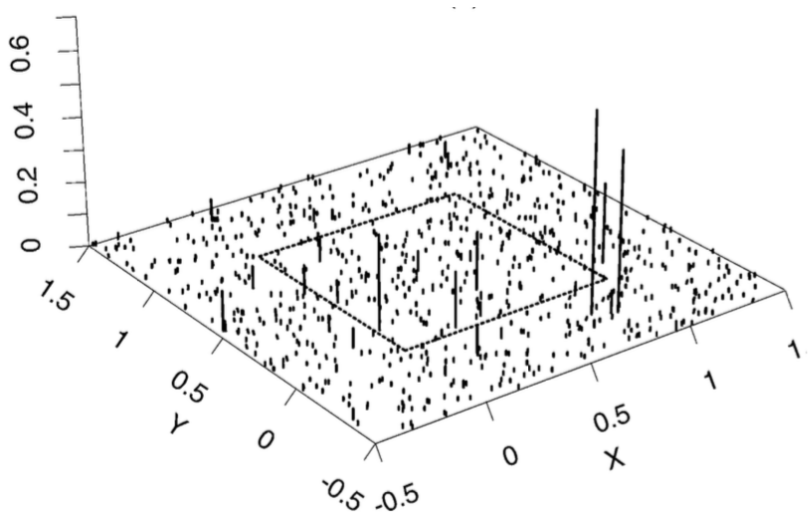
Two main categories of priors depending on parameter spaces

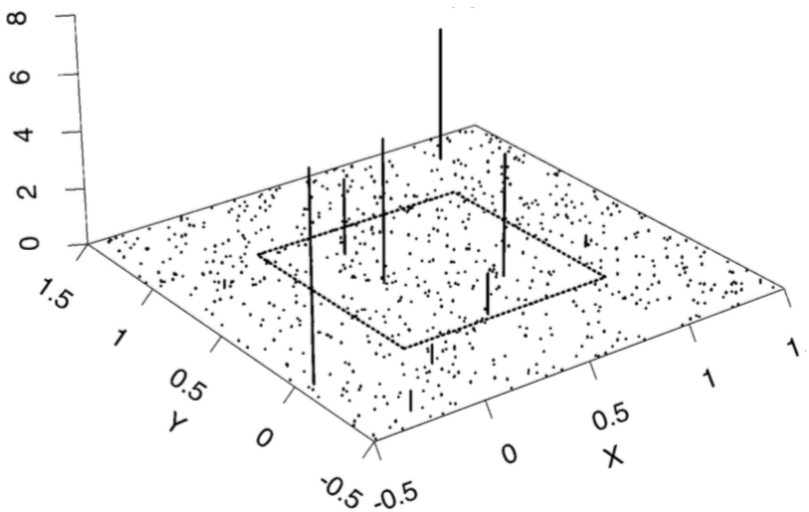
Spaces of functions
random functions

- Continuous stochastic processes
  e.g. Gaussian processes
- Random basis expansions
- Random densities (expon.)

Spaces of probability measures
random probability measures (RPM)

- Often discrete proba. measures
  Cornerstone: Dirichlet process
  We'll see others: Pitman–Yor, Normalized generalized gamma process, Normalized stable process, Gibbs-type processes, Normalized random measures, etc





(Brix, 1999)

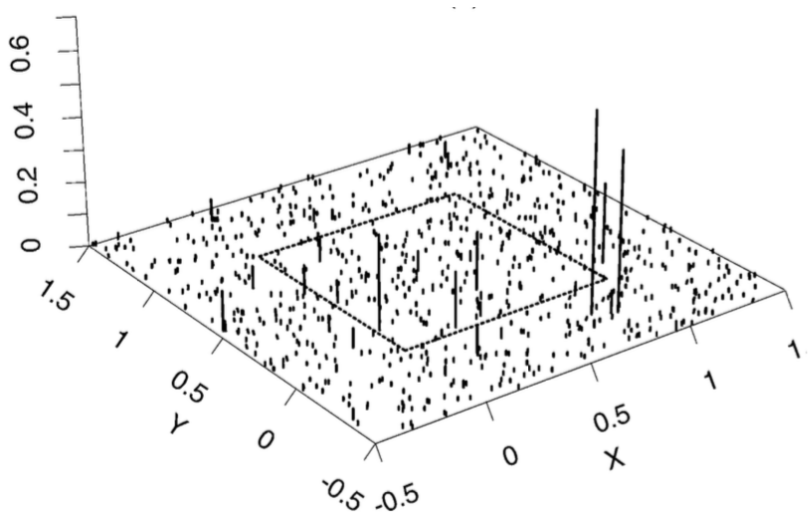
Two main categories of priors depending on parameter spaces

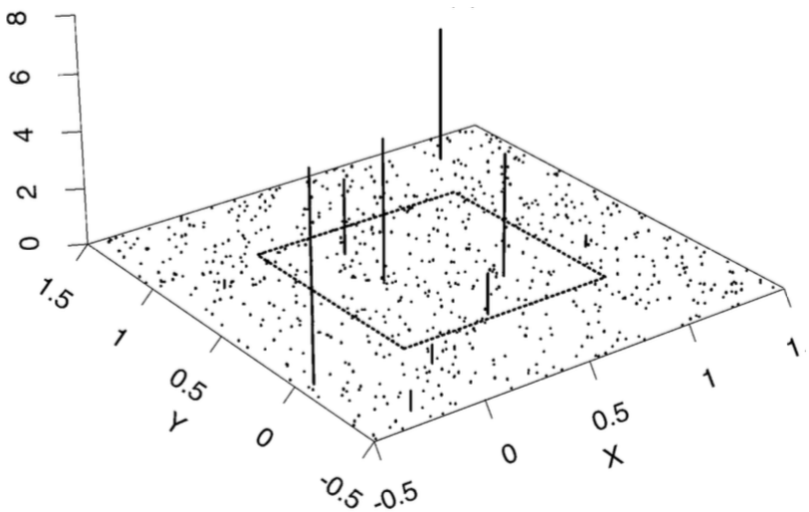
Spaces of functions
random functions

- Continuous stochastic processes
  e.g. Gaussian processes
- Random basis expansions
- Random densities (expon.)

Spaces of probability measures
random probability measures (RPM)

- Often discrete proba. measures
  Cornerstone: Dirichlet process
  We'll see others: Pitman–Yor, Normalized generalized gamma process, Normalized stable process, Gibbs-type processes, Normalized random measures, etc





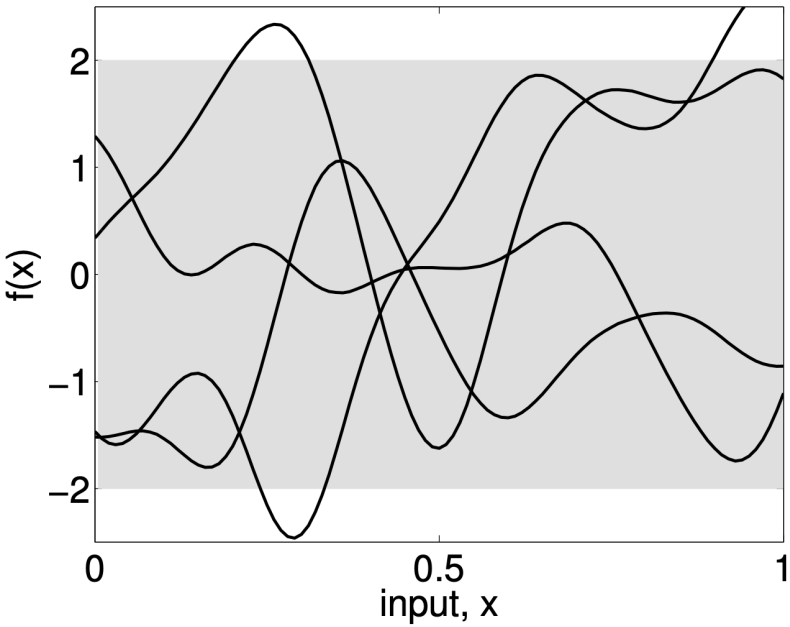
(Brix, 1999)

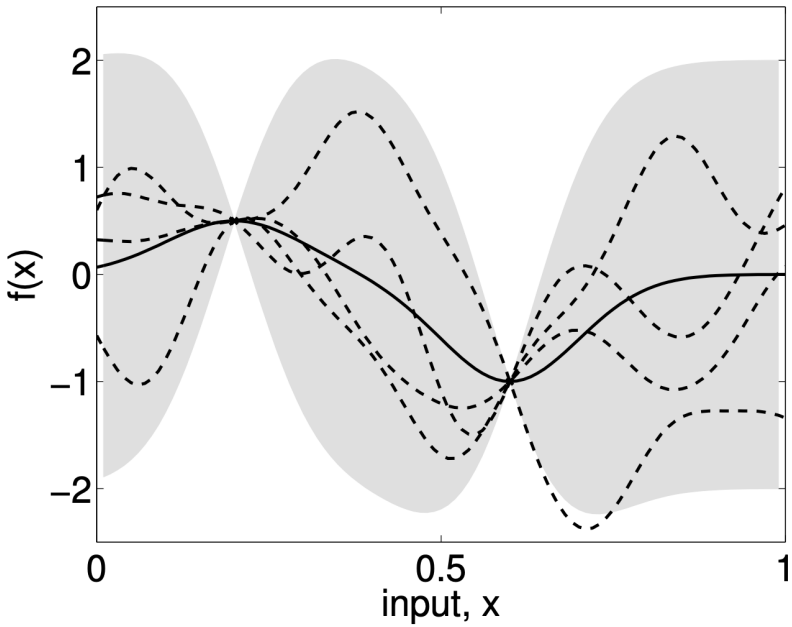
Outline

Outline

What comes to *your* mind when you hear "Gaussian processes"?

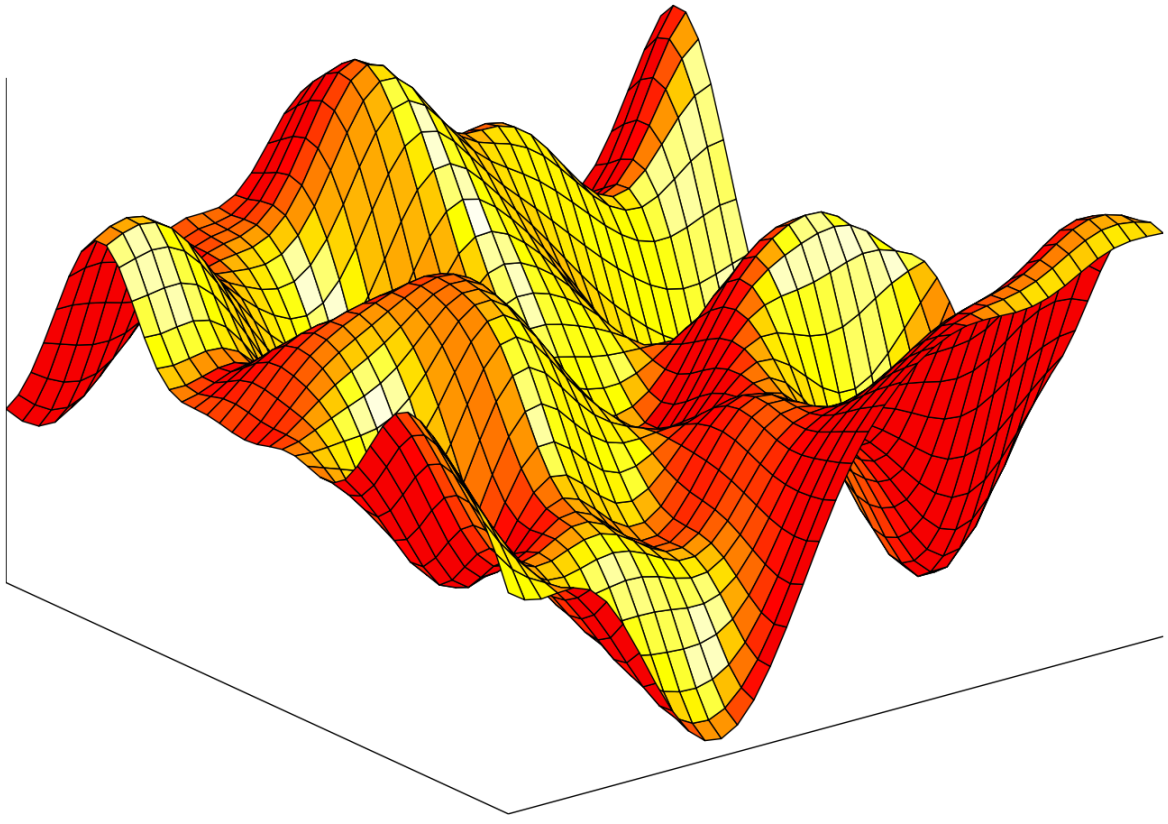
Gaussian processes





From Rasmussen and Williams (2006)

Gaussian processes



From Rasmussen and Williams (2006)

Links with other chapters:

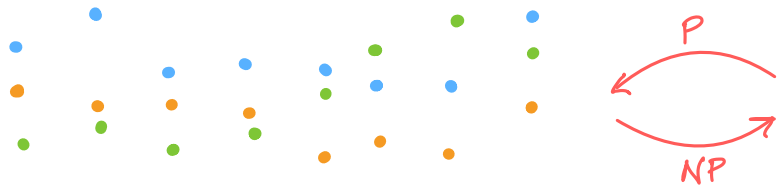
- GPs are used are BNP priors on curves
- As such, the properties of the induced posterior are studied in the section on asymptotics
- Wide limit in Bayesian neural networks
- SGD with constant learning rate
- GPs are the nonparametric counterpart of the multivariate Gaussian distribution, just like the Dirichlet process is the nonparametric counterpart of the Dirichlet distribution

Gaussian process, Dirichlet process, and their parametric counterparts

Multivariate Gaussian

random vector

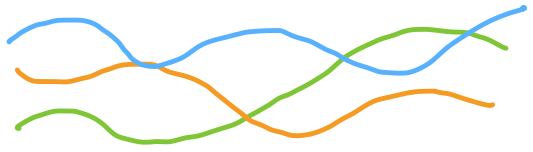
$$\begin{cases} W = (W_1, \dots, W_k) \in \mathbb{R}^k \\ W \sim N(\mu, \Sigma) \end{cases}$$



Gaussian Process

random function

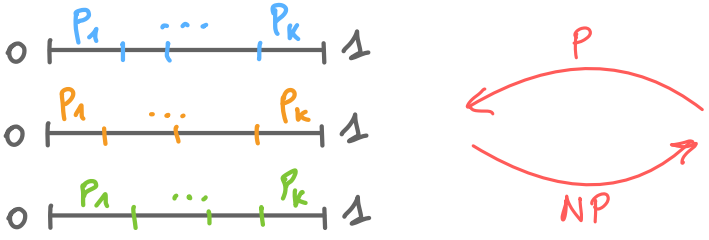
$$\begin{cases} W = (W_t)_{t \in \mathbb{R}}, \ W \sim GP \\ \text{Margins}(W) \sim N(\mu, \Sigma) \end{cases}$$



Dirichlet Distribution

random proba. vector

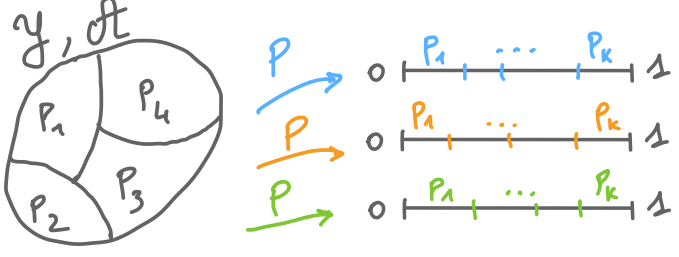
$$\begin{cases} P = (P_1, \dots, P_k) \in S^k \text{ simplex} \\ P \sim \text{Dir}(\alpha), \ \alpha = (\alpha_1, \dots, \alpha_k) \end{cases}$$



Dirichlet Process

random proba. measure

$$\begin{cases} P = (P_A)_{A \in \text{partitions}(Y)}, \ P \sim DP \\ \text{Margins}(P_A) = (P_1, \dots, P_k) \sim \text{Dir} \end{cases}$$



References

- Main reference on GPs: Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. DOI: 10.1.1.86.3414
- GPs in Bayesian inference: Chapter 11 of Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press, 2017
- Chapter 18 on Gaussian processes of Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL: http://probml.github.io/book2

Supervized learning

Two common approaches to supervized learning:

- restrict the class of functions considered, for example only linear functions of the input
- give a prior probability to every possible function, where higher probabilities are given to functions that we consider to be more likely

Definition (Rasmussen and Williams, 2006)

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Definition (Ghosal and Van der Vaart, 2017)

A Gaussian process is a stochastic process $W = (W_t : t \in T)$ indexed by an arbitrary set T such that the vector $(W_{t_1}, \ldots, W_{t_k})$ possesses a multivariate normal distribution, for every $t_i \in T$ and $k \in \mathbb{N}$. A Gaussian process W indexed by $\mathbb{R}^d$ is called:

- self-similar of index α if $(W_{\sigma t} : t \in \mathbb{R}^d)$ is distributed like $(\sigma^\alpha W_t : t \in \mathbb{R}^d)$, for every $\sigma > 0$, and
- stationary if $(W_{t+h} : t \in \mathbb{R}^d)$ has the same distribution of $(W_t : t \in \mathbb{R}^d)$, for every $h \in \mathbb{R}^d$.

Definition (Rasmussen and Williams, 2006)

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Definition (Ghosal and Van der Vaart, 2017)

A Gaussian process is a stochastic process $W = (W_t : t \in T)$ indexed by an arbitrary set T such that the vector $(W_{t_1}, \ldots, W_{t_k})$ possesses a multivariate normal distribution, for every $t_i \in T$ and $k \in \mathbb{N}$. A Gaussian process W indexed by $\mathbb{R}^d$ is called:

- self-similar of index α if $(W_{\sigma t} : t \in \mathbb{R}^d)$ is distributed like $(\sigma^\alpha W_t : t \in \mathbb{R}^d)$, for every $\sigma > 0$, and
- stationary if $(W_{t+h} : t \in \mathbb{R}^d)$ has the same distribution of $(W_t : t \in \mathbb{R}^d)$, for every $h \in \mathbb{R}^d$.

Mean function and covariance kernel

Vectors $(W_{t_1}, \ldots, W_{t_k})$ are called marginals, and their distributions marginal distributions or finite-dimensional distributions

Mean function and covariance kernel

Finite-dimensional distributions are determined by the mean function and covariance kernel, defined by

$$\mu(t) = \mathrm{E}(W_t), \quad K(s,t) = \mathrm{Cov}(W_s, W_t), \quad s,t \in T.$$

Vectors $(W_{t_1}, \ldots, W_{t_k})$ are called marginals, and their distributions marginal distributions or finite-dimensional distributions

Mean function and covariance kernel

Finite-dimensional distributions are determined by the mean function and covariance kernel, defined by

$$\mu(t) = \mathrm{E}(W_t), \quad K(s,t) = \mathrm{Cov}(W_s, W_t), \quad s,t \in T.$$

Scaling

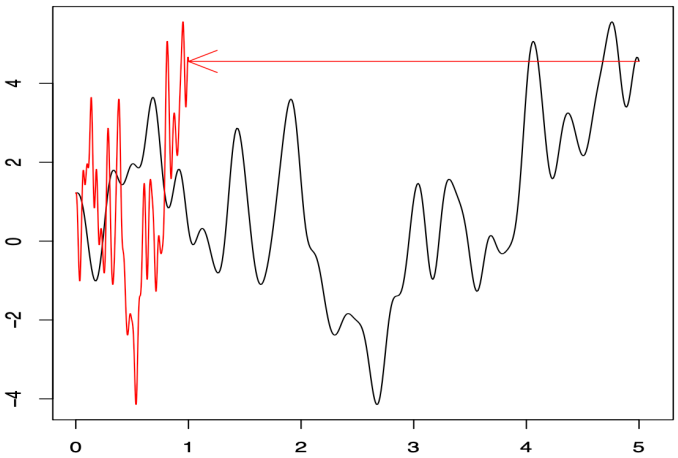
If $W = (W_t : t \in \mathbb{R}^d)$ is a Gaussian process with covariance kernel K, then the process $(W_{\sigma t} : t \in \mathbb{R}^d)$ is another Gaussian process, with covariance kernel $K(\sigma s, \sigma t)$, for any $\sigma > 0$. A scaling factor $\sigma > 1$ shrinks the sample paths, whereas a factor $\sigma < 1$ stretches them.

From Ghosal and Van der Vaart (2017)

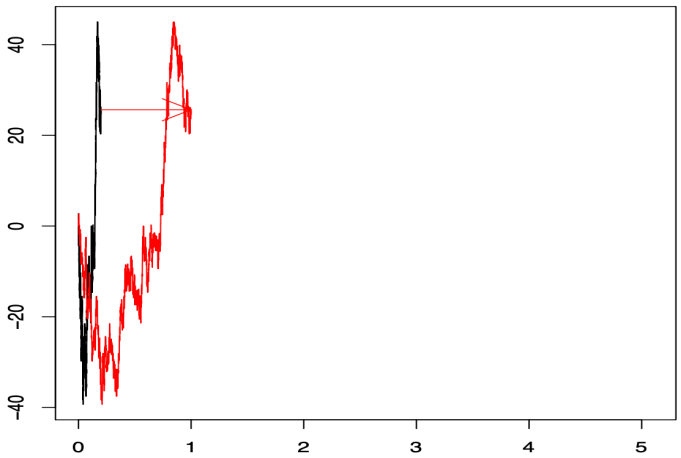
Scaling

If $W = (W_t : t \in \mathbb{R}^d)$ is a Gaussian process with covariance kernel K, then the process $(W_{\sigma t} : t \in \mathbb{R}^d)$ is another Gaussian process, with covariance kernel $K(\sigma s, \sigma t)$, for any $\sigma > 0$. A scaling factor $\sigma > 1$ shrinks the sample paths, whereas a factor $\sigma < 1$ stretches them.

$\sigma > 1$



$\sigma < 1$



From Ghosal and Van der Vaart (2017)

Outline

Examples

Random series

If $Z_1, \ldots, Z_m \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $a_1, \ldots, a_m$ are [deterministic] functions, then the *Random series* $W_t = \sum_{i=1}^{m} a_i(t) Z_i$ defines a Gaussian process with:

$\mu(t) =$

$K(s,t) =$

Brownian motion (or Wiener process)

The *Brownian motion* is the zero-mean Gaussian process, say on $[0, \infty)$, with continuous sample paths and covariance function $K(s, t) = \min(s, t)$.

Brownian motion properties

Let B_t be a Brownian motion, then $\forall s < t$:

- Stationarity: $B_t - B_s \sim \mathcal{N}(0, t - s)$
- Independent increments: $B_t - B_s \perp\!\!\!\perp (B_u, u \leq s)$

Thus it is a Lévy process.

- Self-similar of index $1/2$.

Brownian motion (or Wiener process)

The *Brownian motion* is the zero-mean Gaussian process, say on $[0, \infty)$, with continuous sample paths and covariance function $K(s, t) = \min(s, t)$.

Brownian motion properties

Let B_t be a Brownian motion, then $\forall s < t$:

- Stationarity: $B_t - B_s \sim \mathcal{N}(0, t - s)$
- Independent increments: $B_t - B_s \perp\!\!\!\perp (B_u, u \leq s)$

Thus it is a Lévy process.

- Self-similar of index $1/2$.

Ornstein–Uhlenbeck

The standard *Ornstein–Uhlenbeck process* with parameter $\theta > 0$ is a mean-zero, stationary GP with time set $T = [0, \infty)$, continuous sample paths, and covariance function

$$K(s,t) = (2\theta)^{-1} \exp\left(-\theta|t-s|\right).$$

Properties of Ornstein–Uhlenbeck process

The standard Ornstein–Uhlenbeck process with parameter $\theta > 0$ can be constructed from a Brownian motion B through the relation

$$W_t = (2\theta)^{-1/2} \exp\left(-\theta t\right) B_{e^{2\theta t}}.$$

Relationship between [fixed learning rate] stochastic gradient descent (SGD) and Markov chain Monte Carlo (MCMC) through the Ornstein–Uhlenbeck process: see Mandt, Hoffman, and David M. Blei (2017).

Ornstein–Uhlenbeck

The standard *Ornstein–Uhlenbeck process* with parameter $\theta > 0$ is a mean-zero, stationary GP with time set $T = [0, \infty)$, continuous sample paths, and covariance function

$$K(s,t) = (2\theta)^{-1} \exp\left(-\theta|t-s|\right).$$

Properties of Ornstein–Uhlenbeck process

The standard Ornstein–Uhlenbeck process with parameter $\theta > 0$ can be constructed from a Brownian motion B through the relation

$$W_t = (2\theta)^{-1/2} \exp\left(-\theta t\right) B_{e^{2\theta t}}.$$

Relationship between [fixed learning rate] stochastic gradient descent (SGD) and Markov chain Monte Carlo (MCMC) through the Ornstein–Uhlenbeck process: see Mandt, Hoffman, and David M. Blei (2017).

Ornstein–Uhlenbeck

The standard *Ornstein–Uhlenbeck process* with parameter $\theta > 0$ is a mean-zero, stationary GP with time set $T = [0, \infty)$, continuous sample paths, and covariance function

$$K(s,t) = (2\theta)^{-1} \exp\left(-\theta|t-s|\right).$$

Properties of Ornstein–Uhlenbeck process

The standard Ornstein–Uhlenbeck process with parameter $\theta > 0$ can be constructed from a Brownian motion B through the relation

$$W_t = (2\theta)^{-1/2} \exp\left(-\theta t\right) B_{e^{2\theta t}}.$$

Relationship between [fixed learning rate] stochastic gradient descent (SGD) and Markov chain Monte Carlo (MCMC) through the Ornstein–Uhlenbeck process: see Mandt, Hoffman, and David M. Blei (2017).

Examples

Square exponential

GP with covariance function (a.k.a. radial basis function kernel)

$$K(s,t) = \exp\left(-\frac{\|t-s\|^2}{2\ell^2}\right).$$

Parameter ℓ is called the *characteristic length-scale*.

Fractional Brownian motion

The *fractional Brownian motion* (fBm) with *Hurst parameter* $\alpha \in (0,1)$ is the mean zero Gaussian process $W = (W_t : t \in [0,1])$ with continuous sample paths and covariance function

$$K(s,t) = \frac{1}{2}\left(s^{2\alpha} + t^{2\alpha} - |t-s|^{2\alpha}\right).$$

- $\alpha = 1/2$ yields the standard Brownian motion.

Practical

- Try practical on Gaussian process sampling, gaussian-process-sampling.ipynb.

Outline

Conditional distributions of a multivariate Gaussian

Let an N-dimensional Gaussian vector $\mathbf{x}$ be partitioned as:
$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$ with sizes q and $N - q$, and accordingly $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are partitioned as
$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$ with sizes q and $N - q$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times q & q \times (N-q) \\ (N-q) \times q & (N-q) \times (N-q) \end{bmatrix}.$$

Then

- the conditional distribution of $\mathbf{x}_1$, conditioned on $\mathbf{x}_2 = a$, is multivariate normal $\mathcal{N}_q(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ where $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{a} - \boldsymbol{\mu}_2)$ and $\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$;
- the marginal (unconditional) distribution of $\mathbf{x}_1$ is multivariate normal $\mathcal{N}_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

Gaussian process regression without noise

Let a regression function modeled by a Gaussian process as follows:

$$\mathbf{f} \mid \mathbf{X} \sim GP(\mathbf{f} \mid \boldsymbol{\mu}, \mathbf{K}),$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ represents the observed data points, $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]$ the function values, $\boldsymbol{\mu} = [m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n)]$ the mean function, and $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ the kernel function. Assume $\boldsymbol{\mu} = 0$. We want to predict $\mathbf{f}(\mathbf{X}_*)$ at new points $\mathbf{X}_*$. The joint distribution of $\mathbf{f}$ and $\mathbf{f}_*$ is expressed as:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^\mathsf{T} & \mathbf{K}_{**} \end{bmatrix}\right),$$

where $\mathbf{K} = K(\mathbf{X}, \mathbf{X})$, $\mathbf{K}_* = K(\mathbf{X}, \mathbf{X}_*)$ and $\mathbf{K}_{**} = K(\mathbf{X}_*, \mathbf{X}_*)$. The conditional distribution of interest is:

$$\mathbf{f}_* \mid \mathbf{f}, \mathbf{X}, \mathbf{X}_* \sim \mathcal{N}\left(\mathbf{K}_*^\mathsf{T} \mathbf{K}^{-1} \mathbf{f},\ \mathbf{K}_{**} - \mathbf{K}_*^\mathsf{T} \mathbf{K}^{-1} \mathbf{K}_*\right).$$

Gaussian process regression with noise

Oftentimes, we only have access to noisy versions of the true function, $y = f(x) + \epsilon$, where ϵ represents additive i.i.d. Gaussian noise with variance σ^2. Then $\text{cov}(y) = \mathbf{K} + \sigma^2\mathbf{I}$. The joint distribution of the observed values and the function values at new testing points is:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma^2\mathbf{I} & \mathbf{K}_* \\ \mathbf{K}_*^\mathsf{T} & \mathbf{K}_{**} \end{bmatrix}\right) .$$

Using the conditional distribution, we now get:

$$\mathbf{f}_* \mid \mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}\left(\mathbf{K}_*^\mathsf{T}[\mathbf{K} + \sigma^2\mathbf{I}]^{-1}\mathbf{y}, \mathbf{K}_{**} - \mathbf{K}_*^\mathsf{T}[\mathbf{K} + \sigma^2\mathbf{I}]^{-1}\mathbf{K}_*\right) .$$

Practical

- Try practical on Gaussian process regression, gaussian-process-regression.ipynb.

Outline

Reproducing kernel Hilbert space

To every Gaussian process corresponds a Hilbert space, determined by its covariance kernel. This space determines the support and shape of the process, and therefore is crucial for the properties of the Gaussian process as a prior.

Let's break down what each part means:

Reproducing Kernel: This refers to a mathematical function that takes in two inputs and calculates a measure of similarity or distance between them. It's called reproducing because it has a special property related to the inner product (a mathematical operation) that allows it to reproduce functions.

Hilbert Space: This is a mathematical concept from functional analysis, which is basically a fancy way of saying a space where functions live. A Hilbert space is a mathematical structure that generalizes the notion of Euclidean space to infinite-dimensional spaces (formal def: inner product space that is complete wrt the distance function induced by the inner product).

Reproducing kernel Hilbert space

To every Gaussian process corresponds a Hilbert space, determined by its covariance kernel. This space determines the support and shape of the process, and therefore is crucial for the properties of the Gaussian process as a prior.

Let's break down what each part means:

Reproducing Kernel: This refers to a mathematical function that takes in two inputs and calculates a measure of similarity or distance between them. It's called reproducing because it has a special property related to the inner product (a mathematical operation) that allows it to reproduce functions.

Hilbert Space: This is a mathematical concept from functional analysis, which is basically a fancy way of saying a space where functions live. A Hilbert space is a mathematical structure that generalizes the notion of Euclidean space to infinite-dimensional spaces (formal def: inner product space that is complete wrt the distance function induced by the inner product).

Reproducing kernel Hilbert space

For a Gaussian process $W = (W_t : t \in T)$, let $\overline{\text{lin}}(W)$ be the closure of the set of all linear combinations $\sum_I \alpha_i W_{t_i}$ in the L_2-space of square-integrable variables. The space $\overline{\text{lin}}(W)$ is a Hilbert space.

Definition

The *reproducing kernel Hilbert space* (RKHS) of the mean-zero, Gaussian process $W = (W_t : t \in T)$ is the set $\mathbb{H}$ of all functions $z_H : T \to \mathbb{R}$ defined by $z_H(t) = \mathbb{E}(W_t H)$, for H ranging over $\overline{\text{lin}}(W)$. The corresponding inner product is

$$\langle z_{H_1}, z_{H_2} \rangle_{\mathbb{H}} = \mathbb{E}(H_1 H_2).$$

Reproducing kernel Hilbert space

For a Gaussian process $W = (W_t : t \in T)$, let $\overline{\text{lin}}(W)$ be the closure of the set of all linear combinations $\sum_I \alpha_i W_{t_i}$ in the L_2-space of square-integrable variables. The space $\overline{\text{lin}}(W)$ is a Hilbert space.

Definition

The *reproducing kernel Hilbert space* (RKHS) of the mean-zero, Gaussian process $W = (W_t : t \in T)$ is the set $\mathbb{H}$ of all functions $z_H : T \to \mathbb{R}$ defined by $z_H(t) = \mathrm{E}(W_t H)$, for H ranging over $\overline{\text{lin}}(W)$. The corresponding inner product is

$$\langle z_{H_1}, z_{H_2} \rangle_{\mathbb{H}} = \mathrm{E}(H_1 H_2).$$

Properties of RKHS

- Correspondance $z_H \leftrightarrow H$ is an isometry (by def of inner product), so the definition is well-posed (the correspondence is one-to-one), and H is indeed a Hilbert space.
- Function corresponding to $H = \sum_I \alpha_i W_{s_i}$ is $z_H =$
- For any $s \in T$, function $K(s, \cdot)$ is in RKHS $\mathbb{H}$ associated with $H = W_s$.

Reproducing kernel Hilbert space II

Reproducing formula

For a general function $z_H \in \mathbb{H}$ we have

$$\langle z_H, K(s,\cdot)\rangle_{\mathbb{H}} = \mathrm{E}(HW_s) = z_H(s).$$

That is to say, for any function $h \in \mathbb{H}$,

$$h(t) = \langle h, K(t,\cdot)\rangle_{\mathbb{H}}.$$

Example of RKHS: Euclidean space

Let's consider a simple Euclidean space, such as $\mathbb{R}$. In this case, the RKHS corresponds to a space of functions defined on this real line.

Consider a set of data points on the real line, represented by pairs (x_i, y_i) where x_i are the input values and y_i are the corresponding output values.

Let's model this data with a function $f(x)$. In an RKHS framework, we can express this function as a linear combination of basis functions $k(x, x_i)$, where k is a kernel function that measures the similarity between input points x and x_i.

The RKHS then consists of all functions that can be expressed as:

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$$

where α_i are coefficients that determine the weighting of each basis function.

Common choices for the kernel function in this context might be the linear kernel $k(x, x_i) = x \cdot x_i$, the polynomial kernel $k(x, x_i) = (x \cdot x_i + c)^d$, or the Gaussian kernel $k(x, x_i) = \exp(-\|x - x_i\|^2/2\sigma^2)$.

The RKHS provides a flexible framework for modeling functions in this space using different kernel functions, allowing us to capture various types of relationships between input and output data points.

Outline

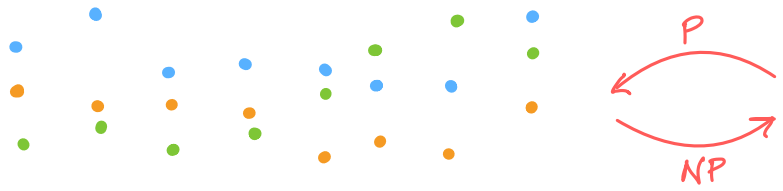
Outline

Gaussian process, Dirichlet process, and their parametric counterparts

Multivariate Gaussian

random vector

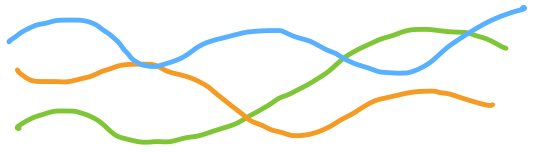
$$\begin{cases} W = (W_1, \dots, W_k) \in \mathbb{R}^k \\ W \sim N(\mu, \Sigma) \end{cases}$$



Gaussian Process

random function

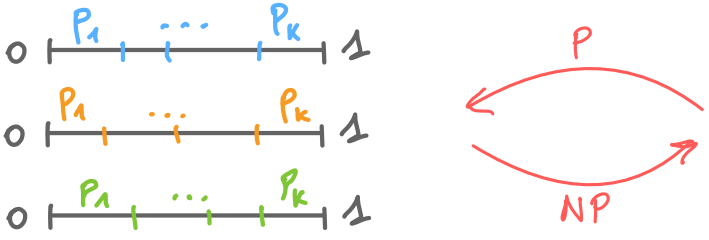
$$\begin{cases} W = (W_t)_{t \in \mathbb{R}}, \ W \sim GP \\ \text{Margins}(W) \sim N(\mu, \Sigma) \end{cases}$$



Dirichlet Distribution

random proba. vector

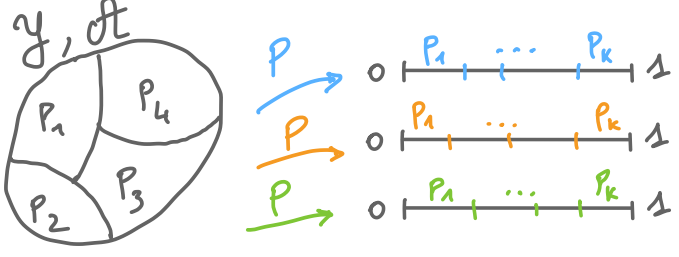
$$\begin{cases} P = (P_1, \dots, P_k) \in S^k \text{ simplex} \\ P \sim \text{Dir}(\alpha), \ \alpha = (\alpha_1, \dots, \alpha_k) \end{cases}$$



Dirichlet Process

random proba. measure

$$\begin{cases} P = (P_A)_{A \in \text{partitions}(Y)}, \ P \sim DP \\ \text{Margins}(P_A) = (P_1, \dots, P_k) \sim \text{Dir} \end{cases}$$



References

- One of the first textbooks: J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. New York: Springer, 2003
- One that reads very well: Nils Lid Hjort et al. *Bayesian nonparametrics*. Vol. 28. Cambridge University Press, Apr. 2010. URL: http://www.cambridge.org/us/academic/subjects/statistics-probability/statistical-theory-and-methods/bayesian-nonparametrics
- Quite a comprehensive one on the theory side: Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press, 2017
- Chapter 31 on Nonparametric Bayesian models of Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL: http://probml.github.io/book2 (as of today, the full version of this chapter can be found in the supplementary of the book)

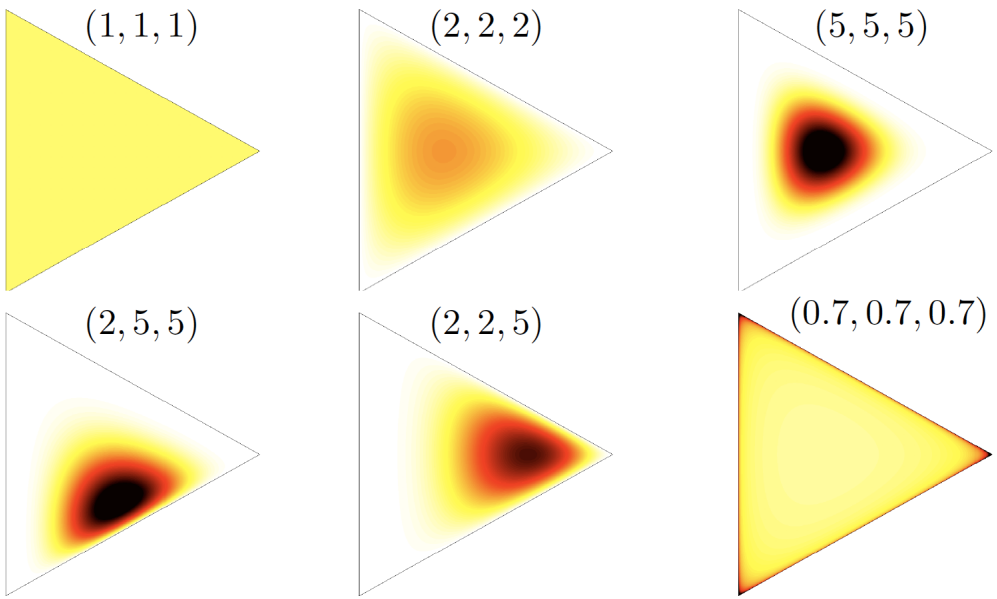
Outline

Dirichlet distribution

The Dirichlet distribution on the simplex Δ_K is a probability distribution with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ with $\alpha_j > 0$ and density function, for $\boldsymbol{x} = (x_1, \ldots, x_K) \in \Delta_K$,

$$f(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}.$$

The Dirichlet distribution is conjugate for the multinomial distribution.



[Image by Y.W. Teh]

Reminder on useful distributions

Let $X \sim \text{Be}(a, b)$, then $E[X] = \dfrac{a}{a+b}$ and $\text{Var}[X] = \dfrac{ab}{(a+b)^2(a+b+1)}$.

Let $X = (X_1, \ldots, X_K) \sim \text{Dir}(\boldsymbol{\alpha})$ and $|\alpha| = \sum_{i=1}^{K} \alpha_i$.

Then for any i, $X \sim \text{Be}(\alpha_i, |\alpha| - \alpha_i)$, and $\text{E}[X_i] = \dfrac{\alpha_i}{|\alpha|}$, and

$\text{Var}[X_i] = \dfrac{\alpha_i(|\alpha| - \alpha_i)}{|\alpha|^2(|\alpha| + 1)}$.

If $i \neq j$, then $\text{Cov}[X_i, X_j] = \dfrac{-\alpha_i \alpha_j}{|\alpha|^2(|\alpha| + 1)}$.

Dirichlet process

A central Bayesian nonparametric prior (Ferguson, 1973).

Definition (Dirichlet process)

A Dirichlet process on the space $\mathcal{Y}$ is a random process P such that there exist $\alpha > 0$ (precision parameter) and P_0 (base/centering distribution) such that for any finite partition $\{A_1, \ldots, A_k\}$ of $\mathcal{Y}$, the random vector $(P(A_1), \ldots, P(A_k))$ is Dirichlet distributed

$$(P(A_1), \ldots, P(A_k)) \sim \text{Dir}(\alpha P_0(A_1), \ldots, \alpha P_0(A_k))$$

Notation: $P \sim \text{DP}(\alpha, P_0)$

Dirichlet process

A central Bayesian nonparametric prior (Ferguson, 1973).

Definition (Dirichlet process)

A Dirichlet process on the space $\mathcal{Y}$ is a random process P such that there exist $\alpha > 0$ (precision parameter) and P_0 (base/centering distribution) such that for any finite partition $\{A_1, \ldots, A_k\}$ of $\mathcal{Y}$, the random vector $(P(A_1), \ldots, P(A_k))$ is Dirichlet distributed

$$(P(A_1), \ldots, P(A_k)) \sim \text{Dir}(\alpha P_0(A_1), \ldots, \alpha P_0(A_k))$$

Notation: $P \sim \text{DP}(\alpha, P_0)$

Dirichlet process

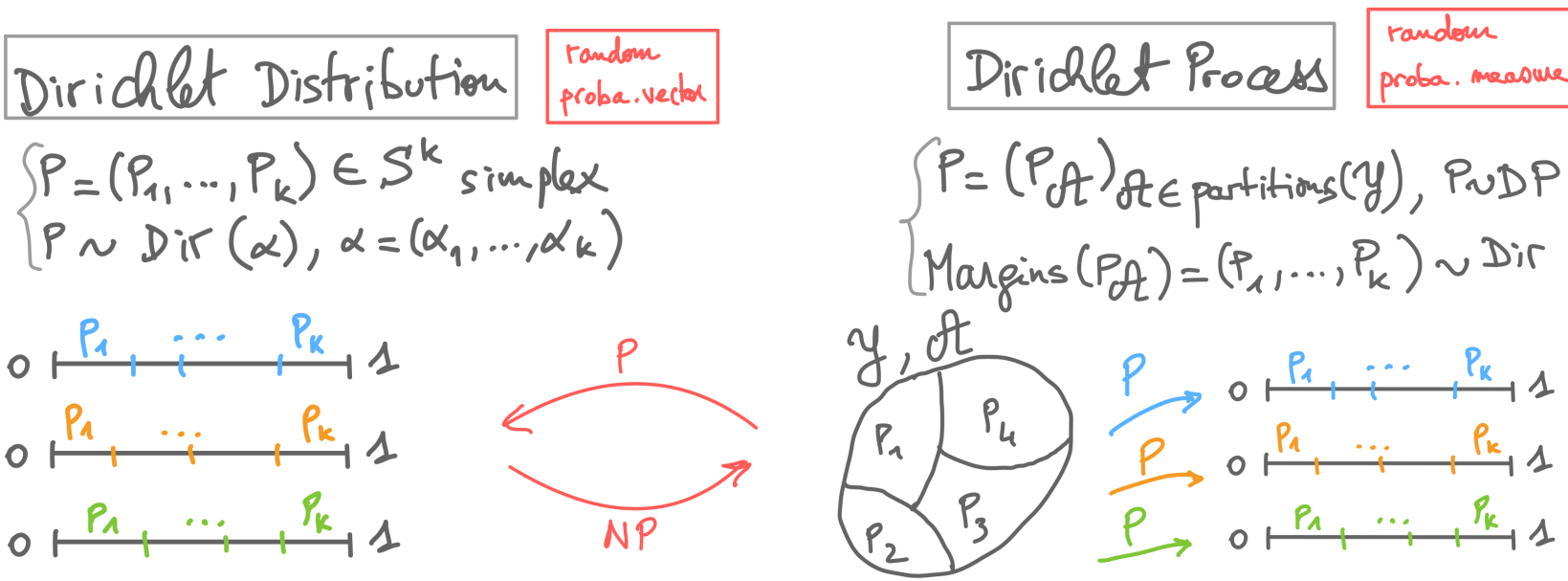
A central Bayesian nonparametric prior (Ferguson, 1973).

Definition (Dirichlet process)

A Dirichlet process on the space $\mathcal{Y}$ is a random process P such that there exist $\alpha > 0$ (precision parameter) and P_0 (base/centering distribution) such that for any finite partition $\{A_1, \ldots, A_k\}$ of $\mathcal{Y}$, the random vector $(P(A_1), \ldots, P(A_k))$ is Dirichlet distributed

$$(P(A_1), \ldots, P(A_k)) \sim \text{Dir}(\alpha P_0(A_1), \ldots, \alpha P_0(A_k))$$

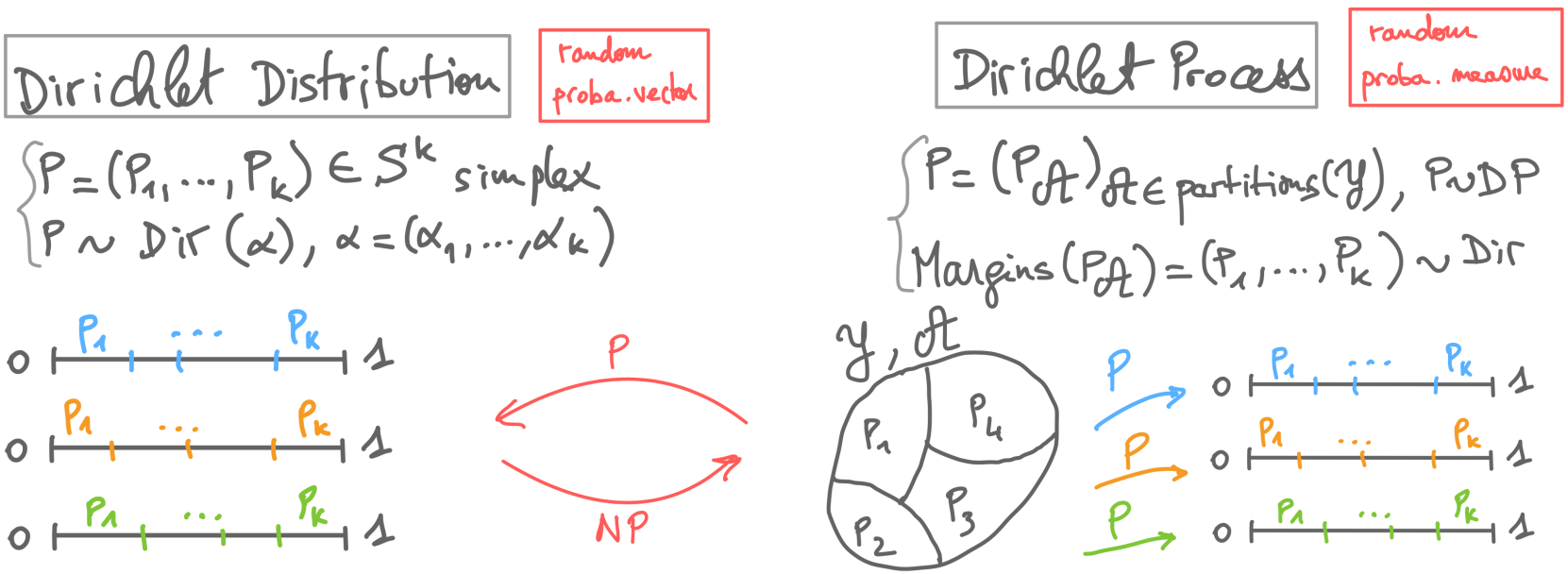
Notation: $P \sim \text{DP}(\alpha, P_0)$



Notations

Two notations are valid ones: $P \sim \mathrm{DP}(\alpha, P_0)$, as well as $P \sim \mathrm{DP}(\alpha P_0)$. The only difference is that the second one take a single parameter in argument, in the form of a measure (not necessarily a probability measure).

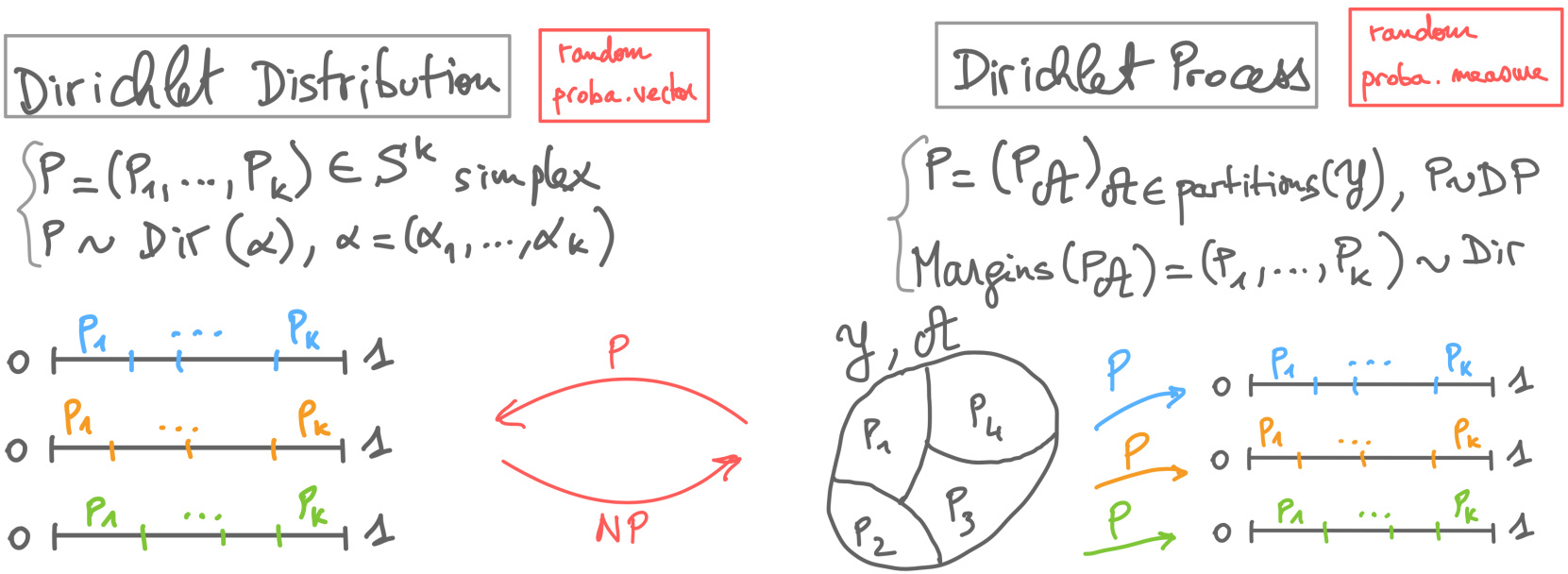
Can you move from one to the other?For instance, what are the (α, P_0) parameters associated to the notation $P \sim \mathrm{DP}(G_0)$, with G_0 a measure.



Notations

Two notations are valid ones: $P \sim \mathrm{DP}(\alpha, P_0)$, as well as $P \sim \mathrm{DP}(\alpha P_0)$. The only difference is that the second one take a single parameter in argument, in the form of a measure (not necessarily a probability measure).

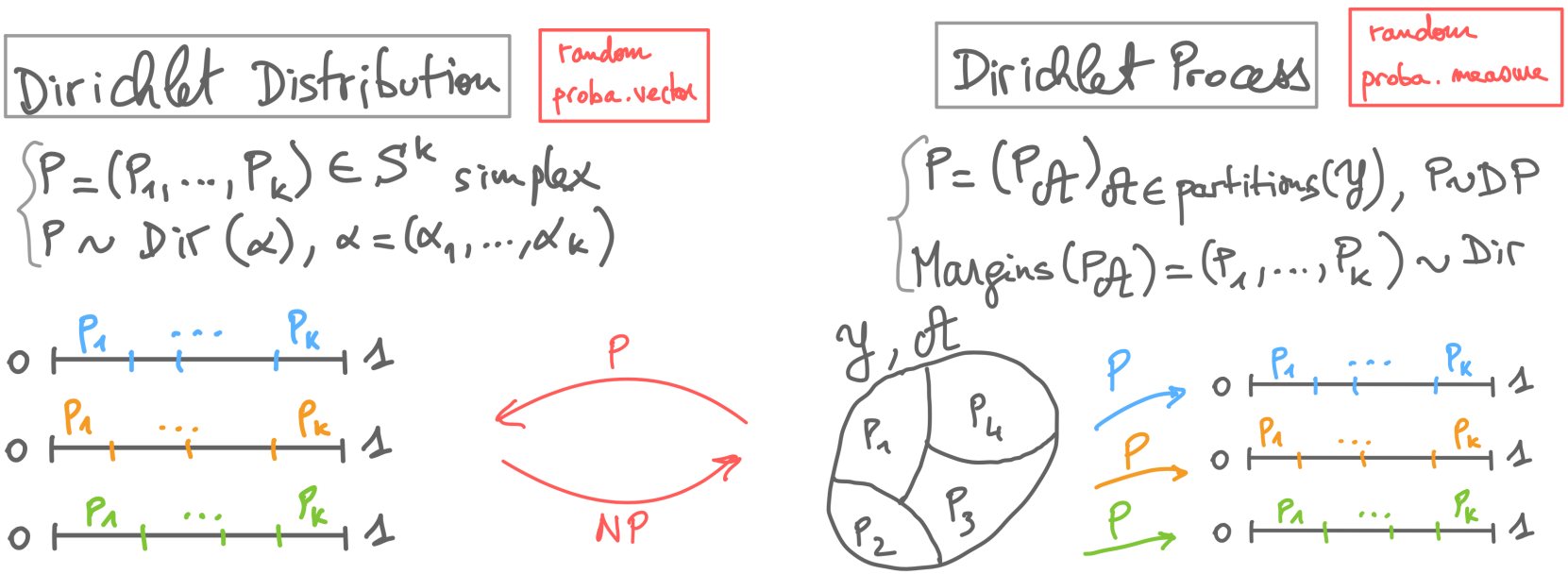
Can you move from one to the other? For instance, what are the (α, P_0) parameters associated to the notation $P \sim \mathrm{DP}(G_0)$, with G_0 a measure.



Notations

Two notations are valid ones: $P \sim \mathrm{DP}(\alpha, P_0)$, as well as $P \sim \mathrm{DP}(\alpha P_0)$. The only difference is that the second one take a single parameter in argument, in the form of a measure (not necessarily a probability measure).

Can you move from one to the other?For instance, what are the $((\alpha, P_0)$ parameters associated to the notation $P \sim \mathrm{DP}(G_0)$, with G_0 a measure.



Moments of Dirichlet process I

Proposition

Let $P \sim DP(\alpha, P_0)$ then for every measurable sets A, B we have

$$\mathrm{E}[P(A)] = P_0(A),$$

$$\mathrm{Var}[P(A)] = \frac{P_0(A)(1 - P_0(A))}{1 + \alpha},$$

$$\mathrm{Cov}(P(A), P(B)) = \frac{P_0(A \cap B) - P_0(A)P_0(B)}{1 + \alpha}.$$

Moments of Dirichlet process II

Proof. We make use of $P(A) \sim \text{Beta}(\alpha P_0(A), \alpha(1 - P_0(A)))$. From this we obtain

$$\mathrm{E}(P(A)) = \frac{\alpha P_0(A)}{\alpha(P_0(A) + 1 - P_0(A))} = P_0(A)$$

and

$$\text{Var}(P(A)) = \frac{\alpha^2 P_0(A)(1 - P_0(A))}{\alpha^2(\alpha + 1)}.$$

We derive the covariance result by developing the terms of $\text{Cov}(P(A), P(B)) = \text{Cov}(P(A \cap B) + P(A \cap B^c), P(B \cap A) + P(B \cap A^c))$. $\square$

Marginalizing out the DP

A Dirichlet process model can be constructed as a two level sampling model:

$$\begin{cases} P \sim \mathrm{DP}(\alpha, P_0) \\ X|P \sim P, \end{cases}$$

i.e. a probability measure P is first sampled from the Dirichlet process and then given P, we sample a random variable X.

This defines a joint distribution on (P, X).

Marginalizing out P, we obtain the marginal distribution of X. Property $\mathrm{E}[P(A)] = P_0(A)$ can be written equivalently as

$$\mathrm{E}(P(A)) = P_0(A) = \int P(A)\mathrm{dDP}(P),$$

which yields that marginally,

$$X \sim P_0.$$

Marginalizing out the DP

A Dirichlet process model can be constructed as a two level sampling model:

$$\begin{cases} P \sim \mathrm{DP}(\alpha, P_0) \\ X|P \sim P, \end{cases}$$

i.e. a probability measure P is first sampled from the Dirichlet process and then given P, we sample a random variable X.

This defines a joint distribution on (P, X).

Marginalizing out P, we obtain the marginal distribution of X. Property $\mathrm{E}[P(A)] = P_0(A)$ can be written equivalently as

$$\mathrm{E}(P(A)) = P_0(A) = \int P(A)\mathrm{dDP}(P),$$

which yields that marginally,

$$X \sim P_0.$$

Posterior distribution

Let $X_{1:n} := (X_1, \dots, X_n)$ be sampled from the hierarchical model

$$\begin{cases} P \sim \mathrm{DP}(\alpha, P_0) \\ X_{1:n} | P \overset{\text{iid}}{\sim} P \end{cases}$$

to be used as a building block of a larger hierarchical model (e.g. mixture a model).

Theorem (DP posterior distribution)

The DP is conjugate, with posterior equal to

$$P | X_{1:n} \sim DP(\alpha P_0 + \sum_{i=1}^{n} \delta_{X_i}) = DP(\alpha_n, P_n)$$

with

$$\alpha_n = \alpha + n, \quad P_n = \frac{\alpha}{\alpha + n} P_0 + \frac{n}{\alpha + n} \tilde{P}_n,$$

where $\tilde{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ is the data empirical measure. The predictive distribution, called Pólya urn or Blackwell–MacQueen scheme, is given by

$$X_{n+1} | X_{1:n} \sim P_n.$$

Posterior distribution

Let $X_{1:n} := (X_1, \dots, X_n)$ be sampled from the hierarchical model

$$\begin{cases} P \sim \mathrm{DP}(\alpha, P_0) \\ X_{1:n} | P \overset{\text{iid}}{\sim} P \end{cases}$$

to be used as a building block of a larger hierarchical model (e.g. mixture a model).

Theorem (DP posterior distribution)

The DP is conjugate, with posterior equal to

$$P | X_{1:n} \sim DP(\alpha P_0 + \sum_{i=1}^{n} \delta_{X_i}) = DP(\alpha_n, P_n)$$

with

$$\alpha_n = \alpha + n, \quad P_n = \frac{\alpha}{\alpha + n} P_0 + \frac{n}{\alpha + n} \tilde{P}_n,$$

where $\tilde{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ *is the data empirical measure. The predictive distribution, called Pólya urn or Blackwell–MacQueen scheme, is given by*

$$X_{n+1} | X_{1:n} \sim P_n.$$

Posterior distribution. Proof

Proof. The posterior distribution of $\boldsymbol{a} = (a_1, \ldots, a_k) = (P(A_1), \ldots, P(A_k))$ depends on the observations only via their cell counts $\boldsymbol{N} = (N_1, \ldots, N_k)$, $N_j = \#\{i : X_i \in A_j\}$ (it comes from tail–free property), so

$$\boldsymbol{a}|X_{1:n} \sim \boldsymbol{a}|N_{1:k}.$$

The prior and model are

$$\begin{cases} \boldsymbol{a} \sim \mathrm{Dir}_k(\alpha P_0(A_1), \ldots, \alpha P_0(A_k)) \\ \boldsymbol{N}|P \sim \mathrm{Multinom}_k(\boldsymbol{a}). \end{cases}$$

This results in the posterior of form

$$\begin{aligned} P(\boldsymbol{a}|\boldsymbol{N}) &\propto a_1^{\alpha P_0(A_1)+N_1-1} \cdots a_k^{\alpha P_0(A_k)+N_k-1} \\ &= \mathrm{Dir}_k\big(\alpha P_0(A_1) + N_1, \ldots, \alpha P_0(A_k) + N_k\big). \end{aligned}$$

□

Number of distinct values

Let $D_i = \mathbb{I}(X_i \text{ is a new value})$ and denote by $K_n = \sum_{i=1}^{n} D_i$ the number of distinct values in $X_1, \ldots, X_n$.

Denote its distribution $\mathcal{L}(K_n)$. Assuming that the base measure P_0 is non-atomic, then with probability 1: $X_i \notin \{X_1, \ldots, X_{i-1}\} \Leftrightarrow X_i \sim P_0$.

Proposition (Asymptotics for K_n)

Random variables D_i are Bernoulli$(\alpha/(\alpha + i - 1))$. For $n \to \infty$:

i) $\mathrm{E}K_n \sim \alpha \log n \sim \mathrm{Var}(K_n)$

ii) $K_n / \log(n) \xrightarrow{a.s.} \alpha$

iii) $(K_n - \mathrm{E}K_n)/sd(K_n) \to N(0,1)$

iv) $d_{TV}(\mathcal{L}(K_n), Poisson(\mathrm{E}K_n)) = o(1/\log(n))$ *where*

$$d_{TV}(P, Q) = sup|P(A) - Q(A)|$$

over measurable partitions A.

Number of distinct values

Let $D_i = \mathbb{I}(X_i \text{ is a new value})$ and denote by $K_n = \sum_{i=1}^{n} D_i$ the number of distinct values in $X_1, \ldots, X_n$.

Denote its distribution $\mathcal{L}(K_n)$. Assuming that the base measure P_0 is non-atomic, then with probability 1: $X_i \notin \{X_1, \ldots, X_{i-1}\} \Leftrightarrow X_i \sim P_0$.

Proposition (Asymptotics for K_n)

Random variables D_i are Bernoulli$(\alpha/(\alpha + i - 1))$. For $n \to \infty$:

i) $\mathrm{E}K_n \sim \alpha \log n \sim \mathrm{Var}(K_n)$

ii) $K_n / \log(n) \xrightarrow{a.s.} \alpha$

iii) $(K_n - \mathrm{E}K_n)/sd(K_n) \to N(0,1)$

iv) $d_{TV}\big(\mathcal{L}(K_n), Poisson(\mathrm{E}K_n)\big) = o\big(1/\log(n)\big)$ *where*

$$d_{TV}(P, Q) = sup|P(A) - Q(A)|$$

over measurable partitions A.

Number of distinct values

Proof.

i) $\mathrm{E}K_n = \sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1}$ and $\mathrm{Var}(K_n) = \sum_{i=1}^{n} \frac{\alpha(i-1)}{(\alpha+i-1)^2}$.

ii) Since D_i's are $\mathbb{I}$ one may use Kolmogorov law of strong numbers and

$$\sum_{i=1}^{\infty} \frac{\mathrm{Var}(D_i)}{(\log i)^2} = \sum_{i=1}^{\infty} \frac{\alpha(i-1)}{(\alpha+i-1)^2(\log i)^2} < \infty$$

by e.g. the fact that $\sum_i (1/i(\log i)^2)$ converges.

iii) By Lindeberg central limit theorem.

iv) This is implied from Chein–Stein approximation.

□

A central limit theorem for independent random variables (possibly not identically distributed).

Theorem (Lindeberg central limit theorem)

Suppose X_i are i.i.d. such that $\mathrm{E}X_i = \mu_i$ and $\mathrm{Var}X_i = \sigma_i^2 < \infty$. Define $Y_i = X_i - \mu_i$, $T_n = \sum_{i=1}^n Y_i$, $s_n^2 = \mathrm{Var}(T_n) = \sum_{i=1}^n \sigma_i^2$. Then provided that

$$\forall \epsilon > 0 \quad \frac{1}{s_n^2} \sum_{i=1}^{n} \mathrm{E}\big(Y_i^2 \mathbb{I}(|Y_i| > \epsilon s_n)\big) \xrightarrow{n\to\infty} 0 \quad \textit{[Lindeberg condition]},$$

we have the central limit theorem: $T_n/s_n \xrightarrow{d} N(0,1)$.

Distribution of distinct values

We have now the limits of K_n and we know approximations of its distribution $\mathcal{L}(K_n)$. The exact distribution of K_n is:

Proposition (Distribution of K_n)

If P_0 is non-atomic then

$$\mathbb{P}(K_n = k) = \mathfrak{C}_n(k) n! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \tag{1}$$

where

$$\mathfrak{C}_n(k) = \frac{1}{n!} \sum_{S \in \mathfrak{J}_n(k)} \prod_{j \in S} j \tag{2}$$

and $\mathfrak{J}_n(k) = \{S \subset \{1, \ldots, n-1\},\ |S| = n - k\}$.

Recall the definition of the Gamma function $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$.

Proof.

Let us consider when we may deal with events $K_n = k$: we have two cases

$$\begin{cases} K_{n-1} = k-1 \text{ and } X_n \text{ is a new value} \\ K_{n-1} = k \text{ and } X_n \text{ is not a new value.} \end{cases}$$

This results in

$$p_n(k,\alpha) := \mathbb{P}(k_n = k|\alpha) = \frac{\alpha}{\alpha+n-1} p_{n-1}(k-1,\alpha) + \frac{n-1}{\alpha+n-1} p_{n-1}(k,\alpha). \tag{3}$$

Now let us remark that $\mathfrak{C}_n(k) = p_n(k, \alpha = 1)$. Therefore

$$\mathfrak{C}_n(k) = \frac{1}{n}\mathfrak{C}_{n-1}(k-1) + \frac{n-1}{n}\mathfrak{C}_{n-1}(k). \tag{4}$$

By induction over n: first we check case $n = 1$:

$$p_1(1,\alpha) = \mathfrak{C}_1(1)\frac{\alpha}{\alpha} = \mathfrak{C}_1(1).$$

Distribution of distinct values

To check case $n > 1$ we use (1) and then (3):

$$
\begin{aligned}
p_n(k,\alpha) &= \frac{\alpha}{\alpha+n-1} p_{n-1}(k-1,\alpha) + \frac{n-1}{\alpha+n-1} p_{n-1}(k,\alpha) \\
&= \frac{\alpha}{\alpha+n-1} \mathfrak{C}_{n-1}(k-1)(n-1)!\alpha^{k-1} \frac{\Gamma(\alpha)}{\Gamma(\alpha+n-1)} + \\
&+ \frac{n-1}{\alpha+n-1} \mathfrak{C}_{n-1}(k)(n-1)!\alpha^{k} \frac{\Gamma(\alpha)}{\Gamma(\alpha+n-1)} \\
&= \frac{\alpha^k}{\alpha+n-1}(n-1)! \frac{\Gamma(\alpha)}{\Gamma(\alpha+n-1)} n \left(\frac{1}{n} \mathfrak{C}_{n-1}(k-1) + \frac{n-1}{n} \mathfrak{C}_{n-1}(k) \right) \\
&= \mathfrak{C}_n(k) n! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)},
\end{aligned}
$$

which proves property (1).

To prove (2) let us define a polynomial $A_n(s)$ as $A_n(s) = \sum_{i=1}^{\infty} \mathfrak{C}_n(k)s^k$. Then using (4) polynomial $A_n(s)$ can be written as

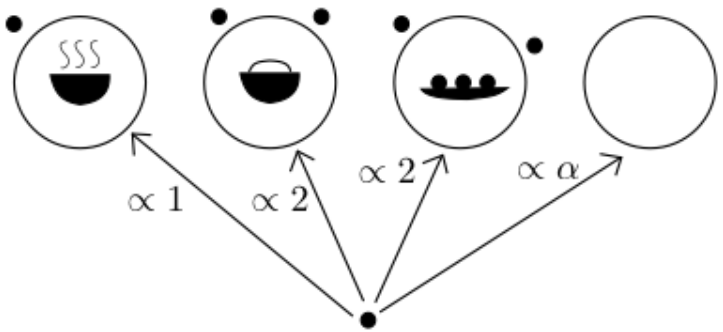
$$\begin{aligned} A_n(s) &= \sum_{k=1}^{\infty} \left(\frac{1}{n} \mathfrak{C}_{n-1}(k-1) + \frac{n-1}{n} \mathfrak{C}_{n-1}(k) \right) s_k \\ &= \frac{1}{n}(sA_{n-1}(s) + (n-1)A_{n-1}(s)) = \frac{s+n-1}{n} A_{n-1}(s) \\ &= \ldots = A_1(s) \prod_{j=2}^{n} \frac{s+j-1}{j} = \frac{s(s+1)\cdot \ldots \cdot (s+n-1)}{n!}. \end{aligned}$$

Last equality implies from the fact that $\mathfrak{C}_1(k) = \mathbf{1}\{k = 1\}$ and hence $A_1(s) = s$. Checking terms after the expansion finishes the proof of (2).

Chinese Restaurant process

This is a culinary metaphor of the random partition induced by the DP.

Generative process: customers join tables with probability proportional to n_j, the number of clients already sitting, or sit at new table with probability proportional to α.



Proposition (Chinese Restaurant process)

A random sample $X_{1:n}$ from a DP with precision parameter α induces a partition of $\{1, \ldots, n\}$ into k sets of sizes $n_1, \ldots, n_k$ with probability

$$p(n_1, \ldots, n_k) = p(\{n_1, \ldots, n_k\}) = \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \prod_{j=1}^{k} \Gamma(n_j).$$

Chinese Restaurant process. Proof

Proof. We use the Pólya urn scheme slightly changed by using $n_1, \dots, n_k$

$$\mathbb{P}(X_{n+1}|X_{1:n}) = \frac{\alpha}{\alpha+n}P_0 + \frac{1}{\alpha+n}\sum_{j=1}^{k} n_j \delta_{X_j^*}.$$

By exchangeability, the distribution of $\{n_1, \dots, n_k\}$ does not depend on the order of the observations. Let's compute $p(n_1, \dots, p_k)$ as the probability of one draw where the first table consists of first n_1 observations etc.

To proceed, let us use Pólya urn scheme: we denote $\bar{n}_j = \sum_{i=1}^{j} n_i$ and hence $\bar{n}_k = n$, the total number of observations. We can observe the following pattern: first ball open new table, following $n_j - 1$ ones fill in that table and so forth. That quantity can be rewritten as

$$\frac{\alpha^k}{\alpha(\alpha+1)\dots(\alpha+n-1)}\prod_{j=1}^{k}(n_j-1)! = \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}\prod_{j=1}^{k}\Gamma(n_j).$$

Note that for ordered partitions we have

$$\bar{p}(n_1, \dots, n_k) = \frac{p(n_1, \dots, n_k)}{k!}.$$

□

Ewens sampling formula

Ewens sampling formula (ESF), presented originally by Ewens (1972), is the distribution of multiplicities $m = (m_1, \ldots, m_n)$, m_ℓ is the number of groups of size ℓ. Also known as allelic partitions in population genetics, when there is no selective difference between types: null hypothesis in non Darwinian theory. See also Antoniak (1974).

Proposition (Ewens sampling formula)

The distribution of the multiplicities $(m_1, \ldots, m_n)$ *induced by a DP is*

$$p(m_1, \ldots, m_n) = \frac{\alpha^k}{\alpha_{(n)}} \frac{n!}{\prod_{\ell=1}^{n} \ell^{m_\ell} m_\ell!}.$$

Notation $n_{(k)} := n(n+1)\cdots(n+k-1)$.

Ewens sampling formula

Proof. Two steps: 1) Compute probability of particular sequence of $X_1, \ldots, X_n$ in given class $(m_1, \ldots, m_n)$, note that all such sequences are equally likely and 2) multiply obtained quantity by the number of such sequences.

1) Consider a sequence $X_1, \ldots, X_n$ such that $X_1, \ldots, X_{m_1}$ occur each only once, then the next m_2 occur only twice and so on. This sequence has probability which may be obtained by the Pólya Urn scheme in the same fashion as CRP:

$$\frac{\alpha^{m_1}(\alpha \cdot 1)^{m_2} \ldots \big(\alpha \cdot 1 \cdot \ldots \cdot (n-1)\big)^{m_n}}{\alpha_{(n)}} = \frac{\alpha^k}{\alpha_{(n)}} \prod_{\ell=1}^{n} ((\ell-1)!)^{m_\ell}.$$

2) Number of sequences $X_1, \ldots, X_n$ with frequencies $(m_1, \ldots, m_n)$ is a number of ways of putting n distinct objects into bins, so called multinomial coefficient. Since ordering of the m_ℓ bins of frequency ℓ is irrelevant, divide by $m_\ell!$:

$$\frac{1}{\prod_{l=1}^{n}(m_\ell)!} \binom{n}{1 \times \#m_1, 2 \times \#m_2, \ldots, n \times \#m_n} = \frac{n!}{\prod_{\ell=1}^{n} m_\ell!(\ell!)^{m_\ell}}$$

To finish one needs to multiply results obtained in 1) and 2). □

Stick-breaking representation

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

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

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

Stick-breaking representation

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

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

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

Stick-breaking representation

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

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

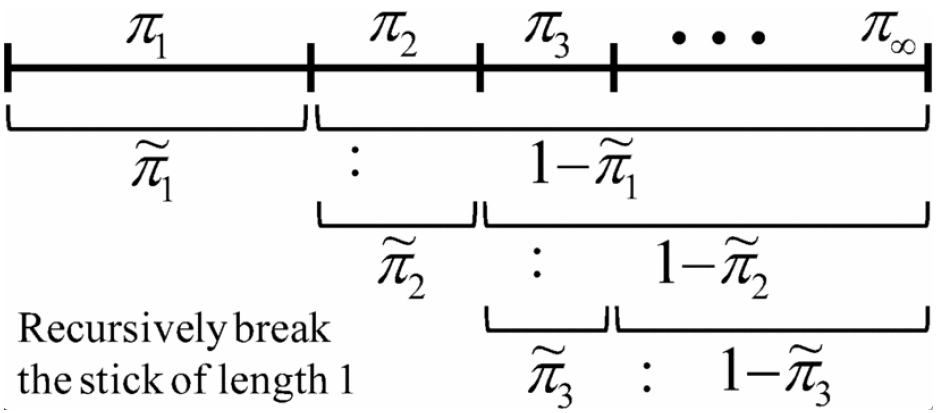
- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $\pi_j = \tilde{\pi}_j \prod_{l<j}(1 - \tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} \mathit{Beta}(1, \alpha)$,

Stick-breaking representation

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

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

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

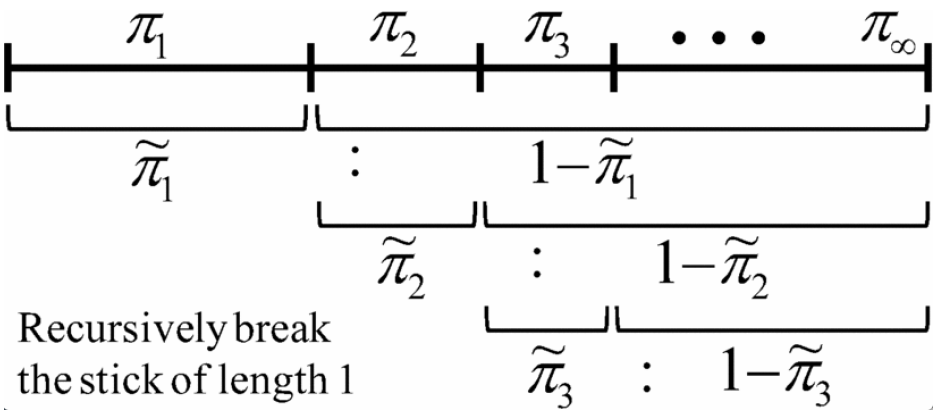


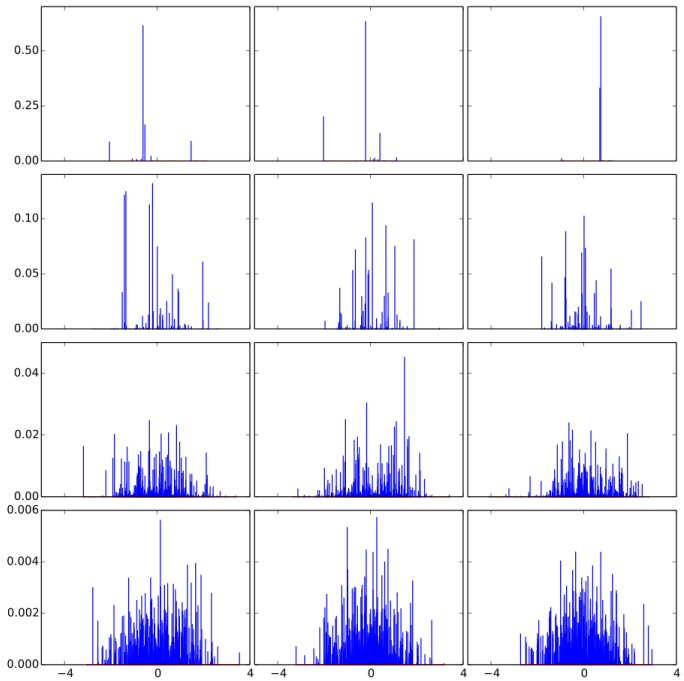
Stick-breaking representation

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

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

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





Stick-breaking representation

A constructive representation of the DP due to Sethuraman (1994).

Theorem (Stick-breaking)

If $V_1, V_2, \ldots \overset{iid}{\sim} Be(1, \alpha)$ *and* $\phi_1, \phi_2, \ldots \overset{iid}{\sim} P_0$ *are independent random variables, and* $p_1 = V_1$ *and*

$$p_j = V_j \prod_{l<j} (1 - V_l),$$

then

$$P = \sum_{i=1}^{\infty} p_i \delta_{\phi_i} \sim DP(\alpha, P_0).$$

The vector $\boldsymbol{p} = (p_1, p_2, \ldots)$ constructed in this way is an infinite random probability vector. Its distribution is called the GEM distribution, named after Griffiths, Engen and McClosky, denoted by $\boldsymbol{p} \sim \text{GEM}(\alpha)$.

Stick-breaking representation. Proof

The main ingredient of the proof is:

Lemma

For independent $\phi \sim P_0$ and $V \sim Be(1, \alpha)$, the DP is the only solution of the distributional equation

$$P \stackrel{d}{=} V\delta_\phi + (1 - V)P.$$

Proof of the Lemma.

Existence. The DP is indeed a solution by combining its definition with the regenerative property:

If $p \sim Dir(\alpha)$, $N \sim Multinoulli(\alpha)$ and $V \sim Beta(1, |\alpha|)$ are independent, then

$$Vn + (1 - V)P \sim Dir(\alpha).$$

Uniqueness. Comes from this lemma:

Lemma

Let X and $V \in [-1, 1]$ be random variables such that $P(|V| < 1) = 1$. Then the distribution of any random variable Y that is independent of (X, V) and satisfies the distributional equation $Y \stackrel{d}{=} X + VY$ is unique.

Stick-breaking representation. Proof

The main ingredient of the proof is:

Lemma

For independent $\phi \sim P_0$ and $V \sim Be(1, \alpha)$, the DP is the only solution of the distributional equation

$$P \stackrel{d}{=} V\delta_\phi + (1 - V)P.$$

Proof of the Lemma.

Existence. The DP is indeed a solution by combining its definition with the regenerative property:

If $p \sim Dir(\boldsymbol{\alpha})$, $N \sim Multinoulli(\boldsymbol{\alpha})$ and $V \sim Beta(1, |\boldsymbol{\alpha}|)$ are independent, then

$$Vn + (1 - V)P \sim Dir(\boldsymbol{\alpha}).$$

Uniqueness. Comes from this lemma:

Lemma

Let X and $V \in [-1, 1]$ be random variables such that $P(|V| < 1) = 1$. Then the distribution of any random variable Y that is independent of (X, V) and satisfies the distributional equation $Y \stackrel{d}{=} X + VY$ is unique.

Stick-breaking representation. Proof

Proof of the SB Theorem. 1) The weights $(p_1, p_2, \ldots)$ need to form a probability vector. The leftover mass at stage j is

$$1 - \left(\sum_{i=1}^{j} p_i \right) = \prod_{i=1}^{j} (1 - V_i) =: R_j.$$

Notice that R_j is decreasing and for every j we have $R_j \in [0, 1]$. Hence almost sure convergence is equivalent with convergence in mean. We can check convergence in mean easily:

$$\mathrm{E} R_j = \mathrm{E} \prod_{j} (1 - V_j) = \prod_{j} \mathrm{E}(1 - V_j) = \left(\frac{\alpha}{\alpha + 1} \right)^j \to 0.$$

So $(p_1, p_2, \ldots)$ is a probability vector almost surely and P is a probability measure almost surely.

2) Now write

$$P = p_1\delta_{\phi_1} + \sum_{j=2}^{\infty} p_j\delta_{\phi_j} = V_1\delta_{\phi_1} + (1 - V_1)\sum_{j=1}^{\infty} \tilde{p}_j\delta_{\tilde{\phi}_j},$$

where

$$\tilde{p}_j = \frac{p_{j+1}}{1 - V_1} = V_{j+1}\prod_{l=2}^{j}(1 - V_l), \quad \tilde{\phi}_j = \phi_{j+1}.$$

Then $(\tilde{p}_j)$ and $(\tilde{\phi}_j)$ satisfy the same distributional definitions as (p_j) and (ϕ_j), hence $\tilde{P} \stackrel{d}{=} P$ in distribution.

So P is a solution of the Lemma equation (4) whose only solution is the DP. $\square$

DP as a normalized Gamma process I

The DP can be obtained by normalizing a Gamma process. It is a generic way to obtain random probability measures from almost surely finite random measures. Let us restrict to $\mathcal{Y} = \mathbb{R}$.

Definition

Gamma process on $\mathbb{R}_+$ is a process $(S(u) : u \geq 0)$ with independent increments satisfying

$$\forall u_1 : 0 \leq u_1 \leq u_2 : \quad S(u_2) - S(u_1) \overset{\text{ind}}{\sim} Ga(u_2 - u_1, 1).$$

This ensures that the process has non-decreasing right continuous sample path $u \mapsto S(u)$.

Theorem

For every $\alpha > 0$ and for every cumulative distribution function G, a random cumulative distribution function such that

$$F(t) = \frac{S(\alpha G(t))}{S(\alpha)}$$

is the distribution of a $\text{DP}(\alpha, G)$.

DP as a normalized Gamma process II

Proof. For any set of t_i satisfying $-\infty = t_0 < t_1 < \ldots < t_k = \infty$ we have

$$S(\alpha G(t_i)) - S(\alpha G(t_{i-1})) \sim Ga\big(\alpha G(t_i) - \alpha G(t_{i-1}), 1\big).$$

Use property that if $Y_i \overset{\text{ind}}{\sim} Ga(\alpha_i, 1)$ then
$(Y_1, \ldots, Y_n)/\sum_i Y_i \sim \text{Dir}_n(\alpha_1, \ldots, \alpha_n)$ to obtain

$$\big(F(t_1)-F(t_0), \ldots, F(t_k)-F(t_{k-1})\big) \sim \text{Dir}_k\big(\alpha G(t_1)-\alpha G(t_0), \ldots, \alpha G(t_k)-\alpha G(t_{k-1})\big).$$

Hence the definition of DP holds for every partition in intervals. These form a measure determining class, so that the definition holds for every partition in general. □

Definition via the Pólya urn scheme

A Pólya sequence with parameter αP_0 is a sequence of random variables $X_1, \ldots, X_n$ whose joint distribution satisfies

$$X_1 \sim P_0, \;\; X_{n+1}|X_1, \ldots, X_n \sim \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i}.$$

Theorem

If $X_1, X_2, \ldots$ is a Pólya sequence then there exists a random probability measure P such that $X_i|P \overset{\text{iid}}{\sim} P$ and $P \sim DP(\alpha, P_0)$.

Proof. Considering a Pólya sequence as the outcome of a Pólya urn, we see that it is exchangeable. By de Finetti's theorem there exists such a probability measure P such that $X_i|P \overset{\text{iid}}{\sim} P$. So far we have proved existence of the DP and know that the DP generates a Pólya sequence. Since the random probability measure given by de Finetti's theorem is unique this proves that $P \sim \text{DP}(\alpha, P_0)$. □

Definition via the Pólya urn scheme

A Pólya sequence with parameter αP_0 is a sequence of random variables $X_1, \ldots, X_n$ whose joint distribution satisfies

$$X_1 \sim P_0, \;\; X_{n+1}|X_1, \ldots, X_n \sim \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i}.$$

Theorem

If $X_1, X_2, \ldots$ is a Pólya sequence then there exists a random probability measure P such that $X_i|P \overset{iid}{\sim} P$ and $P \sim DP(\alpha, P_0)$.

Proof. Considering a Pólya sequence as the outcome of a Pólya urn, we see that it is exchangeable. By de Finetti's theorem there exists such a probability measure P such that $X_i|P \overset{\text{iid}}{\sim} P$. So far we have proved existence of the DP and know that the DP generates a Pólya sequence. Since the random probability measure given by de Finetti's theorem is unique this proves that $P \sim \text{DP}(\alpha, P_0)$. □

Definition via the Pólya urn scheme

A Pólya sequence with parameter αP_0 is a sequence of random variables $X_1, \ldots, X_n$ whose joint distribution satisfies

$$X_1 \sim P_0, \;\; X_{n+1}|X_1, \ldots, X_n \sim \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i}.$$

Theorem

If $X_1, X_2, \ldots$ is a Pólya sequence then there exists a random probability measure P such that $X_i|P \overset{\text{iid}}{\sim} P$ and $P \sim DP(\alpha, P_0)$.

Proof. Considering a Pólya sequence as the outcome of a Pólya urn, we see that it is exchangeable. By de Finetti's theorem there exists such a probability measure P such that $X_i|P \overset{\text{iid}}{\sim} P$. So far we have proved existence of the DP and know that the DP generates a Pólya sequence. Since the random probability measure given by de Finetti's theorem is unique this proves that $P \sim \text{DP}(\alpha, P_0)$. □

Outline

Parametric mixtures

A mixture model with K components has the form

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

where the vector of weights $\pi = (\pi_1, \ldots, \pi_K)$ is a probability vector.

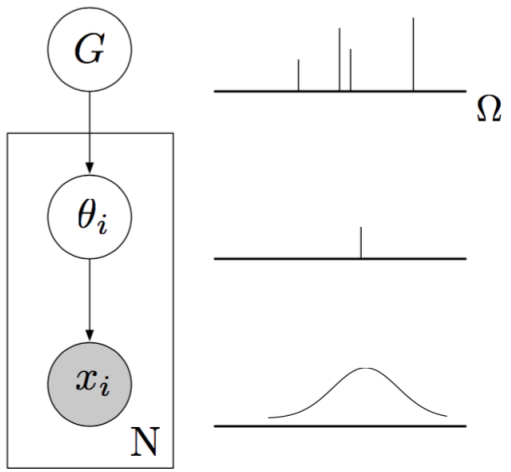
It mixes the parametric kernel (density) $p(\cdot|\phi)$ with the mixing measure

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

where δ_{ϕ_k} is a point mass (Dirac Measure) at ϕ_k.

Then

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



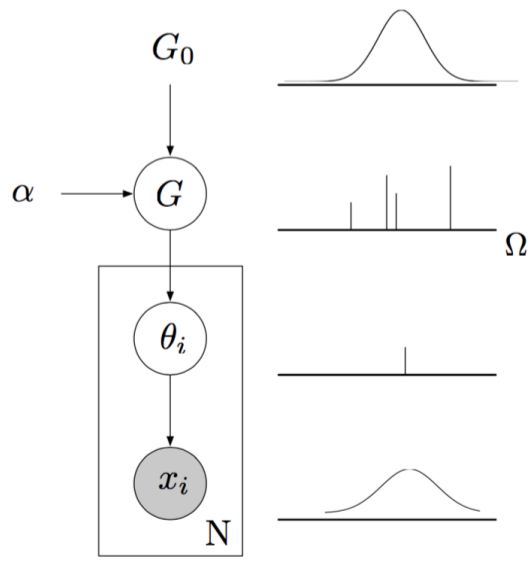
Bayesian parametric mixtures

For Bayesian mixture models, we need a distribution over the probability measure G, that is a distribution over weights $\pi = (\pi_1, \ldots, \pi_K)$ and over cluster-specific parameters ϕ_k (eg mean and covariance $\phi_k = (\mu_k, \Sigma_k)$):

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

This makes $G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$ a random probability measure (with finitely many atoms).

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



Choosing the number of components K

There are several options for choosing the number of components K

- Model selection with information criteria: AIC, BIC, or cross-validation, etc
- Hierarchical model, with a prior on K
- Nonparametric model, and let K get large... or even possibly infinite.

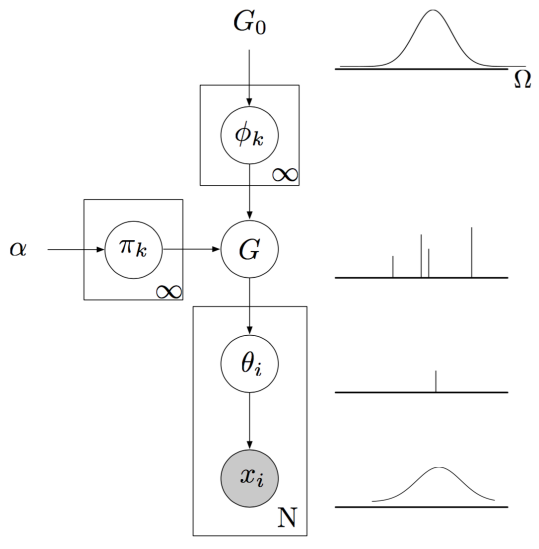
A Bayesian nonparametric approach

Bayesian nonparametric mixture models.

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

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

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



Posterior sampling

Markov chain Monte Carlo (MCMC) methods:

- Marginal methods: marginalizing over the posterior DP P, and sampling using the posterior Pólya urn scheme (easy in conjugate case, see Neal, 2000)
- Conditional methods: sampling a finite but sufficient number of parameters (Ishwaran and James, 2001). BNPdensity R package (Arbel et al., 2021).

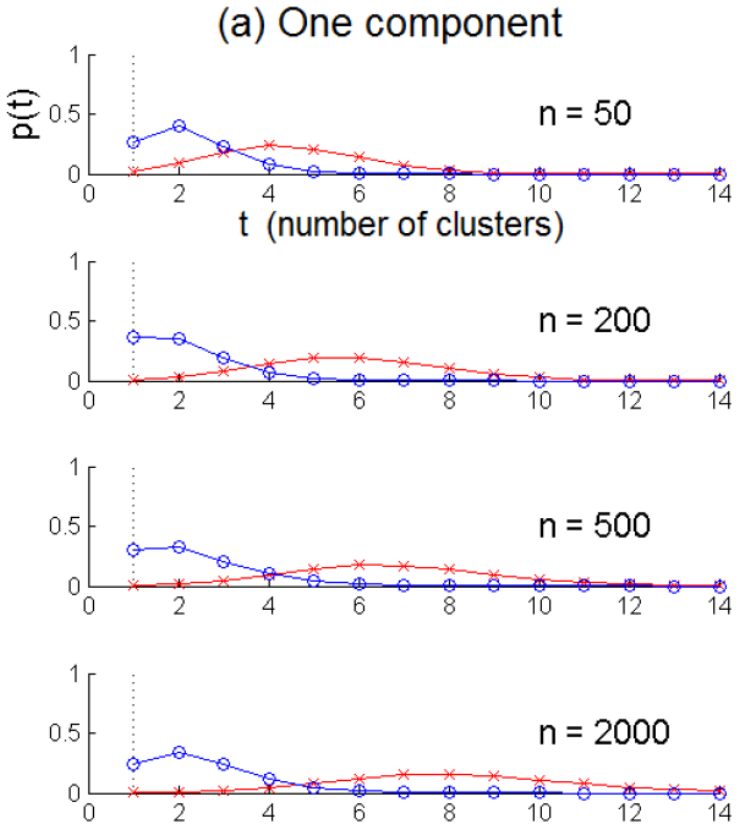
Variational approximations (David M Blei, Jordan, et al., 2006)

Warning on interpretation of K_n I

Consider a simple DP mixture model with

- Gaussian base measure,
- Gaussian kernel,
- where data are sampled iid from some distribution.

Then the posterior on K_n is inconsistent (Miller and Harrison, 2013).



Warning on interpretation of K_n II

From Miller and Harrison (2013) (here K_n is denoted T_n):

Theorem 4.1. *If $X_1, X_2, \ldots \in \mathbb{R}$ are i.i.d. from any distribution with $\mathbb{E}|X_i| < \infty$, then with probability 1, under the standard normal DPM with $\alpha = 1$ as defined above, $p(T_n = 1 \mid X_{1:n})$ does not converge to 1 as $n \to \infty$.*

Theorem 5.1. *If $X_1, X_2, \ldots \sim \mathcal{N}(0, 1)$ i.i.d. then*

$$p(T_n = 1 \mid X_{1:n}) \xrightarrow{\Pr} 0 \quad \text{as } n \to \infty$$

under the standard normal DPM with concentration parameter $\alpha = 1$.

But there is some hope...

Bayesian decision theory

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

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

Examples with Euclidean parameter spaces:

- L^2, squared loss $\longrightarrow$ posterior mean
- L^1, absolute loss $\longrightarrow$ posterior median
- $0-1$ loss $\longrightarrow$ mode a posteriori (MAP)

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

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

Examples with Euclidean parameter spaces:

- L^2, squared loss $\longrightarrow$ posterior mean
- L^1, absolute loss $\longrightarrow$ posterior median
- $0-1$ loss $\longrightarrow$ mode a posteriori (MAP)

Deriving an optimal clustering

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

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

and the decision is taken by choosing the best

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

Several losses have been considered:

- 0-1 loss (Rajkowski, 2019),
- Binder loss (Dahl, 2006),
- Variation of information (Wade and Ghahramani, 2018).

Deriving an optimal clustering

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

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

and the decision is taken by choosing the best

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

Several losses have been considered:

- 0-1 loss (Rajkowski, 2019),
- Binder loss (Dahl, 2006),
- Variation of information (Wade and Ghahramani, 2018).

Simplest loss: L_{0-1}

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

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

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

Negative results by Rajkowski (2019) show that the maximum a posteriori (MAP) is inconsistent.

Simplest loss: L_{0-1}

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

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

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

Negative results by Rajkowski (2019) show that the maximum a posteriori (MAP) is inconsistent.

Simplest loss: L_{0-1}

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

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

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

Negative results by Rajkowski (2019) show that the maximum a posteriori (MAP) is inconsistent.

Outline

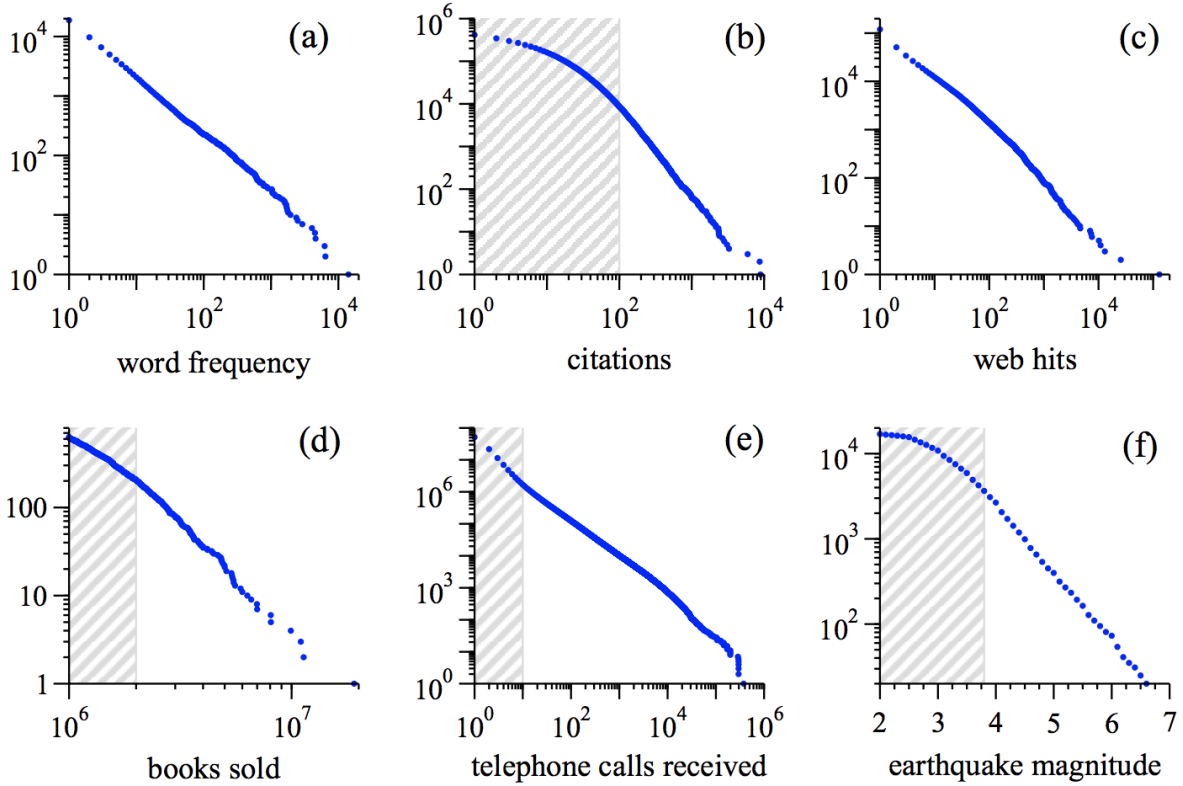
Need for a power-law for K_n

Newman (2005) and Clauset, Shalizi, and Newman (2009) show that "*Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena*".

Hence the need to depart from $K_n \sim \alpha \log n$ induced by a Dirichlet process.

Need for a power-law for K_n

Newman (2005) and Clauset, Shalizi, and Newman (2009) show that "*Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena*".



[Image from Newman (2005)]

Hence the need to depart from $K_n \sim \alpha \log n$ induced by a Dirichlet process.

Chinese restaurant process

Consider discrete data $X_1, \dots X_n | P \overset{\text{iid}}{\sim} P$, and $P \sim \mathcal{Q}$

Features $k_n \leq n$ unique values $X_1^*, \dots, X_{k_n}^*$ with resp. frequencies $n_1, \dots, n_{k_n}$

Discrete random probability measures are characterized by predictive distr.

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

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

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

Product form exchangeable partition probability function

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

Chinese restaurant process

Consider discrete data $X_1, \dots X_n | P \overset{\text{iid}}{\sim} P$, and $P \sim \mathcal{Q}$

Features $k_n \leq n$ unique values $X_1^*, \dots, X_{k_n}^*$ with resp. frequencies $n_1, \dots, n_{k_n}$

Discrete random probability measures are characterized by predictive distr.

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

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

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

Product form exchangeable partition probability function

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

Chinese restaurant process

Consider discrete data $X_1, \dots X_n | P \overset{\text{iid}}{\sim} P$, and $P \sim \mathcal{Q}$

Features $k_n \leq n$ unique values $X_1^*, \dots, X_{k_n}^*$ with resp. frequencies $n_1, \dots, n_{k_n}$

Discrete random probability measures are characterized by predictive distr.

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

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

Rate for number of clusters $k_n \asymp \begin{cases} K \text{ random variable a.s. finite if } \sigma < 0 \\ \alpha \log n \text{ if } \sigma = 0 \\ S n^\sigma \text{ if } \sigma \in (0,1), (S \text{ random variable}). \end{cases}$

Product form exchangeable partition probability function

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

Beyond the DP from predictive function viewpoint

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

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

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

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

Beyond the DP from predictive function viewpoint

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

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

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

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

Beyond the DP from predictive function viewpoint

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

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

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

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

Beyond the DP from predictive function viewpoint

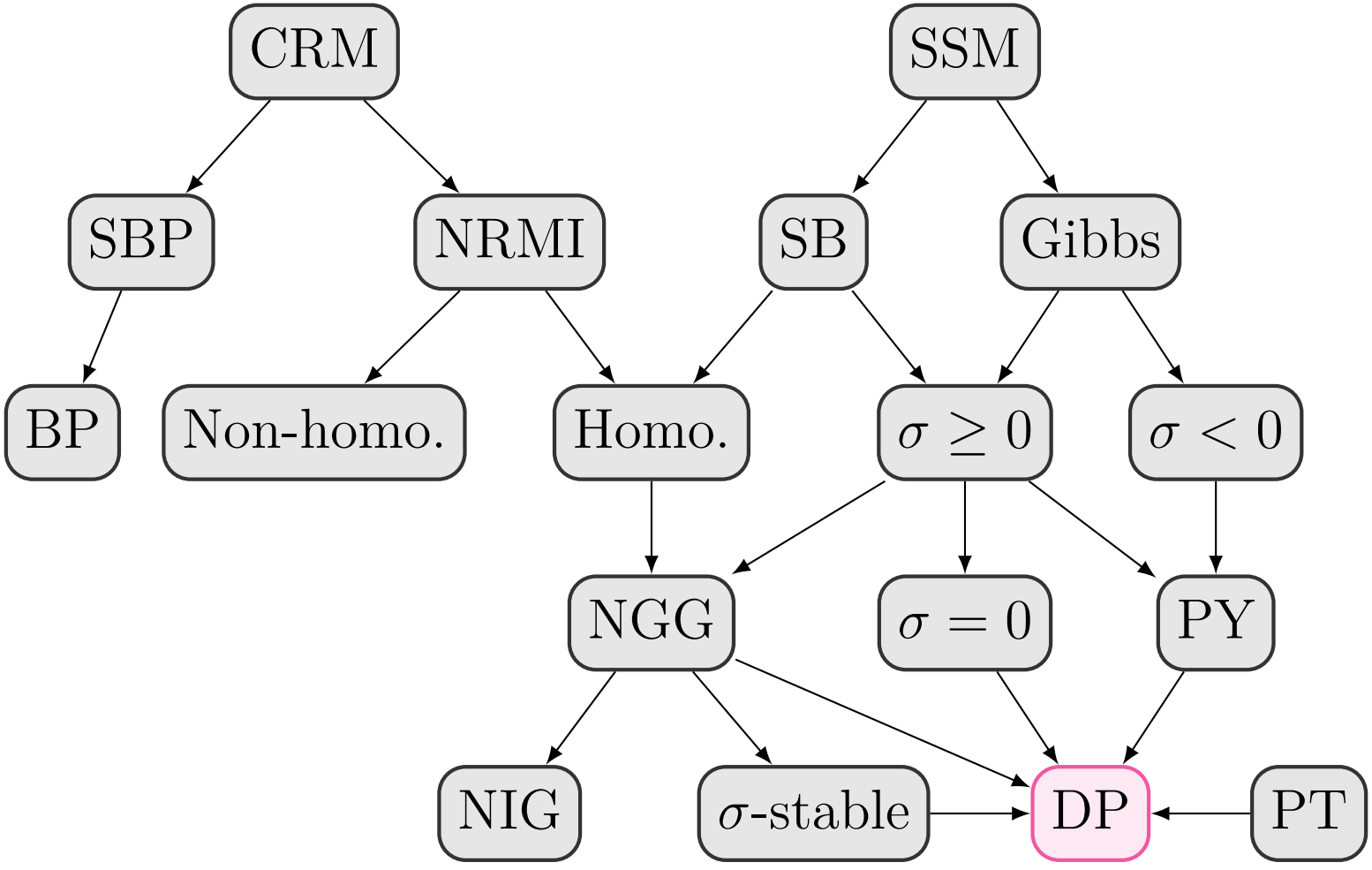
A discrete random probability measure P can be classified in 3 main categories according to $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n]$

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

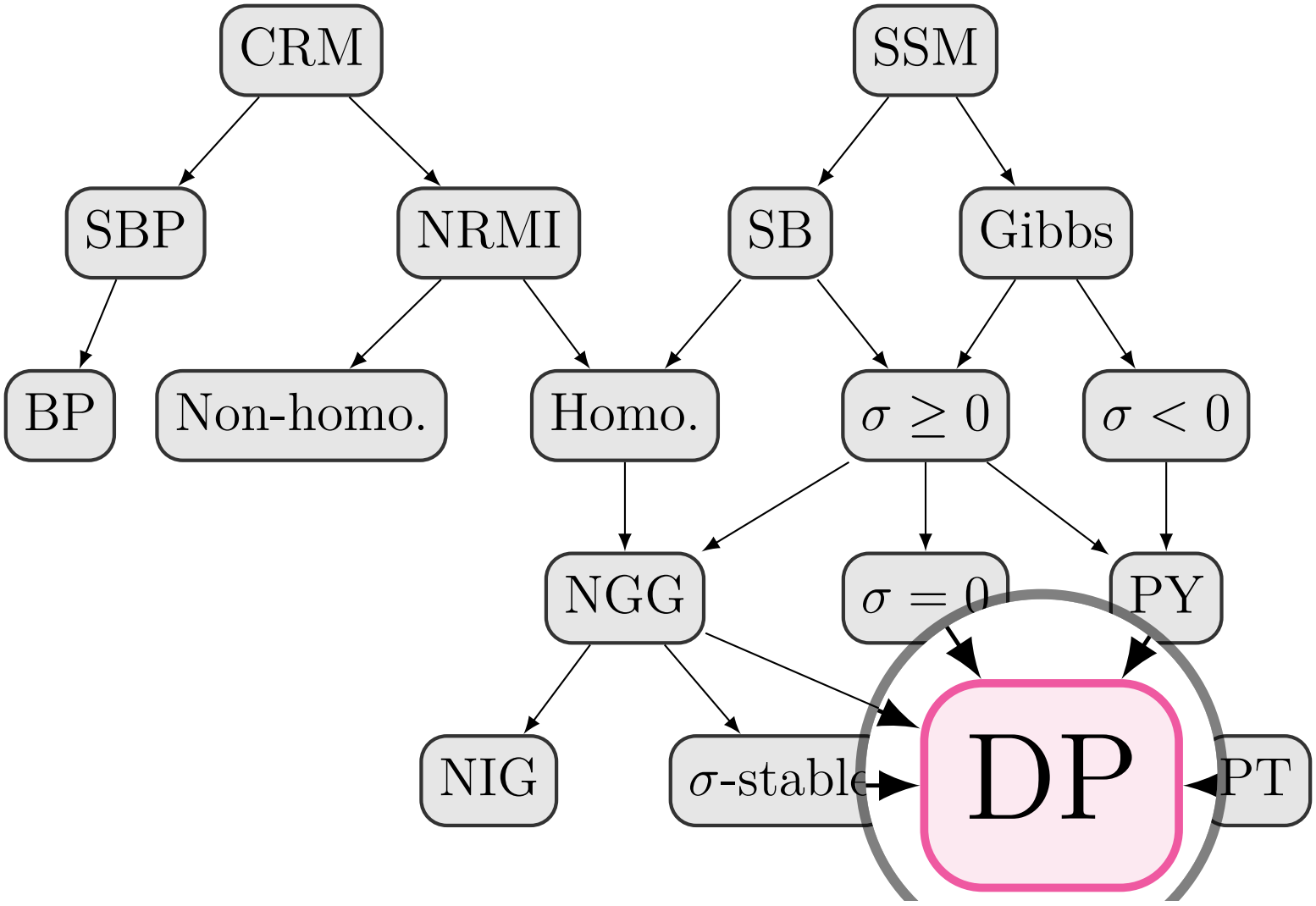
2) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, \text{model parameters})$
$\iff$ depends on n and k_n but not on $(n_1, \ldots, n_{k_n})$
$\iff$ Gibbs-type prior (Pitman, 2003);

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

Tree of discrete random probability measures



Tree of discrete random probability measures



Proposition (Pitman Sampling formula)

The multiplicities $(m_1, \ldots, m_n)$ *in* $X_1, \ldots, X_n | P \overset{iid}{\sim} P,\ P \sim PY(\sigma, \alpha, P_0)$ *have distribution*

$$p(m_1, \ldots, m_n) = \frac{n!}{(1+\alpha)_{(n-1)}} (\alpha+\sigma) \cdots (\alpha+(k-1)\sigma) \prod_{\ell=1}^{n} \frac{1}{m_\ell!} \left(\frac{(1-\sigma)_{(\ell-1)}}{\ell!} \right)^{m_\ell}$$

Proof. Same technique as for the DP Ewens sampling formula.

Pitman–Yor process II

Proposition (Power law and σ-diversity)

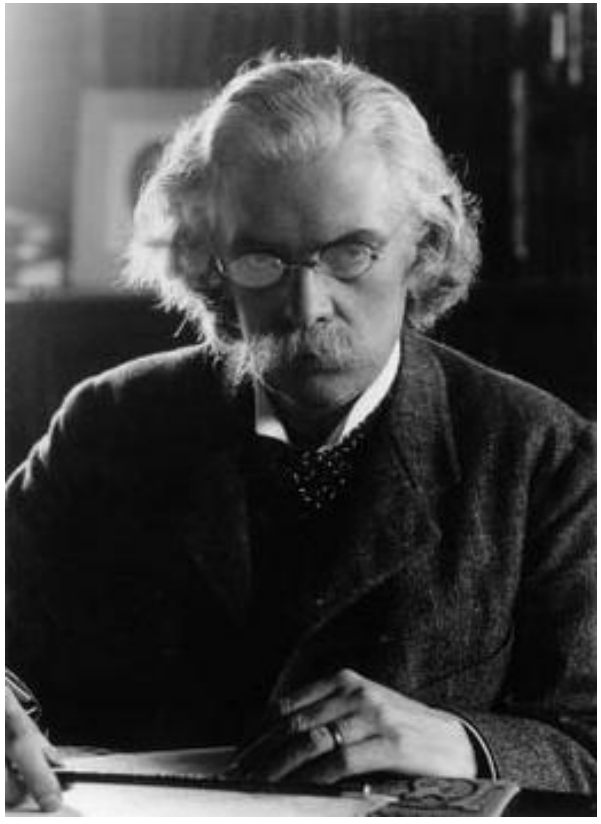
For $\sigma > 0$ we have the almost sure convergence

$$n^{-\sigma} K_n \to S_{\sigma,\alpha},$$

where $S_{\sigma,\alpha}$ is called σ-diversity of the PY, whose density is a polynomially tilted Mittag–Leffler density (ML):

$$g_{\sigma,\alpha}(x) \propto x^{\alpha/\sigma} g_\alpha(x),$$

and g_α is ML density.



[Image: Wikipedia]

Pitman–Yor process III

Theorem (Stick breaking representation for PY)

If $V_j \overset{ind}{\sim} Be(1-\sigma, \alpha + j\sigma)$ *and* $p_1 = V_1,\ p_j = V_j \prod_{l<j}(1-V_l)$ *and further we have* $\phi_j \overset{iid}{\sim} P_0$ *then*

$$P = \sum_{j=1}^{\infty} p_j \delta_{\phi_j} \sim PY(\sigma, \alpha P_0).$$

Pitman–Yor process IV

Proposition (Moments of PY)

If $P \sim PY(\sigma, \alpha, P_0)$, then for every measurable sets A, B we have

1) $\mathrm{E}[P(A)] = P_0(A)$,
2) $\mathrm{E}[P(A)P(B)] = (1-\sigma)/(1+\alpha)P_0(A \cap B) + (\alpha+\sigma)/(1+\alpha)P_0(A)P_0(B)$,
3) $\mathrm{Cov}[P(A), P(B)] = (1-\sigma)/(1+\alpha)\big(P_0(A \cap B) - P_0(A)P_0(B)\big)$.

Proof.

1) We use the stick-breaking representation:

$$\mathrm{E}P(A) = \sum_j \mathrm{E}p_j \mathrm{E}\delta_{\phi_j} = \sum_j \mathrm{E}(p_j) P_0(A) = P_0(A)\mathrm{E}(\sum_j p_j) = P_0(A).$$

2) Let $X_1, X_2 | P \overset{\text{iid}}{\sim} P$, then

$$\mathrm{E}(P(A)P(B)) = \mathbb{P}(X_1 \in A, X_2 \in B) = \mathbb{P}(X_1 \in A)\mathbb{P}(X_2 \in B | X_1 \in A).$$

Lets investigate two terms above: from 1) we know that $\mathbb{P}(X_1 \in A) = P_0(A)$. We know the predictive of PY:

$$X_2 | X_1 \sim \frac{\alpha + \sigma}{\alpha + 1} P_0 + \frac{1 - \sigma}{\alpha + 1} \delta_{X_1},$$

and hence

$$\mathbb{P}(X_2 \in B | X_1 \in A) = \frac{\alpha + \sigma}{\alpha + 1} P_0(B) + \frac{1 - \sigma}{\alpha + 1} P_{0A}(B),$$

when we used notation $P_{0A}(B) = P_0(B|A) = P_0(A \cap B)/P_0(A)$ for a conditional measure.

3) It is straightforward combination of 1) and 2).

Unlike the DP, the PY is not conjugate. However, the posterior can be explicited.

Theorem (Posterior distribution of PY)

If $P \sim PY(\sigma, \alpha, P_0)$ then the posterior of P based on observations $X_{1:n}|P \overset{iid}{\sim} P$ has the distribution of the random probability measure

$$(1 - q_n)P_n + q_n \sum_{j=1}^{K_n} p_j^* \delta_{X_j^*},$$

where $X_{1:n}^*$ are the K_n distinct values in $X_{1:n}$, frequencies are denoted $n_1, \ldots, n_{K_n}$ and

- $q_n \sim Beta(n - K_n\sigma, \alpha + K_n\sigma)$,
- $(p_1^*, \ldots, p_{K_n}^*) \sim Dir(n_1 - \sigma, \ldots, n_{K_n} - \sigma)$,
- $P_n \sim PY(\sigma, (\alpha + \sigma K_n)P_0)$.

Impact of the stability parameter σ

Prior distribution of the number of clusters K_n

- α controls the location (as for the DP)
- σ controls the flatness (or variability)

Impact of the stability parameter σ

Prior distribution of the number of clusters K_n

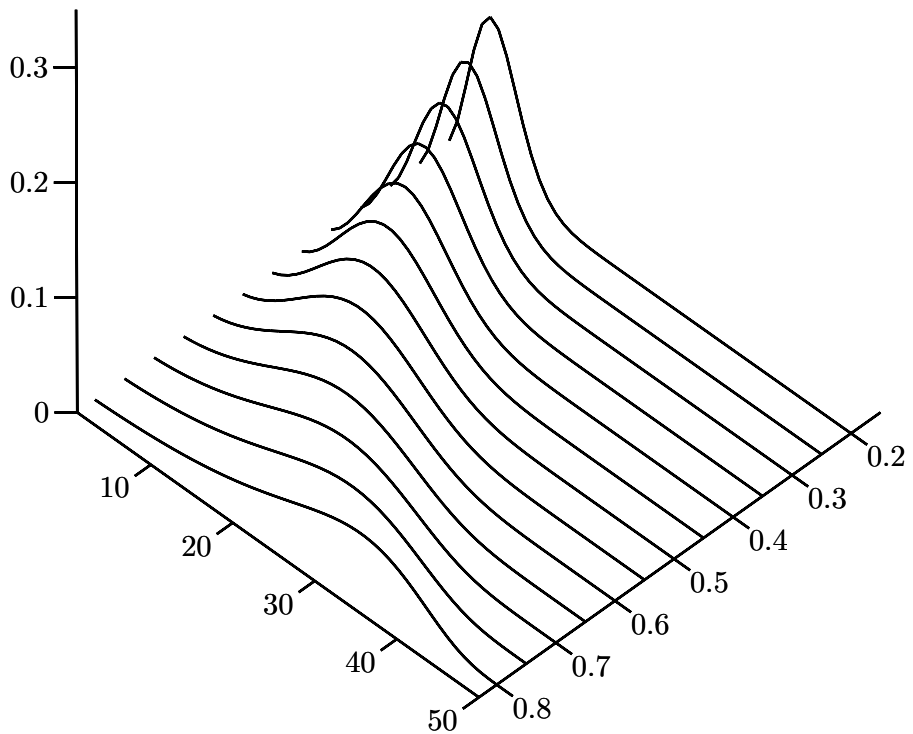
- α controls the location (as for the DP)
- σ controls the flatness (or variability)

Impact of the stability parameter σ

Prior distribution of the number of clusters K_n

- α controls the location (as for the DP)
- σ controls the flatness (or variability)

With $n = 50, \alpha = 1$ and $\sigma = 0.2, 0.3, \ldots, 0.8$ [Image from De Blasi et al. (2015)]



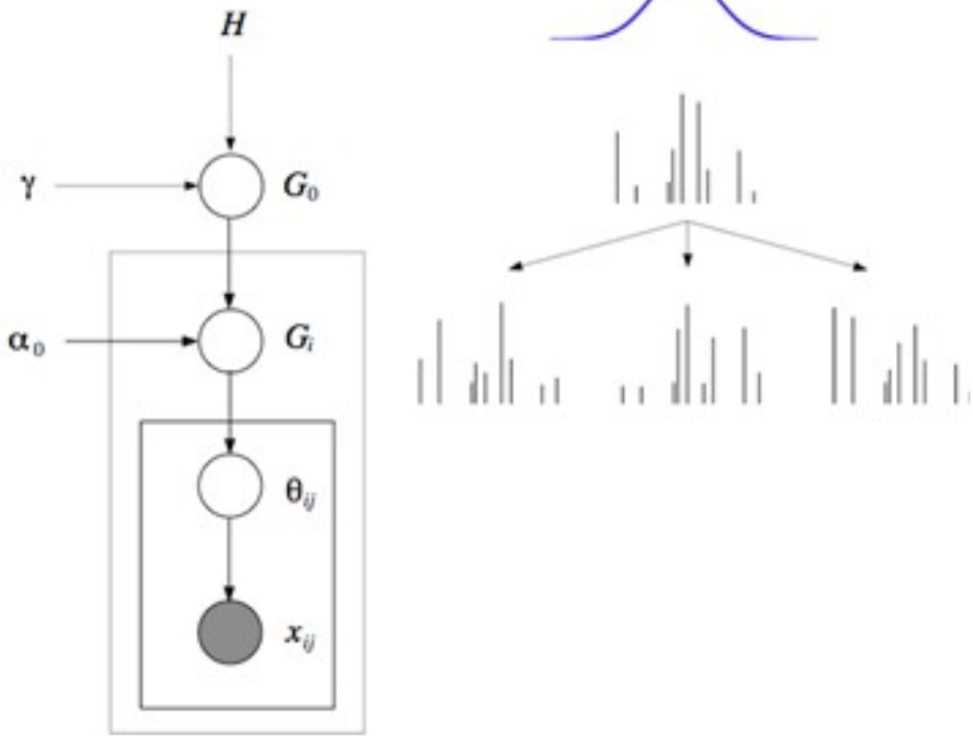
Outline

Hierarchical Dirichlet process

Remember that Latent Dirichlet Allocation (LDA, David M Blei, Ng, and Jordan, 2003) is a probabilistic model used for topic modeling. It assumes that each document is a mixture of various topics, and each topic is a mixture of various words. The goal of LDA is to uncover these topics given a collection of documents.

Hierarchical Dirichlet Process (HDP) due to Teh et al. (2006) is an extension of the Dirichlet process, which is a way of modeling distributions over an unknown number of groups or clusters. The HDP allows for an infinite number of topics to be inferred from a collection of documents, meaning it can automatically determine the appropriate number of topics rather than requiring it to be specified beforehand.

Hierarchical Dirichlet process



$$
\begin{aligned}
G_0|\gamma, H &\sim DP(\gamma H) \\
G_i|\alpha, G_0 &\sim DP(\alpha_0 G_0) \\
\theta_{ij}|G_i &\sim G_i \\
x_{ij}|\theta_{ij} &\sim F(x_{ij}|\theta_{ij})
\end{aligned}
$$

[Image by M. Jordan]

HDP. Chinese Restaurant Franchise

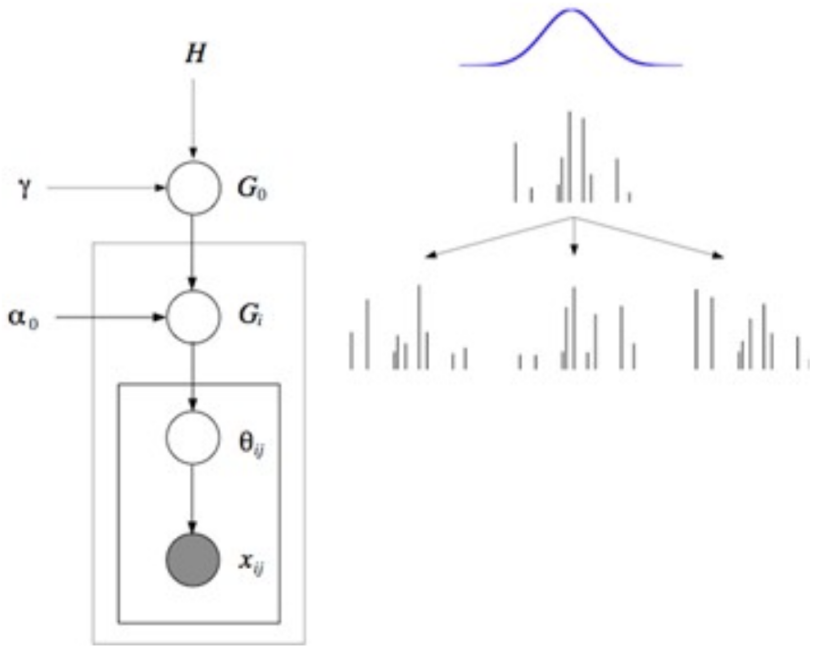
The partition distribution of the HDP is called Chinese Restaurant Franchise. It is corresponds to the marginal clustering distribution obtained once the process is integrated out. The generative process is as follows:

- Franchise Aspect: each restaurant can be seen as its own "franchise" of the overall restaurant chain, with its own set of customers.
- Restaurant Metaphor: each restaurant has an infinite number of tables, each representing a different topic or cluster. Each table can accommodate an infinite number of customers.
- Customers: these represent data points, such as words in documents. Each customer needs to choose a table to sit at.
- Choosing a Table: in the same way as for the Chinese Restaurant process.

HDP. Stick-breaking representation

Each process of the processes of the HDP admit a stick-breaking representation, both the first layer G_0 and the second-layer G_i's.

- What can you say of the SB of the second-layer G_i's?
- Hint: G_0 is itself a draw from a DP, meaning it has its own SB representation.

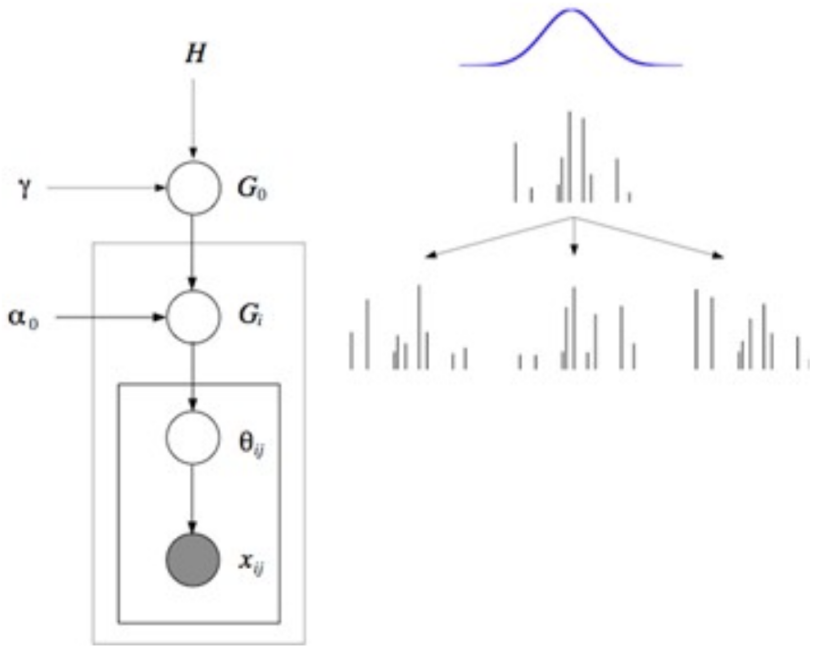


$$
\begin{aligned}
G_0|\gamma, H &\sim DP(\gamma H) \\
G_i|\alpha, G_0 &\sim DP(\alpha_0 G_0) \\
\theta_{ij}|G_i &\sim G_i \\
x_{ij}|\theta_{ij} &\sim F(x_{ij}|\theta_{ij})
\end{aligned}
$$

HDP. Stick-breaking representation

Each process of the processes of the HDP admit a stick-breaking representation, both the first layer G_0 and the second-layer G_i's.

- What can you say of the SB of the second-layer G_i's?
- Hint: G_0 is itself a draw from a DP, meaning it has its own SB representation.

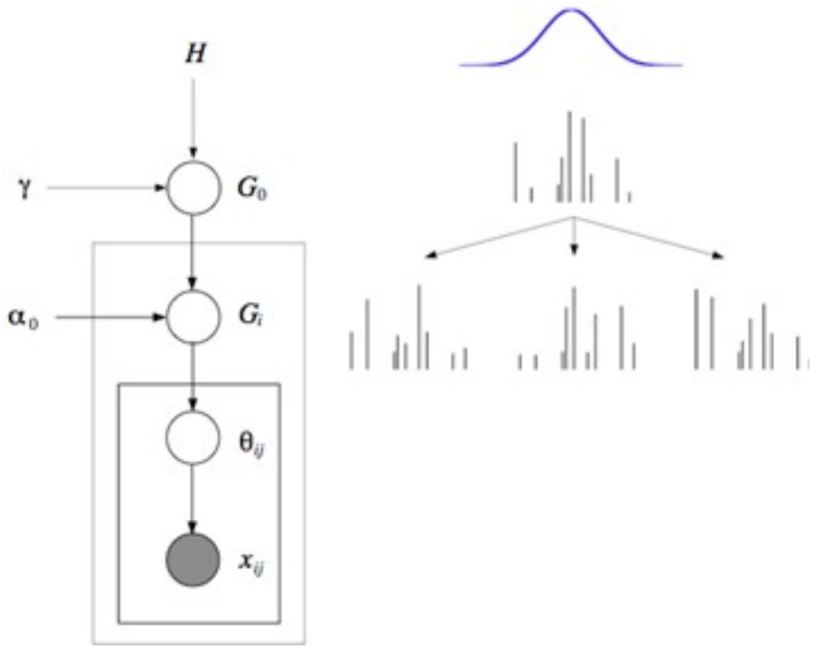


$$
\begin{aligned}
G_0|\gamma, H &\sim DP(\gamma H) \\
G_i|\alpha, G_0 &\sim DP(\alpha_0 G_0) \\
\theta_{ij}|G_i &\sim G_i \\
x_{ij}|\theta_{ij} &\sim F(x_{ij}|\theta_{ij})
\end{aligned}
$$

HDP. Stick-breaking representation

Each process of the processes of the HDP admit a stick-breaking representation, both the first layer G_0 and the second-layer G_i's.

- What can you say of the SB of the second-layer G_i's?
- Hint: G_0 is itself a draw from a DP, meaning it has its own SB representation.



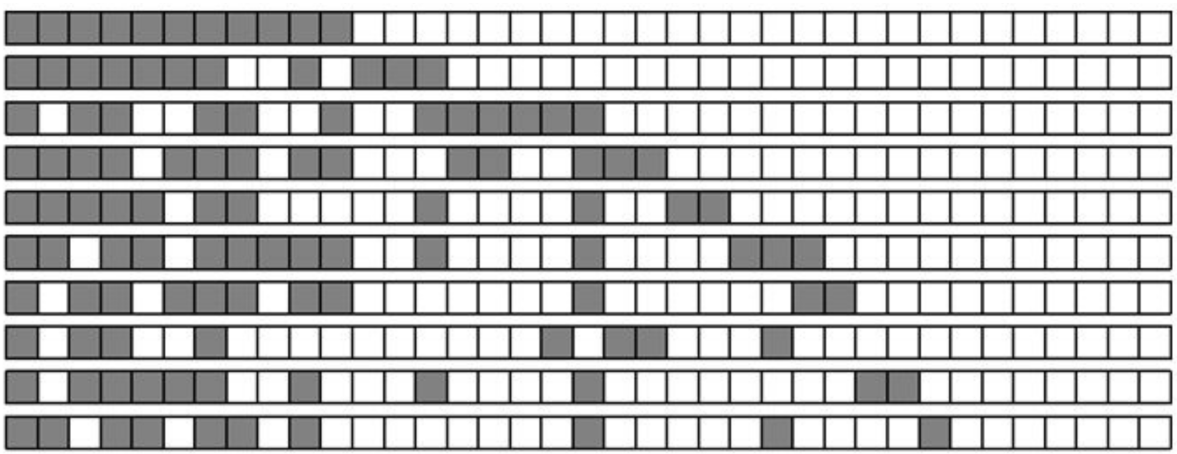
$$
\begin{aligned}
G_0|\gamma, H &\sim DP(\gamma H) \\
G_i|\alpha, G_0 &\sim DP(\alpha_0 G_0) \\
\theta_{ij}|G_i &\sim G_i \\
x_{ij}|\theta_{ij} &\sim F(x_{ij}|\theta_{ij})
\end{aligned}
$$

Indian Buffet process

Feature allocation model by Ghahramani and Griffiths (2006), where observations may share several features. Note how different this is with clustering where each observation is assigned to one and only one cluster.

Generative model is as follows:

- first customer samples Poisson(γ) dishes.
- second customer chooses every dish of first customer wp $1/2$, plus Poisson($\gamma/2$) new dishes.
- ...
- nth step: K dishes have been sampled, each by $m_1, \ldots, m_K$ previous customers, the customer chooses jth dish wp m_j/n, plus Poisson(γ/n) new dishes.



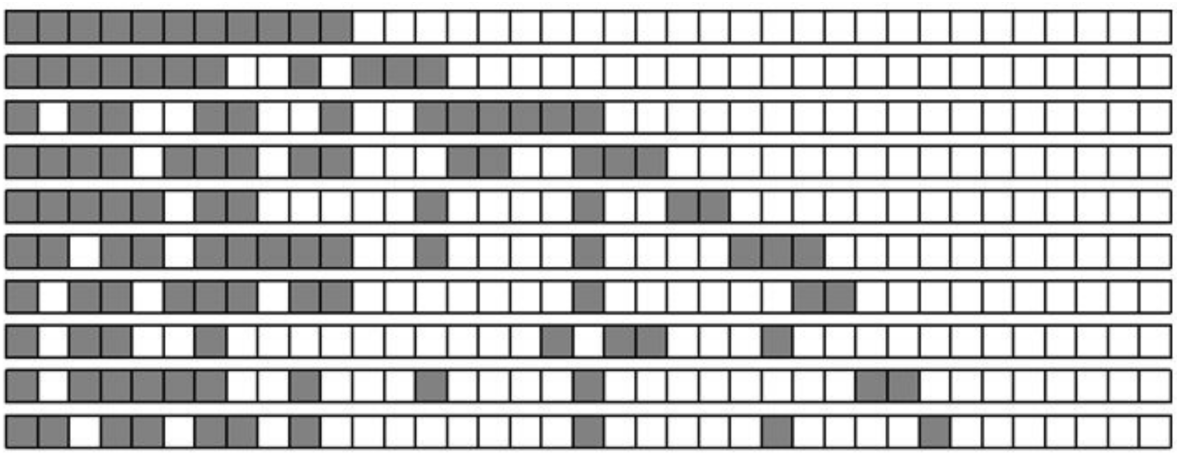
[Image by M. Jordan]

Indian Buffet process

Feature allocation model by Ghahramani and Griffiths (2006), where observations may share several features. Note how different this is with clustering where each observation is assigned to one and only one cluster.

Generative model is as follows:

- first customer samples Poisson(γ) dishes.
- second customer chooses every dish of first customer *wp* 1/2, plus Poisson($\gamma/2$) new dishes.
- ...
- ith step: K dishes have been sampled, each by $m_1, \ldots, m_K$ previous customers. ith customer chooses jth dish *wp* m_j/i, plus Poisson(γ/i) new dishes.



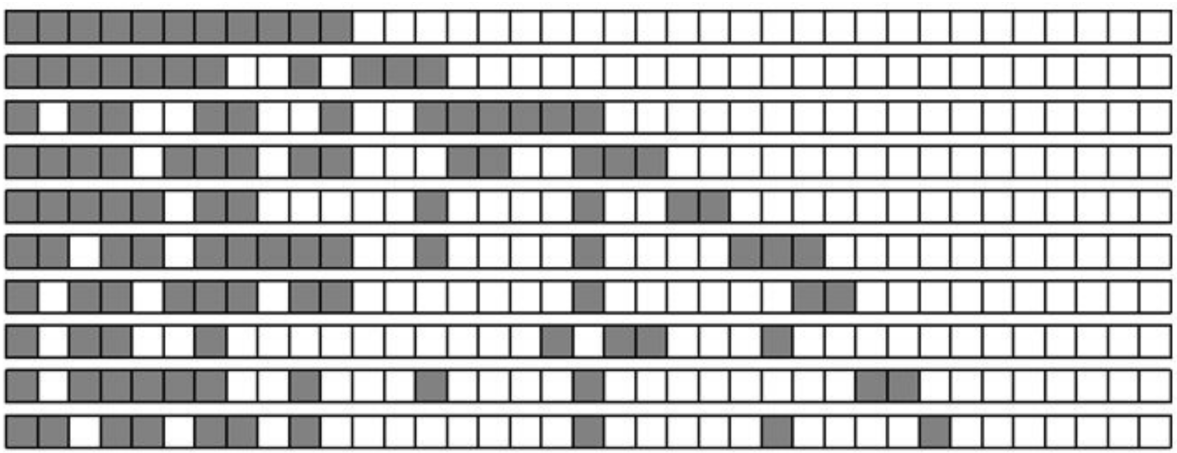
[Image by M. Jordan]

Indian Buffet process

Feature allocation model by Ghahramani and Griffiths (2006), where observations may share several features. Note how different this is with clustering where each observation is assigned to one and only one cluster.

Generative model is as follows:

- first customer samples Poisson(γ) dishes.
- second customer chooses every dish of first customer *wp* $1/2$, plus Poisson($\gamma/2$) new dishes.
- ...
- ith step: K dishes have been sampled, each by $n_1, \ldots, n_K$ customers; ith customer chooses jth dish *wp* n_j/i, plus Poisson(γ/i) new dishes.



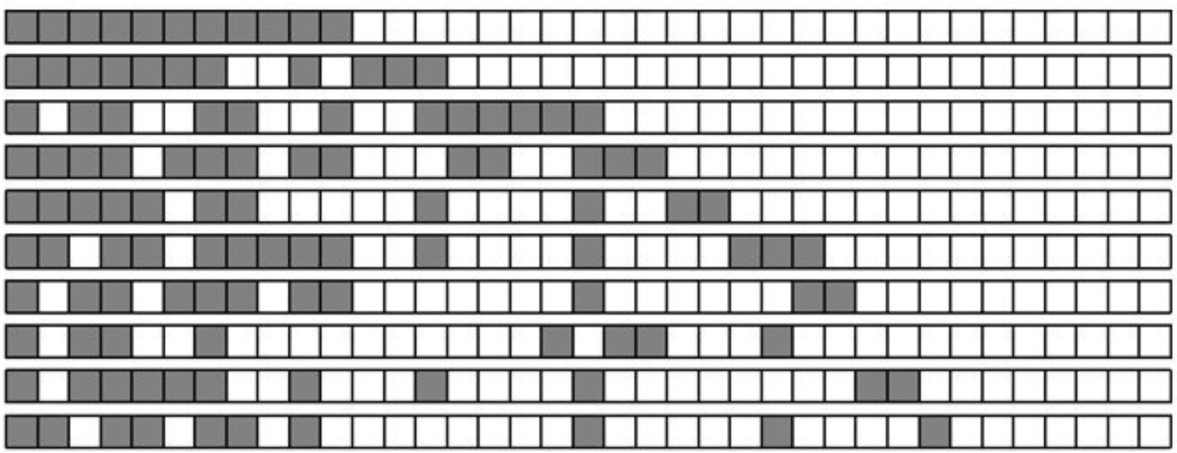
[Image by M. Jordan]

Indian Buffet process

Feature allocation model by Ghahramani and Griffiths (2006), where observations may share several features. Note how different this is with clustering where each observation is assigned to one and only one cluster.

Generative model is as follows:

- first customer samples Poisson(γ) dishes.
- second customer chooses every dish of first customer *wp* $1/2$, plus Poisson($\gamma/2$) new dishes.
- ...
- ith step: K dishes have been sampled, each by $n_1, \ldots, n_K$ customers; ith customer chooses jth dish *wp* n_j/i, plus Poisson(γ/i) new dishes.



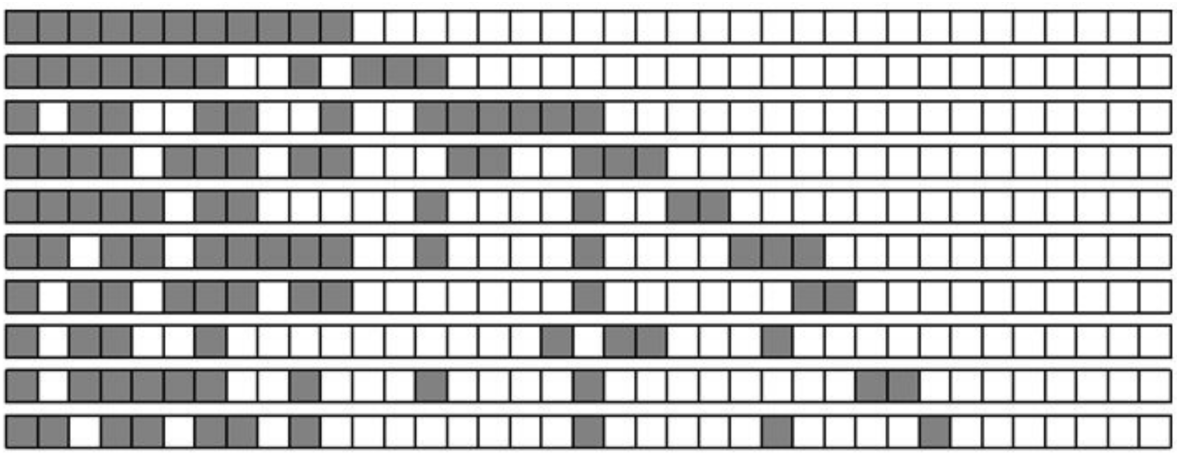
[Image by M. Jordan]

Indian Buffet process

Feature allocation model by Ghahramani and Griffiths (2006), where observations may share several features. Note how different this is with clustering where each observation is assigned to one and only one cluster.

Generative model is as follows:

- first customer samples Poisson(γ) dishes.
- second customer chooses every dish of first customer *wp* $1/2$, plus Poisson($\gamma/2$) new dishes.
- ...
- ith step: K dishes have been sampled, each by $n_1, \ldots, n_K$ customers; ith customer chooses jth dish *wp* n_j/i, plus Poisson(γ/i) new dishes.



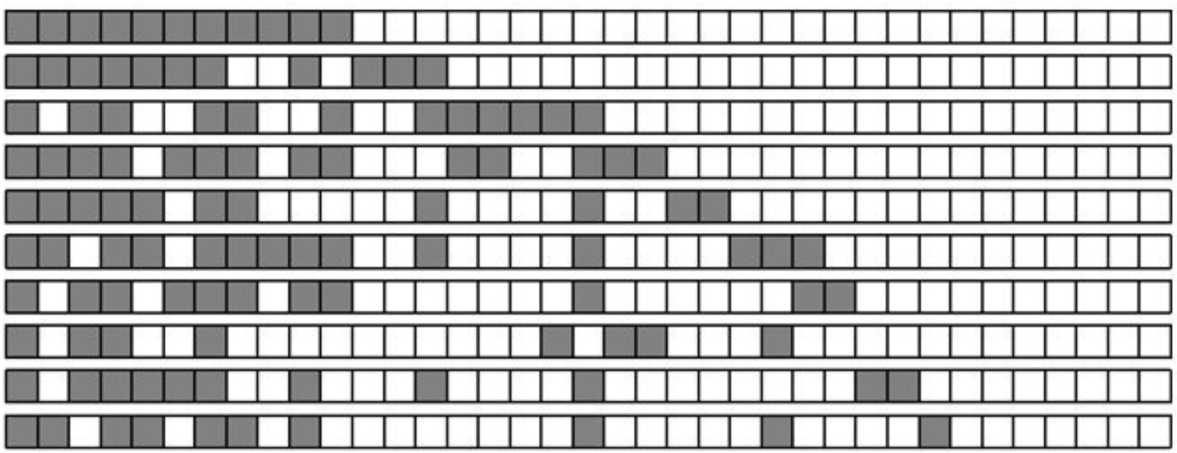
[Image by M. Jordan]

Indian Buffet process

Feature allocation model by Ghahramani and Griffiths (2006), where observations may share several features. Note how different this is with clustering where each observation is assigned to one and only one cluster.

Generative model is as follows:

- first customer samples Poisson(γ) dishes.
- second customer chooses every dish of first customer *wp* $1/2$, plus Poisson($\gamma/2$) new dishes.
- ...
- ith step: K dishes have been sampled, each by $n_1, \ldots, n_K$ customers; ith customer chooses jth dish *wp* n_j/i, plus Poisson(γ/i) new dishes.



[Image by M. Jordan]

IBP. Growth of K_n

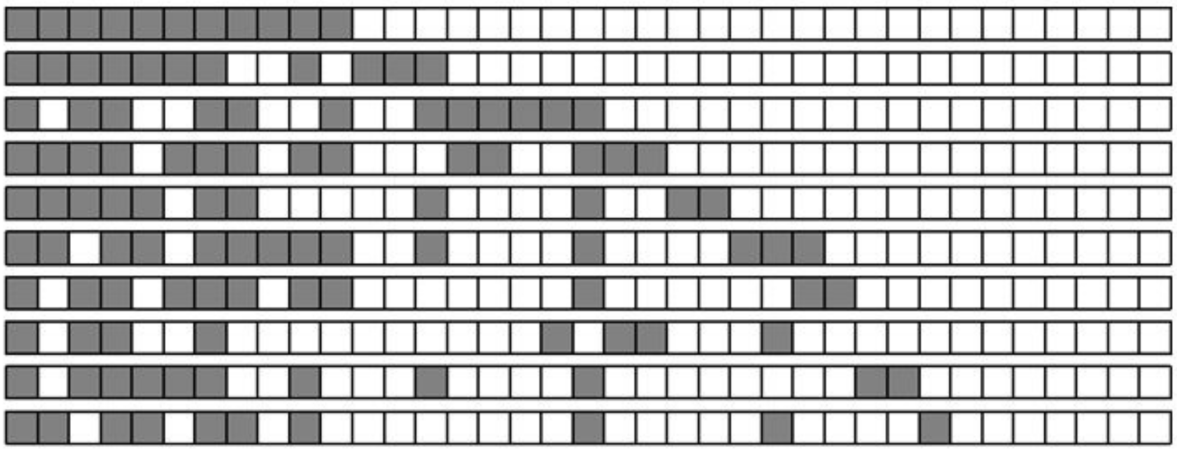
Each customer samples Poisson(γ) dishes.

How does K_n, the number of different dishes, grow as $n \to \infty$?

Use the additivity property of Poisson to derive:

$$K_n \sim \text{Poisson}(\gamma) + \text{Poisson}(\gamma/2) + \cdots + \text{Poisson}(\gamma/n)$$

$$= \text{Poisson}\left(\gamma\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right)\right) \asymp \text{Poisson}(\gamma \log n).$$



[Image by M. Jordan]

IBP. Growth of K_n

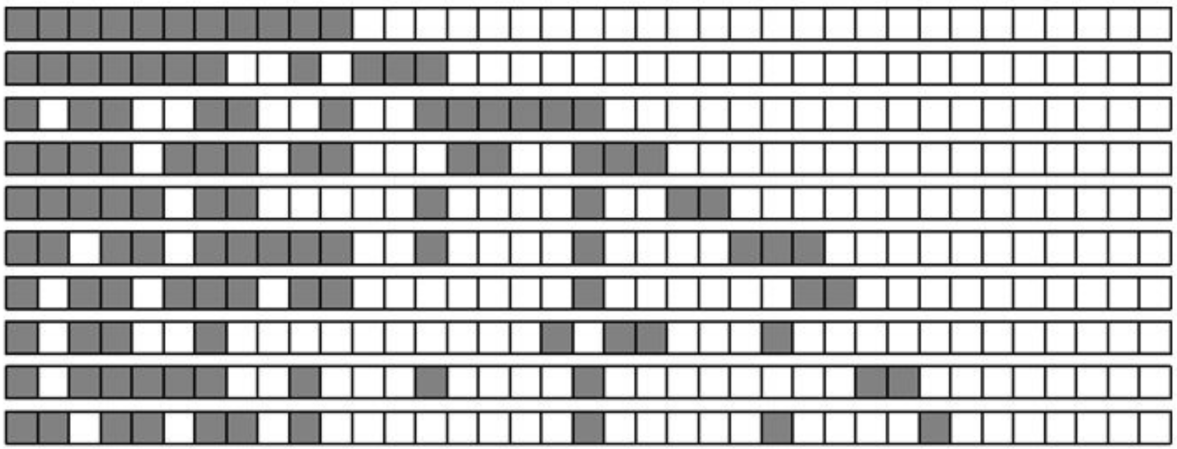
Each customer samples Poisson(γ) dishes.

How does K_n, the number of different dishes, grow as $n \to \infty$?

Use the additivity property of Poisson to derive:

$$K_n \sim \text{Poisson}(\gamma) + \text{Poisson}(\gamma/2) + \cdots + \text{Poisson}(\gamma/n)$$

$$= \text{Poisson}\left(\gamma\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right)\right) \asymp \text{Poisson}(\gamma \log n).$$



[Image by M. Jordan]

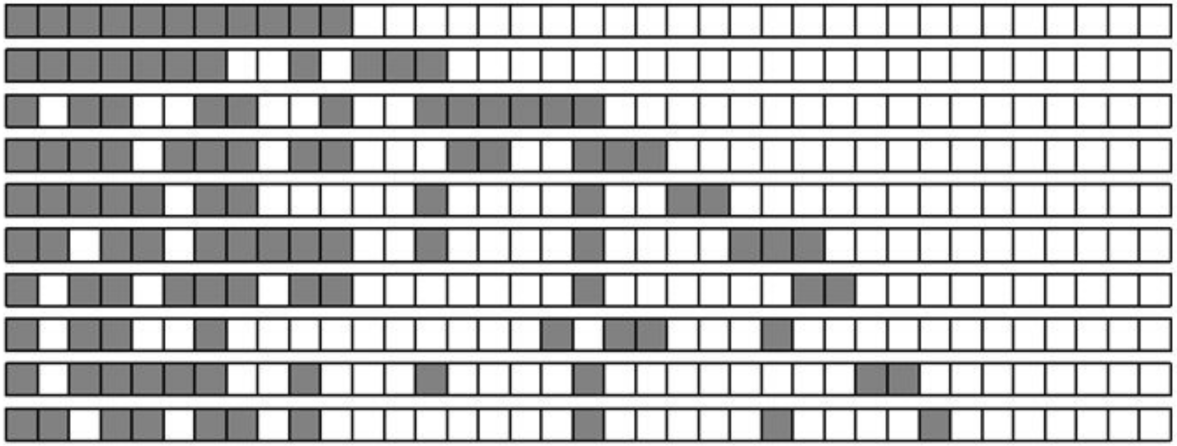
IBP. Growth of K_n

Each customer samples Poisson(γ) dishes.

How does K_n, the number of different dishes, grow as $n \to \infty$?

Use the additivity property of Poisson to derive:

$$K_n \sim \text{Poisson}(\gamma) + \text{Poisson}(\gamma/2) + \cdots + \text{Poisson}(\gamma/n)$$
$$= \text{Poisson}\left(\gamma\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right)\right) \asymp \text{Poisson}(\gamma \log n).$$



[Image by M. Jordan]

Three-parameter IBP

In real-data applications, the log-growth of the number of dishes may not be adapted. How to tweak the IBP generative process in order to sample new dishes more often (ideally, at a power-law rate)?

Think as for the PY and Gibbs-type processes with the addition of a discount parameter $\sigma \in (0, 1)$.

Three-parameter IBP

In real-data applications, the log-growth of the number of dishes may not be adapted. How to tweak the IBP generative process in order to sample new dishes more often (ideally, at a power-law rate)?

Think as for the PY and Gibbs-type processes with the addition of a discount parameter $\sigma \in (0, 1)$.

Outline

Outline

What comes to *your* mind when you hear "Asymptotics"?

Why Asymptotics

- Construction of a prior on a nonparametric space is difficult
- We cannot hope to cover all the space of density (for example) with our prior (the prior does not have full support)
- We need to check that our inference is not completely off!

Parametric setting

We have the celebrated Bernstein-von Mises theorem that implies that the effect of the prior on the posterior inference vanishes when the amount of information grows.

This is not true anymore in a nonparametric setting!

Why Asymptotics

- Construction of a prior on a nonparametric space is difficult
- We cannot hope to cover all the space of density (for example) with our prior (the prior does not have full support)
- We need to check that our inference is not completely off!

Parametric setting

We have the celebrated Bernstein-von Mises theorem that implies that the effect of the prior on the posterior inference vanishes when the amount of information grows.

This is not true anymore in a nonparametric setting!

Why Asymptotics

- Construction of a prior on a nonparametric space is difficult
- We cannot hope to cover all the space of density (for example) with our prior (the prior does not have full support)
- We need to check that our inference is not completely off!

Parametric setting

We have the celebrated Bernstein-von Mises theorem that implies that the effect of the prior on the posterior inference vanishes when the amount of information grows.

This is not true anymore in a nonparametric setting!

Why Asymptotics

- Construction of a prior on a nonparametric space is difficult
- We cannot hope to cover all the space of density (for example) with our prior (the prior does not have full support)
- We need to check that our inference is not completely off!

Parametric setting

We have the celebrated Bernstein-von Mises theorem that implies that the effect of the prior on the posterior inference vanishes when the amount of information grows.

This is not true anymore in a nonparametric setting!

Why Asymptotics

- Construction of a prior on a nonparametric space is difficult
- We cannot hope to cover all the space of density (for example) with our prior (the prior does not have full support)
- We need to check that our inference is not completely off!

Parametric setting

We have the celebrated Bernstein-von Mises theorem that implies that the effect of the prior on the posterior inference vanishes when the amount of information grows.

This is not true anymore in a nonparametric setting!

Why Asymptotics

A first order approximation is to consider the asymptotic setting:

- Adopt a Frequentist point of view: "There exists a *true* parameter θ_0, and we study the posterior distribution with data generated w.r.t. θ_0."
- Ideally, the posterior distribution will concentrate around θ_0 when $n \to \infty$.

Why Asymptotics

A first order approximation is to consider the asymptotic setting:

- Adopt a Frequentist point of view: "There exists a *true* parameter θ_0, and we study the posterior distribution with data generated w.r.t. θ_0."
- Ideally, the posterior distribution will concentrate around θ_0 when $n \to \infty$.

Why Asymptotics

A first order approximation is to consider the asymptotic setting:

- Adopt a Frequentist point of view: "There exists a *true* parameter θ_0, and we study the posterior distribution with data generated w.r.t. θ_0."
- Ideally, the posterior distribution will concentrate around θ_0 when $n \to \infty$.

References

- J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. New York: Springer, 2003
- Nils Lid Hjort et al. *Bayesian nonparametrics*. Vol. 28. Cambridge University Press, Apr. 2010. URL: http://www.cambridge.org/us/academic/subjects/statistics-probability/statistical-theory-and-methods/bayesian-nonparametrics
- Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press, 2017

Outline

Consistency

Setting:

- $\forall n \in \mathbb{N}$, let X^n be some observations in a sample space $\{\mathcal{X}^n, \mathcal{A}^n\}$ with distribution P_θ
- $\theta \in \Theta$ with (Θ, d) a (semi-)metric space

Let Π be a prior distribution on Θ and $\Pi(\cdot|X^n)$ a version of its posterior distribution.

Definition (Consistency)

The posterior distribution $\Pi(\cdot|X^n)$ is said to be weakly consistent at θ_0 if for all $\epsilon > 0$

$$\Pi\left(d(\theta, \theta_0) > \epsilon | X^n\right) \xrightarrow[n\to\infty]{P_{\theta_0}} 0.$$

If the convergence is almost sure, then the posterior is said to be strongly consistent.

Consistency

Setting:

- $\forall n \in \mathbb{N}$, let X^n be some observations in a sample space $\{\mathcal{X}^n, \mathcal{A}^n\}$ with distribution P_θ
- $\theta \in \Theta$ with (Θ, d) a (semi-)metric space

Let Π be a prior distribution on Θ and $\Pi(\cdot|X^n)$ a version of its posterior distribution.

Definition (Consistency)

The posterior distribution $\Pi(\cdot|X^n)$ is said to be weakly consistent at θ_0 if for all $\epsilon > 0$

$$\Pi\left(d(\theta, \theta_0) > \epsilon | X^n\right) \xrightarrow[n\to\infty]{P_{\theta_0}} 0.$$

If the convergence is almost sure, then the posterior is said to be strongly consistent.

Point estimators

Naturally one will hope that posterior consistency implies that some summary of the posterior location would be a consistent estimator.

Theorem

Let $\Pi(\cdot|X^n)$ be a posterior distribution on Θ and suppose that it is consistent at θ_0 relative to a metric d on Θ. For $\alpha \in (0,1)$, define $\hat{\theta}_n$ as the centre of the smallest ball containing at least α of the posterior mass. Then

$$d(\hat{\theta}_n, \theta_0) \xrightarrow[n\to\infty]{P_{\theta_0}, \text{ or } P_{\theta_0} \text{ a.s.}} 0.$$

Naturally one will hope that posterior consistency implies that some summary of the posterior location would be a consistent estimator.

Theorem

Let $\Pi(\cdot|X^n)$ be a posterior distribution on Θ and suppose that it is consistent at θ_0 relative to a metric d on Θ. For $\alpha \in (0,1)$, define $\hat{\theta}_n$ as the centre of the smallest ball containing at least α of the posterior mass. Then

$$d(\hat{\theta}_n, \theta_0) \xrightarrow[n\to\infty]{P_{\theta_0}, \text{ or } P_{\theta_0} \text{ a.s.}} 0.$$

Extra notes I

Take $\alpha = 1/2$ for simplicity and consistency in probability. Define $B(\theta, r)$ the closed ball of radius r centred around θ, and let

$$\hat{r}(\theta) = \inf\{r, \Pi(B(\theta, r)|X^n) \geq 1/2\}$$

(and inf over the empty set is ∞). Now let $\hat{\theta}_n$ be such that

$$\hat{r}(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} r(\theta) + 1/n$$

Consistency implies that $\Pi(B(\theta_0, \epsilon)|X^n) \to 1$ so $\hat{r}(\theta_0) \leq \epsilon$ with probability tending to 1. Furthermore, $\hat{r}(\hat{\theta}_n) \leq \hat{r}(\theta_0) + 1/n$ thus $\hat{r}(\hat{\theta}_n) \leq \epsilon + 1/n$ with probability tending to 1.

In addition, $B(\theta_0, \epsilon) \cap B(\hat{\theta}_n, \hat{r}(\hat{\theta}_n)) \neq \emptyset$ otherwise $\Pi(B(\theta_0, \epsilon) \cup B(\hat{\theta}_n, \hat{r}(\hat{\theta}_n))|X^n) = \Pi(B(\theta_0, \epsilon)|X^n) + \Pi(B(\hat{\theta}_n, \hat{r}(\hat{\theta}_n))|X^n) \to 1 + 1/2$. So we have

$$d(\theta_0, \hat{\theta}_n) \leq \hat{r}(\hat{\theta}_n) + \epsilon \leq 2\epsilon + 1/n$$

with probability that goes to 1.

Point estimator

- If Θ is a vector space, then one might want to use the posterior mean.
- But... weak convergence to a Dirac does not imply convergence of moments.
- Consistency of the posterior mean requires additional assumptions such as boundedness of posterior moments in probability or a.s. for some $p > 1$ would be sufficient.

Theorem (Posterior mean)

Assume that the balls of the metric space (Θ, d) are convex. Suppose that for any sequence $\theta_{1,n}, \theta_{2,n}$ in Θ and $\lambda_n \to 0$

$$d(\theta_{1,n}, (1-\lambda_n)\theta_{1,n} + \lambda_n \theta_{2,n}) \to 0$$

Then consistency of the posterior distribution implies consistency of the posterior mean.

Extra notes I

Let $\epsilon > 0$ and write $\hat{\theta}_n = \int \theta \Pi(d\theta|X^n)$. We decompose

$$\hat{\theta}_n = \int_{B(\theta_0,\epsilon)} \theta \Pi(d\theta|X^n) + \int_{B(\theta_0,\epsilon)^c} \theta \Pi(d\theta|X^n) = \theta_{1,n}(1-\lambda_n) + \lambda_n \theta_{2,n}$$

where $\theta_{1,n} = \int_{B(\theta_0,\epsilon)} \theta \frac{\Pi(d\theta|X^n))}{\Pi(B(\theta_0,\epsilon)|X^n)}$, $\lambda_n = \Pi(B(\theta_0,\epsilon)|X^n)$ and similarly for $\theta_{2,n}$ on the complement of $B(\theta_0,\epsilon)$. Using Jensen inequality we have

$$d(\theta_{n,1},\theta_0) \leq \int_{B(\theta_0,\epsilon)} d(\theta,\theta_0) \frac{\Pi(d\theta|X^n))}{\Pi(B(\theta_0,\epsilon)|X^n)} \leq \epsilon$$

In addition we have

$$d(\hat{\theta}_n,\theta_0) \leq d(\theta_{n,1},\theta_0) + d(\theta_{n,1},\theta_{1,n}(1-\lambda_n) + \lambda_n\theta_{2,n}).$$

Using the fact that $\lambda_n \to 0$ since the posterior is consistent, we have the desired result.

Remark

For the condition on d to hold, one can assume it to be convex and uniformly bounded.

A first consistent posterior

Example (Dirichlet process)

Assume the following model

$$X_1, \ldots, X_n \overset{iid}{\sim} P,$$
$$P \sim \mathrm{DP}(M\alpha)$$

Consider the semi-metric $d_A(P, Q) = |P(A) - Q(A)|$ for some measurable event A on Θ, then $\Pi(\cdot|X^n)$ is strongly consistent at any P_0 for d_A.

From this result, we can easily obtain consistency under the weak topology. We could also obtain stronger consistency using Glivenko–Cantelli theorem.

A first consistent posterior

Example (Dirichlet process)

Assume the following model

$$X_1, \ldots, X_n \overset{iid}{\sim} P,$$
$$P \sim \mathrm{DP}(M\alpha)$$

Consider the semi-metric $d_A(P, Q) = |P(A) - Q(A)|$ for some measurable event A on Θ, then $\Pi(\cdot|X^n)$ is strongly consistent at any P_0 for d_A.

From this result, we can easily obtain consistency under the weak topology. We could also obtain stronger consistency using Glivenko–Cantelli theorem.

Extra notes I

Consider $\Pi(|P(A) - P_0(A)| \geq \epsilon | X^n)$ which calls for applying Markov inequality. Properties of the Dirichlet process imply that

$$P|X^n \sim \mathrm{DP}(M\alpha + n\mathbb{P}_n),$$

thus

$$P(A)|X^n \sim \mathrm{Beta}(M\alpha(A) + n\mathbb{P}_n(A), M\alpha(A^c) + n\mathbb{P}_n(A^c)).$$

We thus have

$$\begin{aligned}
\mathrm{E}(P(A)|X^n) &= \frac{M}{M+n}\alpha(A) + \frac{n}{M+n}\mathbb{P}_n(A) := \bar{P}(A) \\
\mathrm{var}(P(A)|X^n) &= \frac{\bar{P}(A)\bar{P}(A^c)}{1+n+M} \leq \frac{1}{4(1+n+M)}.
\end{aligned}$$

Markov inequality gives

$$\begin{aligned}
\Pi(|P(A) - P_0(A)| \geq \epsilon | X^n) &\leq \frac{1}{\epsilon^2}\left(|\bar{P}(A) - P_0(A)|^2 + \mathrm{var}(P(A)|X^n)\right) \\
&\to 0 \ [P_0, a.s.]
\end{aligned}$$

using the law of large numbers on $\mathbb{P}(A)$.

Bayesian modelling perspective

From a Bayesian point of view, a Dirac measure at θ_0 corresponds to perfect knowledge of the parameter.

- Prior and posterior distributions model our knowledge about the parameter.
- Consistency thus implies that when the amount of information grows, we tend towards perfect knowledge of the parameter.

A validation of Bayesian methods

The frequentist setting where there exists a *true* parameter θ_0 that generates the data can be seen as an idealized set-up.

- An experimenter feeds a Bayesian with some data using the same data-generating mechanism.
- When the number of observation grows, a Bayesian should be able to pin-point the data-generating mechanism, whatever their prior.
- A prior that does not lead to a consistent posterior should not be used.

A validation of Bayesian methods

The frequentist setting where there exists a *true* parameter θ_0 that generates the data can be seen as an idealized set-up.

- An experimenter feeds a Bayesian with some data using the same data-generating mechanism.
- When the number of observation grows, a Bayesian should be able to pin-point the data-generating mechanism, whatever their prior.
- A prior that does not lead to a consistent posterior should not be used.

Robustness

Two Bayesians walk into a bar... with *almost* the same prior, then their posterior inference should not differ that much.

- Let Π_1 be the prior of Bayesian number 1
- Bayesian number 2 uses an "ϵ-corrupted" prior $\Pi_2 = (1-\epsilon)\Pi_1 + \epsilon\delta_{p_0}$ for some $p_0 \in \Theta$

The posterior of Bayesian number 2 is consistent at p_0 (to be seen later), now what if Π_1 is not consistent at p_0? Let d_W be the metric for the weak topology, then $d_W(\Pi_1(\cdot|X^n), \Pi_2(\cdot|X^n))$ would not go to 0.

Robustness

Two Bayesians walk into a bar... with *almost* the same prior, then their posterior inference should not differ that much.

- Let Π_1 be the prior of Bayesian number 1
- Bayesian number 2 uses an "ϵ-corrupted" prior $\Pi_2 = (1-\epsilon)\Pi_1 + \epsilon\delta_{p_0}$ for some $p_0 \in \Theta$

The posterior of Bayesian number 2 is consistent at p_0 (to be seen later), now what if Π_1 is not consistent at p_0? Let d_W be the metric for the weak topology, then $d_W(\Pi_1(\cdot|X^n), \Pi_2(\cdot|X^n))$ would not go to 0.

Robustness

Two Bayesians walk into a bar... with *almost* the same prior, then their posterior inference should not differ that much.

- Let Π_1 be the prior of Bayesian number 1
- Bayesian number 2 uses an "ϵ-corrupted" prior $\Pi_2 = (1-\epsilon)\Pi_1 + \epsilon\delta_{p_0}$ for some $p_0 \in \Theta$

The posterior of Bayesian number 2 is consistent at p_0 (to be seen later), now what if Π_1 is not consistent at p_0? Let d_W be the metric for the weak topology, then $d_W(\Pi_1(\cdot|X^n), \Pi_2(\cdot|X^n))$ would not go to 0.

Extra notes I

There exists some $\varepsilon_0 > 0$ such that

$$\Pi_{n,1}(B(\theta_0, \varepsilon_0)|X^n) \not\to 0$$

Thus

$$|\Pi_{n,1}(B(\theta_0, \varepsilon_0)|X^n) - \Pi_{n,2}(B(\theta_0, \varepsilon_0)|X^n)| \not\to 0$$

since $\Pi_{n,2}(B(\theta_0, \varepsilon_0)|X^n) \to 0$.

Doob's Theorem

Can one get general conditions on the prior to ensure that it is consistent?

- A first answer: Doob's Theorem
- The posterior is consistent at every θ Π-a.s.

Consider the case of *i.i.d.* observations

Theorem (Doob's Theorem)

Let $\{\mathcal{X}^n, P_\theta, \Theta\}$ be a statistical model where $\{\mathcal{X}^n, \mathcal{A}^n\}$ is a Polish space with Borel σ-field and Θ a Borel subset of a Polish space. Suppose that the map $\theta \mapsto P_\theta(A)$ is Borel measurable for evey $A \in \mathcal{A}$ and $\theta \mapsto P_\theta$ is one-to-one.

Then for any prior distribution Π on Θ, if $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$, $\theta \sim \Pi$, the posterior is strongly consistent at any θ Π-a.s.

Doob's Theorem

Can one get general conditions on the prior to ensure that it is consistent?

- A first answer: Doob's Theorem
- The posterior is consistent at every θ Π-a.s.

Consider the case of *i.i.d.* observations

Theorem (Doob's Theorem)

Let $\{\mathcal{X}^n, P_\theta, \Theta\}$ be a statistical model where $\{\mathcal{X}^n, \mathcal{A}^n\}$ is a Polish space with Borel σ-field and Θ a Borel subset of a Polish space. Suppose that the map $\theta \mapsto P_\theta(A)$ is Borel measurable for evey $A \in \mathcal{A}$ and $\theta \mapsto P_\theta$ is one-to-one. Then for any prior distribution Π on Θ, if $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$, $\theta \sim \Pi$, the posterior is strongly consistent at any θ Π-a.s.

Doob's Theorem

Can one get general conditions on the prior to ensure that it is consistent?

- A first answer: Doob's Theorem
- The posterior is consistent at every θ Π-a.s.

Consider the case of *i.i.d.* observations

Theorem (Doob's Theorem)

Let $\{\mathcal{X}^n, P_\theta, \Theta\}$ be a statistical model where $\{\mathcal{X}^n, \mathcal{A}^n\}$ is a Polish space with Borel σ-field and Θ a Borel subset of a Polish space. Suppose that the map $\theta \mapsto P_\theta(A)$ is Borel measurable for evey $A \in \mathcal{A}$ and $\theta \mapsto P_\theta$ is one-to-one.

Then for any prior distribution Π on Θ, if $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$, $\theta \sim \Pi$, the posterior is strongly consistent at any θ Π-a.s.

Doob's Theorem

Can one get general conditions on the prior to ensure that it is consistent?

- A first answer: Doob's Theorem
- The posterior is consistent at every θ Π-a.s.

Consider the case of *i.i.d.* observations

Theorem (Doob's Theorem)

Let $\{\mathcal{X}^n, P_\theta, \Theta\}$ be a statistical model where $\{\mathcal{X}^n, \mathcal{A}^n\}$ is a Polish space with Borel σ-field and Θ a Borel subset of a Polish space. Suppose that the map $\theta \mapsto P_\theta(A)$ is Borel measurable for evey $A \in \mathcal{A}$ and $\theta \mapsto P_\theta$ is one-to-one.

Then for any prior distribution Π on Θ, if $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$, $\theta \sim \Pi$, the posterior is strongly consistent at any θ Π-a.s.

Doob's Theorem

Can one get general conditions on the prior to ensure that it is consistent?

- A first answer: Doob's Theorem
- The posterior is consistent at every θ Π-a.s.

Consider the case of *i.i.d.* observations

Theorem (Doob's Theorem)

Let $\{\mathcal{X}^n, P_\theta, \Theta\}$ be a statistical model where $\{\mathcal{X}^n, \mathcal{A}^n\}$ is a Polish space with Borel σ-field and Θ a Borel subset of a Polish space. Suppose that the map $\theta \mapsto P_\theta(A)$ is Borel measurable for evey $A \in \mathcal{A}$ and $\theta \mapsto P_\theta$ is one-to-one.

Then for any prior distribution Π on Θ, if $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$, $\theta \sim \Pi$, the posterior is strongly consistent at any θ Π-a.s.

Some remarks on Doob's Theorem

- The conditions of the theorem are extremely weak
- And no conditions on the prior
- However this is only true Π-almost surely.
- Note: the Π-null set can be quite big! we can be happy with this result only if we are confident that the parameters are on the support of the prior. In general no one can be sure that the parameter generating the data inside the support of the prior, this is a real problem in fact in general the support of the prior can be quite thin.
  An extreme example is the case were the prior is a Dirac on some parameter θ_0. Then Doob's theorem still holds.

Some remarks on Doob's Theorem

- The conditions of the theorem are extremely weak
- And no conditions on the prior
- However this is only true Π-almost surely.
- Note: the Π-null set can be quite big! we can be happy with this result only if we are confident that the parameters are on the support of the prior. In general no one can be sure that the parameter generating the data inside the support of the prior, this is a real problem in fact in general the support of the prior can be quite thin.
  An extreme example is the case were the prior is a Dirac on some parameter θ_0. Then Doob's theorem still holds.

Some remarks on Doob's Theorem

- The conditions of the theorem are extremely weak
- And no conditions on the prior
- However this is only true Π-almost surely.
- Note: the Π-null set can be quite big! we can be happy with this result only if we are confident that the parameters are on the support of the prior. In general no one can be sure that the parameter generating the data inside the support of the prior, this is a real problem in fact in general the support of the prior can be quite thin.
  An extreme example is the case were the prior is a Dirac on some parameter θ_0. Then Doob's theorem still holds.

Some remarks on Doob's Theorem

- The conditions of the theorem are extremely weak
- And no conditions on the prior
- However this is only true Π-almost surely.
- Note: the Π-null set can be quite big! we can be happy with this result only if we are confident that the parameters are on the support of the prior. In general no one can be sure that the parameter generating the data inside the support of the prior, this is a real problem in fact in general the support of the prior can be quite thin.
  An extreme example is the case were the prior is a Dirac on some parameter θ_0. Then Doob's theorem still holds.

Setting

Doob's approach is not enough to show consistency of the posterior. For simplicity we focus on the density estimation setting.

- Θ is the set of probability density functions on $\mathcal{X}$ w.r.t. a common dominating measure ν. We denote the parameter p (instead of θ) and P the associated probability measure.
- Observations follow $X_1, \ldots, X_n \overset{iid}{\sim} p$, and $p \sim \Pi$.

Considering density estimation makes things easier without being too simplistic. The same results can be extended to nonparametric regression.

Setting

Doob's approach is not enough to show consistency of the posterior. For simplicity we focus on the density estimation setting.

- Θ is the set of probability density functions on $\mathcal{X}$ w.r.t. a common dominating measure ν. We denote the parameter p (instead of θ) and P the associated probability measure.
- Observations follow $X_1, \ldots, X_n \overset{iid}{\sim} p$, and $p \sim \Pi$.

Considering density estimation makes things easier without being too simplistic. The same results can be extended to nonparametric regression.

KL property

To achieve consistency, we do not want to require that the true parameter p_0 is inside the support of Π. However we still require some prior mass *near* p_0.

Definition (Kullback–Leibler)

Let p and p_0 be two p.d.f. with respect to a common measure such that $p_0 \ll p$. Then the Kullback–Leibler divergence between p and p_0 is

$$\mathrm{KL}(p, p_0) = \int p_0 \log(p_0/p) d\nu.$$

KL property

Definition (KL property)

We say that a prior distribution Π satisfies the Kullback–Leibler property at p_0 if for every $\epsilon > 0$,

$$\Pi(p : \mathrm{KL}(p, p_0) \geq \epsilon) > 0$$

We note $p_0 \in \mathrm{KL}(\Pi)$ and alternatively will say that p_0 is in the KL-support of Π.

This extends quite a lot the parameters at which the posterior can be consistent.

KL property

Definition (KL property)

We say that a prior distribution Π satisfies the Kullback–Leibler property at p_0 if for every $\epsilon > 0$,

$$\Pi(p : \mathrm{KL}(p, p_0) \geq \epsilon) > 0$$

We note $p_0 \in \mathrm{KL}(\Pi)$ and alternatively will say that p_0 is in the KL-support of Π.

This extends quite a lot the parameters at which the posterior can be consistent.

Existence of tests

The other requirement is that the parameter set is not too complex.

Definition (Exponentially consistent tests)

We say that a sequence of tests ϕ_n for $H_0 : p = p_0$ versus $H_1 : p \in U^c$ is exponentially consistent if

$$P_0^n(\phi_n) \lesssim e^{-Cn}, \quad \sup_{p \in U^c} P^n(1 - \phi_n) \lesssim e^{-Cn}$$

A test is understood as a measurable map $\mathcal{X}^n \to [0, 1]$ and the corresponding statistic $\phi_n(X_1, \ldots, X_n)$. ϕ_n is interpreted as the probability that the null is rejected.

Extra notes I

The existence of tests means that we can differentiate between p_0 and parameter in U^c.

It is enough to have uniformly consistent sequence of test

$$P_0(\phi_n) \to 0, \quad \sup_{p \in U^c} P(1 - \phi_n) \to 0.$$

Since the test is uniformly consistent then there exists $k \in \mathbb{N}$ such that $P_0^k(\phi_k) \leq 1/4$, $P^k(1 - \phi_k) \leq 1/4$. Now for n large, write $n = mk + r$. Slice $X^n = (X_1, \ldots, X_n)$ into m sub-sample of size k $X_l^n = (X_{(l-1)k+1}, \ldots, X_{lk})$ and define $Y_{l,n} = \phi_k(X_l^n)$. Now create a new test $\psi_n = \mathcal{I}\{\bar{Y}_m > 1/2\}$. We have for every $p \in U^c$, $P(1 - Y_j) \leq 1/4$

$$P(\psi_n) = P(\bar{Y} \leq 1/2) = P(1 - \bar{Y} \geq 1/2) = P(1 - \bar{(}Y) \geq 1/2) \leq e^{-2m/16} \lesssim e^{-Cn}$$

Using Hoeffding inequality: $\mathbb{P}(\bar{X} - \mathrm{E}(X) \geq \epsilon) \leq \exp\left\{-2\epsilon^2 m\right\}$.

Schwartz Theorem

Theorem

Let Π be a prior distribution on Θ such that $p_0 \in KL(\Pi)$. Let U be a neighbourhood of p_0 such that there exists an exponentially consistent sequence of tests for p_0 against U^c. Then

$$\Pi(U^c|X^n) \to 0 \ [P_0 a.s].$$

This theorem is not due to Herman Schwarz (without t!), nor to Laurent Schwartz the Fields Medalist! But to Lorraine Schwartz, former student of Lucien Le Cam.

Schwartz Theorem

Theorem

Let Π be a prior distribution on Θ such that $p_0 \in KL(\Pi)$. Let U be a neighbourhood of p_0 such that there exists an exponentially consistent sequence of tests for p_0 against U^c. Then

$$\Pi(U^c|X^n) \to 0 \ [P_0 a.s].$$

This theorem is not due to Herman Schwarz (without t!), nor to Laurent Schwartz the Fields Medalist! But to Lorraine Schwartz, former student of Lucien Le Cam.

Extra notes I

$$\Pi(U^c|X^n) = \frac{\int_{U^c} \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p)}{\int_{\mathcal{P}} \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p)} := \frac{N_n}{D_n}.$$

We first show $\liminf D_n e^{n\epsilon}/\Pi(KL(p, p_0) > \epsilon) \geq 1$, $P_0[a.s.]$. Let $\Pi_0(\cdot) = \Pi(\cdot \cap \mathrm{KL}(p, p_0) > \epsilon)/\Pi(KL(p, p_0) > \epsilon)$. Then

$$\begin{aligned}
\log(D_n) &\geq \log\left(\int_{KL(p,p_0)>\epsilon} \frac{p}{p_0}(X_i) d\Pi_0(p)\right) + \log(\Pi(KL(p, p_0) < \epsilon)) \\
&\geq \int_{KL(p,p_0)>\epsilon} \log\left(\prod_{i=1}^n \frac{p}{p_0}(X_i)\right) d\Pi_0(p) + \log(\Pi(KL(p, p_0) < \epsilon)) \\
&= \sum_{i=1}^n \int \log \frac{p}{p_0}(X_i) d\Pi_0(p) + \log(\Pi(KL(p, p_0) < \epsilon))
\end{aligned}$$

The law of large numbers implies

$$\frac{1}{n}\sum_{i=1}^n \int \log \frac{p}{p_0}(X_i) d\Pi_0(p) \to P_0 \int \frac{p}{p_0}(X_i) d\Pi_0(p), \ P_0[a.s.]$$

Extra notes II

which is $-\int \mathrm{KL}(p, p_0) d\Pi_0(p) > -\epsilon$. Thus

$$\liminf D_n e^{n\epsilon} / \Pi(KL(p, p_0) > \epsilon) \geq 1, \ P_0[a.s.]$$

For n large enough we have the following $P_0[a.s.]$

$$\begin{aligned}\Pi(U^c|X^n) &\leq \phi_n + (1-\phi_n)\frac{N_n}{D_n} \\ &\leq \phi_n + (1-\phi_n) N_n e^{\epsilon n} \Pi(KL(p, p_0) > \epsilon)\end{aligned}$$

Furthermore we have that

$$\begin{aligned}P_0^n N_n (1-\phi_n) &= P_0^n \int_{U^c} (1-\phi_n) \prod_{i=1}^{n} \frac{p}{p_0}(X_i) \Pi(dp) \\ &= \int_{U^c} P^n (1-\phi_n) \Pi(dp) \leq e^{-Cn}\end{aligned}$$

We thus get $P_0 \Pi(U^c|X^n) \leq e^{-C'n}$ for $\epsilon < C$ and for $C' = C - \epsilon$. Using Borel–Cantelli we get that $\Pi(U^c|X^n) \to 0 P_0[a.s.]$.

Schwartz Theorem

- Need to test away all densities in U^c
- Might not be possible for strong neighbourhood of p_0 (L_1 metrics)

Extension of Schwartz theorem

The idea is that not *all* functions in U^c matters and we can discard function with very low prior probabilities.

Theorem

The results of the previous theorem are still valid if we replace the assumption on the existence of tests by:

$$\Theta_n \subset \Theta$$

$$\Pi(\Theta_n) \leq e^{-Cn}, \; P_0^n \phi_n \leq e^{-Cn}, \; \sup_{p \in U^c \cap \Theta_n} P(1 - \phi_n) \leq e^{-Cn}$$

- Need to test away all densities in U^c
- Might not be possible for strong neighbourhood of p_0 (L_1 metrics)

Extension of Schwartz theorem

The idea is that not *all* functions in U^c matters and we can discard function with very low prior probabilities.

Theorem

The results of the previous theorem are still valid if we replace the assumption on the existence of tests by:

$$\Theta_n \subset \Theta$$

$$\Pi(\Theta_n) \leq e^{-Cn}, \; P_0^n \phi_n \leq e^{-Cn}, \; \sup_{p \in U^c \cap \Theta_n} P(1 - \phi_n) \leq e^{-Cn}$$

- Need to test away all densities in U^c
- Might not be possible for strong neighbourhood of p_0 (L_1 metrics)

Extension of Schwartz theorem

The idea is that not *all* functions in U^c matters and we can discard function with very low prior probabilities.

Theorem

The results of the previous theorem are still valid if we replace the assumption on the existence of tests by:

$$\Theta_n \subset \Theta$$

$$\Pi(\Theta_n) \leq e^{-Cn}, \; P_0^n \phi_n \leq e^{-Cn}, \; \sup_{p \in U^c \cap \Theta_n} P(1 - \phi_n) \leq e^{-Cn}$$

Existence of tests

Schwartz' theorem requires the existence of exponentially consistent tests.

- We can differentiate between θ_0 and U^c
- The model is not too complex

Question

When do such tests exist?

Let's see the example of iid observations.

Existence of tests

Schwartz' theorem requires the existence of exponentially consistent tests.

- We can differentiate between θ_0 and U^c
- The model is not too complex

Question

When do such tests exist?

Let's see the example of iid observations.

Existence of tests

Schwartz' theorem requires the existence of exponentially consistent tests.

- We can differentiate between θ_0 and U^c
- The model is not too complex

Question

When do such tests exist?

Let's see the example of iid observations.

Existence of tests

Schwartz' theorem requires the existence of exponentially consistent tests.

- We can differentiate between θ_0 and U^c
- The model is not too complex

Question

When do such tests exist?

Let's see the example of iid observations.

Sketch of the proof

- Cannot directly construct test against $U^c = \{p, d(p, p_0) > \epsilon\}$...
- Construct an exponentially consistent test against a generic ball that is at least at distance ϵ
- Cover U^c with N of these balls, and construct a test from the N corresponding tests.

Sketch of the proof

- Cannot directly construct test against $U^c = \{p, d(p, p_0) > \epsilon\}$...
- Construct an exponentially consistent test against a generic ball that is at least at distance ϵ
- Cover U^c with N of these balls, and construct a test from the N corresponding tests.

Sketch of the proof

- Cannot directly construct test against $U^c = \{p, d(p, p_0) > \epsilon\}$...
- Construct an exponentially consistent test against a generic ball that is at least at distance ϵ
- Cover U^c with N of these balls, and construct a test from the N corresponding tests.

Consistency under Entropy bound

We combine the preceding results to get general conditions on the prior and on the model, that ensure consistency.

Theorem

The posterior is strongly consistent relative to the L_1 distance at every p_0 in the KL-support of the prior if for every $\epsilon > 0$ there exist Θ_n such that for $C > 0$ and $0 < c < 1/2$

$$\Pi(\Theta_n^c) \leq e^{-Cn}, \ \log N(\epsilon, \Theta_n, \|\cdot\|_1) \leq cn\epsilon_n^2,$$

for n large enough.

Consistency under Entropy bound

We combine the preceding results to get general conditions on the prior and on the model, that ensure consistency.

Theorem

The posterior is strongly consistent relative to the L_1 distance at every p_0 in the KL-support of the prior if for every $\epsilon > 0$ there exist Θ_n such that for $C > 0$ and $0 < c < 1/2$

$$\Pi(\Theta_n^c) \leq e^{-Cn}, \ \log N(\epsilon, \Theta_n, \|\cdot\|_1) \leq cn\epsilon_n^2,$$

for n large enough.

Consistency under Entropy bound

We combine the preceding results to get general conditions on the prior and on the model, that ensure consistency.

Theorem

The posterior is strongly consistent relative to the L_1 distance at every p_0 in the KL-support of the prior if for every $\epsilon > 0$ there exist Θ_n such that for $C > 0$ and $0 < c < 1/2$

$$\Pi(\Theta_n^c) \leq e^{-Cn}, \ \log N(\epsilon, \Theta_n, \|\cdot\|_1) \leq cn\epsilon_n^2,$$

for n large enough.

Outline

Definition

Contraction rates are a refinement of posterior consistency.

- How fast posterior concentrates its mass around the true parameter
- Helps to see how much the prior influences the posterior

Definition

Let ϵ_n be a positive sequence. The posterior contracts at the rate ϵ_n at θ_0 if for any $M_n \to \infty$

$$\Pi(d(\theta, \theta_0) > M_n \epsilon_n | X^n) \xrightarrow[n \to \infty]{P_{\theta_0}} 0$$

If all the experiments share the same probability space and the convergence is $P_{\theta_0}[a.s]$ we say that the posterior contracts in the strong sense.

Definition

Contraction rates are a refinement of posterior consistency.

- How fast posterior concentrates its mass around the true parameter
- Helps to see how much the prior influences the posterior

Definition

Let ϵ_n be a positive sequence. The posterior contracts at the rate ϵ_n at θ_0 if for any $M_n \to \infty$

$$\Pi(d(\theta, \theta_0) > M_n \epsilon_n | X^n) \xrightarrow[n \to \infty]{P_{\theta_0}} 0$$

If all the experiments share the same probability space and the convergence is $P_{\theta_0}[a.s]$ we say that the posterior contracts in the strong sense.

Definition

Contraction rates are a refinement of posterior consistency.

- How fast posterior concentrates its mass around the true parameter
- Helps to see how much the prior influences the posterior

Definition

Let ϵ_n be a positive sequence. The posterior contracts at the rate ϵ_n at θ_0 if for any $M_n \to \infty$

$$\Pi(d(\theta, \theta_0) > M_n \epsilon_n | X^n) \xrightarrow[n \to \infty]{P_{\theta_0}} 0$$

If all the experiments share the same probability space and the convergence is $P_{\theta_0}[a.s]$ we say that the posterior contracts in the strong sense.

Definition

Contraction rates are a refinement of posterior consistency.

- How fast posterior concentrates its mass around the true parameter
- Helps to see how much the prior influences the posterior

Definition

Let ϵ_n be a positive sequence. The posterior contracts at the rate ϵ_n at θ_0 if for any $M_n \to \infty$

$$\Pi(d(\theta, \theta_0) > M_n \epsilon_n | X^n) \xrightarrow[n \to \infty]{P_{\theta_0}} 0$$

If all the experiments share the same probability space and the convergence is $P_{\theta_0}[a.s]$ we say that the posterior contracts in the strong sense.

Definition

Contraction rates are a refinement of posterior consistency.

- How fast posterior concentrates its mass around the true parameter
- Helps to see how much the prior influences the posterior

Definition

Let ϵ_n be a positive sequence. The posterior contracts at the rate ϵ_n at θ_0 if for any $M_n \to \infty$

$$\Pi(d(\theta, \theta_0) > M_n \epsilon_n | X^n) \xrightarrow[n \to \infty]{P_{\theta_0}} 0$$

If all the experiments share the same probability space and the convergence is $P_{\theta_0}[a.s]$ we say that the posterior contracts in the strong sense.

Remarks

- Any slower rate than ϵ_n also fits the definition so we will say *a* posterior contraction rate
- We will naturally try to find the fastest possible rate!

Regarding M_n

- The sequence M_n plays virtually no role in the posterior rate. In many cases it can be fixed to a constant M.
- For finite dimensional models M_n must be allowed to grow to obtain the usual $n^{-1/2}$ rate in smooth models.

Remarks

- Any slower rate than ϵ_n also fits the definition so we will say *a* posterior contraction rate
- We will naturally try to find the fastest possible rate!

Regarding M_n

- The sequence M_n plays virtually no role in the posterior rate. In many cases it can be fixed to a constant M.
- For finite dimensional models M_n must be allowed to grow to obtain the usual $n^{-1/2}$ rate in smooth models.

Remarks

- Any slower rate than ϵ_n also fits the definition so we will say *a* posterior contraction rate
- We will naturally try to find the fastest possible rate!

Regarding M_n

- The sequence M_n plays virtually no role in the posterior rate. In many cases it can be fixed to a constant M.
- For finite dimensional models M_n must be allowed to grow to obtain the usual $n^{-1/2}$ rate in smooth models.

Remarks

- Any slower rate than ϵ_n also fits the definition so we will say *a* posterior contraction rate
- We will naturally try to find the fastest possible rate!

Regarding M_n

- The sequence M_n plays virtually no role in the posterior rate. In many cases it can be fixed to a constant M.
- For finite dimensional models M_n must be allowed to grow to obtain the usual $n^{-1/2}$ rate in smooth models.

Remarks

- Any slower rate than ϵ_n also fits the definition so we will say *a* posterior contraction rate
- We will naturally try to find the fastest possible rate!

Regarding M_n

- The sequence M_n plays virtually no role in the posterior rate. In many cases it can be fixed to a constant M.
- For finite dimensional models M_n must be allowed to grow to obtain the usual $n^{-1/2}$ rate in smooth models.

Consequences of posterior contraction

Point Estimator

- Let $\hat{\theta}_n$ = centre of the smallest ball that contains at least $1/2$ of the posterior mass.
- Assume that the posterior contracts at θ_0 with rate ϵ_n for the metric d

Then $d(\hat{\theta}_n, \theta) = O_P(\epsilon_n)$ in P_0 probability (or a.s. if strong contraction).

Consequences of posterior contraction

Point Estimator

- Let $\hat{\theta}_n =$ centre of the smallest ball that contains at least $1/2$ of the posterior mass.
- Assume that the posterior contracts at θ_0 with rate ϵ_n for the metric d

Then $d(\hat{\theta}_n, \theta) = O_P(\epsilon_n)$ in P_0 probability (or a.s. if strong contraction).

Consequences of posterior contraction

Point Estimator

- Let $\hat{\theta}_n =$ centre of the smallest ball that contains at least $1/2$ of the posterior mass.
- Assume that the posterior contracts at θ_0 with rate ϵ_n for the metric d

Then $d(\hat{\theta}_n, \theta) = O_P(\epsilon_n)$ in P_0 probability (or a.s. if strong contraction).

Consequences of posterior contraction

Point Estimator

- Let $\hat{\theta}_n =$ centre of the smallest ball that contains at least $1/2$ of the posterior mass.
- Assume that the posterior contracts at θ_0 with rate ϵ_n for the metric d

Then $d(\hat{\theta}_n, \theta) = O_P(\epsilon_n)$ in P_0 probability (or a.s. if strong contraction).

Consequences of posterior contraction

Posterior mean

If the metric d is bounded and $\theta \mapsto d^s(\theta, \theta_0)$ is convex for some $s \geq 1$ then the posterior mean $\tilde{\theta}_n$ satisfies

$$d(\tilde{\theta}_n, \theta_0) \leq M_n \epsilon_n + \|d\|_\infty^{1/s} \Pi_n(d(\theta, \theta_0) \geq M_n \epsilon_n | X^n)^{1/s}.$$

- First term is the dominating term
- The second term is exponentially small in general

Consequences of posterior contraction

Posterior mean

If the metric d is bounded and $\theta \mapsto d^s(\theta, \theta_0)$ is convex for some $s \geq 1$ then the posterior mean $\tilde{\theta}_n$ satisfies

$$d(\tilde{\theta}_n, \theta_0) \leq M_n \epsilon_n + \|d\|_\infty^{1/s} \Pi_n(d(\theta, \theta_0) \geq M_n \epsilon_n | X^n)^{1/s}.$$

- First term is the dominating term
- The second term is exponentially small in general

Some first Examples - Parametric models

- Let $X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathcal{B}(\theta)$, and $\theta \sim \text{Beta}(\alpha, \beta)$. The posterior contracts at a rate $n^{-1/2}$
- Let $X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathcal{U}([0, \theta])$ and $\pi(\theta) \propto \theta^{-a}$. The posterior contracts at a rate n^{-1}.

Parametric regular models

In fact for all regular finite dimensional models the Bernstein von-Mises theorem implies a posterior rate of $n^{-1/2}$.

Some first Examples - Parametric models

- Let $X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathcal{B}(\theta)$, and $\theta \sim \text{Beta}(\alpha, \beta)$. The posterior contracts at a rate $n^{-1/2}$
- Let $X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathcal{U}([0, \theta])$ and $\pi(\theta) \propto \theta^{-a}$. The posterior contracts at a rate n^{-1}.

Parametric regular models

In fact for all regular finite dimensional models the Bernstein von-Mises theorem implies a posterior rate of $n^{-1/2}$.

Some first Examples - Parametric models

- Let $X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathcal{B}(\theta)$, and $\theta \sim \text{Beta}(\alpha, \beta)$. The posterior contracts at a rate $n^{-1/2}$
- Let $X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathcal{U}([0, \theta])$ and $\pi(\theta) \propto \theta^{-a}$. The posterior contracts at a rate n^{-1}.

Parametric regular models

In fact for all regular finite dimensional models the Bernstein von-Mises theorem implies a posterior rate of $n^{-1/2}$.

- Let $X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathcal{B}(\theta)$, and $\theta \sim \text{Beta}(\alpha, \beta)$. The posterior contracts at a rate $n^{-1/2}$
- Let $X_1, \ldots, X_n | \theta \overset{iid}{\sim} \mathcal{U}([0, \theta])$ and $\pi(\theta) \propto \theta^{-a}$. The posterior contracts at a rate n^{-1}.

Parametric regular models

In fact for all regular finite dimensional models the Bernstein von-Mises theorem implies a posterior rate of $n^{-1/2}$.

Nonparametric example: Dirichlet Process

- $X_1, \ldots, X_n | P \overset{iid}{\sim} P$
- $P \sim \text{DP}(M\alpha)$ for α a probability measure on $\mathcal{X}$.

The posterior distribution is $P|X^n \sim \text{DP}(M\alpha + n\mathbb{P}_n)$.

Local semi-metric[1]

For a measurable set A, let $d(P, Q) = |P(A) - Q(A)|$. The posterior distribution is consistent at P_0 at a rate $n^{-1/2}$.

Global metric

For ν a σ-finite measure and F and G two c.d.f. let $d(F, G) = \|F - G\|_\nu^2 = \int (F(t) - G(t))^2 d\nu(t)$. The posterior contracts at rate $n^{-1/2}$ at P_0 for this metric.

Nonparametric example: Dirichlet Process

- $X_1, \ldots, X_n | P \overset{iid}{\sim} P$
- $P \sim \text{DP}(M\alpha)$ for α a probability measure on $\mathcal{X}$.

The posterior distribution is $P|X^n \sim \text{DP}(M\alpha + n\mathbb{P}_n)$.

Local semi-metric[1]

For a measurable set A, let $d(P, Q) = |P(A) - Q(A)|$. The posterior distribution is consistent at P_0 at a rate $n^{-1/2}$.

Global metric

For ν a σ-finite measure and F and G two c.d.f. let $d(F, G) = \|F - G\|_\nu^2 = \int (F(t) - G(t))^2 d\nu(t)$. The posterior contracts at rate $n^{-1/2}$ at P_0 for this metric.

Nonparametric example: Dirichlet Process

- $X_1, \ldots, X_n | P \overset{iid}{\sim} P$
- $P \sim \text{DP}(M\alpha)$ for α a probability measure on $\mathcal{X}$.

The posterior distribution is $P|X^n \sim \text{DP}(M\alpha + n\mathbb{P}_n)$.

Local semi-metric[1]

For a measurable set A, let $d(P, Q) = |P(A) - Q(A)|$. The posterior distribution is consistent at P_0 at a rate $n^{-1/2}$.

Global metric

For ν a σ-finite measure and F and G two c.d.f. let $d(F, G) = \|F - G\|_\nu^2 = \int (F(t) - G(t))^2 d\nu(t)$. The posterior contracts at rate $n^{-1/2}$ at P_0 for this metric.

Nonparametric example: White Noise

Consider the following model for W_t a white noise

$$X_t = f(t) + n^{-1/2} W_t.$$

Projecting this model onto the Fourier basis if $f \in L_2$, we have the equivalent formulation

$$X_{i,n} = \theta_i + n^{-1/2} \epsilon_i, \ \ i \in \mathbb{N}^*$$

$\theta \in \ell_2(\mathbb{L})$. Assume the following prior

$$\theta_i \overset{ind.}{\sim} \mathcal{N}(0, i^{-2\alpha-1}).$$

If $\theta_0 \in \mathcal{S}_\beta^{2,2}$ then the posterior contracts at θ_0 at the rate $n^{-\min(\alpha,\beta)/(2\alpha+1)}$.

General theorem

- Result similar to Schwartz theorem?
- We focus on the case of i.i.d observations $X_1, \ldots, X_n \overset{iid}{\sim} P$
- The parameter set Θ is the set of probability densities with respect to a common dominating measure μ.

Let Π_n be a sequence of priors. We study the sequence of posterior distributions $\Pi_n(\cdot|X^n)$ under the assumption that the data are generated from P.

General theorem

- Result similar to Schwartz theorem?
- We focus on the case of i.i.d observations $X_1, \ldots, X_n \overset{iid}{\sim} P$
- The parameter set Θ is the set of probability densities with respect to a common dominating measure μ.

Let Π_n be a sequence of priors. We study the sequence of posterior distributions $\Pi_n(\cdot|X^n)$ under the assumption that the data are generated from P.

General theorem

- Result similar to Schwartz theorem?
- We focus on the case of i.i.d observations $X_1, \ldots, X_n \overset{iid}{\sim} P$
- The parameter set Θ is the set of probability densities with respect to a common dominating measure μ.

Let Π_n be a sequence of priors. We study the sequence of posterior distributions $\Pi_n(\cdot|X^n)$ under the assumption that the data are generated from P.

General theorem

- Result similar to Schwartz theorem?
- We focus on the case of i.i.d observations $X_1, \ldots, X_n \overset{iid}{\sim} P$
- The parameter set Θ is the set of probability densities with respect to a common dominating measure μ.

Let Π_n be a sequence of priors. We study the sequence of posterior distributions $\Pi_n(\cdot|X^n)$ under the assumption that the data are generated from P.

General theorem

- Result similar to Schwartz theorem?
- We focus on the case of i.i.d observations $X_1, \ldots, X_n \overset{iid}{\sim} P$
- The parameter set Θ is the set of probability densities with respect to a common dominating measure μ.

Let Π_n be a sequence of priors. We study the sequence of posterior distributions $\Pi_n(\cdot|X^n)$ under the assumption that the data are generated from P.

General Theorem

We follow the same steps as for Schwartz' Theorem:

- Existence of tests to separate p_0 from the complement of balls
- KL condition: the prior puts enough mass on neighbourhood of p_0

Define $V_{2,0}$, the 2nd KL variation

$$V_2 = P_0 \left(\log^2 \left(\frac{p_0}{p}(X) \right) \right),$$

and define two KL neighbourhoods as

$$B_0(p_0, \epsilon) = \{p, \mathrm{KL}(p_0, p) \leq \epsilon^2\},$$
$$B_2(p_0, \epsilon) = \{p, \mathrm{KL}(p_0, p) \leq \epsilon^2, V_2(p_0, p) \leq \epsilon^2\}.$$

General theorem

Theorem (Ghosal, Ghosh and van der Vaart)

Let $d \leq h$ be a metric on Θ for which balls are convex, and let $\Theta_n \subset \Theta$. The posterior contracts at a rate ϵ_n for all ϵ_n such that $n\epsilon_n^2 \to \infty$ and such that for positive constants c_1, c_2 and any $\underline{\epsilon}_n \leq \epsilon_n$

$$\log N(\epsilon_n, \Theta_n, d) \leq c_1 n \epsilon_n^2,$$

$$\Pi_n(B_{2,0}(p_0, \underline{\epsilon}_n^2)) \geq e^{-c_2 n \underline{\epsilon}_n^2}$$

$$\Pi(\Theta_n^c) \leq e^{-(c_2+3) n \underline{\epsilon}_n^2}$$

General Theorem

- The KL condition can be refined, but the idea is basically the same
- Entropy condition is useful for the existence of tests
- Entropy condition can be replaced by a local entropy, which is more like a *dimension of* Θ_n

Interpretation

Assume that d and KL are equivalent

- We need $e^{n\varepsilon_n^2}$ balls to cover Θ_n
- If the prior spread evenly the mass on these balls, we have $\pi(B) \geq e^{-n\varepsilon_n^2}$ mass on each of these balls thus KL condition is satisfied
- If the spread is uneven, then KL condition might not be satisfied for some p_0

General Theorem

- The KL condition can be refined, but the idea is basically the same
- Entropy condition is useful for the existence of tests
- Entropy condition can be replaced by a local entropy, which is more like a *dimension of* Θ_n

Interpretation

Assume that d and KL are equivalent

- We need e^{nr} balls to cover Θ_n
- If the prior spread evenly the mass on these balls, we have e^{nr} mass on each of these balls thus KL condition is satisfied
- If the spread is uneven, then KL condition might not be satisfied for some μ's

General Theorem

- The KL condition can be refined, but the idea is basically the same
- Entropy condition is useful for the existence of tests
- Entropy condition can be replaced by a local entropy, which is more like a *dimension of* Θ_n

Interpretation

Assume that d and KL are equivalent

- We need $e^{n\varepsilon_n^2}$ balls to cover Θ_n.
- If the prior spread evenly the mass on these balls, we have enough mass on each of these balls thus KL condition is satisfied
- If the spread is uneven, then KL condition might not be satisfied for some p_θ.

General Theorem

- The KL condition can be refined, but the idea is basically the same
- Entropy condition is useful for the existence of tests
- Entropy condition can be replaced by a local entropy, which is more like a *dimension of* Θ_n

Interpretation

Assume that d and KL are equivalent

- We need $e^{n\epsilon_n^2}$ balls to cover Θ_n.
- If the prior spread evenly the mass on these balls, we have $e^{-Cn\epsilon_n^2}$ mass on each of these balls thus KL condition is satisfied
- If the spread is uneven, then KL condition might not be satisfied for some p_0.

General Theorem

- The KL condition can be refined, but the idea is basically the same
- Entropy condition is useful for the existence of tests
- Entropy condition can be replaced by a local entropy, which is more like a *dimension of* Θ_n

Interpretation

Assume that d and KL are equivalent

- We need $e^{n\epsilon_n^2}$ balls to cover Θ_n.
- If the prior spread evenly the mass on these balls, we have $e^{-Cn\epsilon_n^2}$ mass on each of these balls thus KL condition is satisfied
- If the spread is uneven, then KL condition might not be satisfied for some p_0.

General Theorem

- The KL condition can be refined, but the idea is basically the same
- Entropy condition is useful for the existence of tests
- Entropy condition can be replaced by a local entropy, which is more like a *dimension of* Θ_n

Interpretation

Assume that d and KL are equivalent

- We need $e^{n\epsilon_n^2}$ balls to cover Θ_n.
- If the prior spread evenly the mass on these balls, we have $e^{-Cn\epsilon_n^2}$ mass on each of these balls thus KL condition is satisfied
- If the spread is uneven, then KL condition might not be satisfied for some p_0.

General Theorem

- The KL condition can be refined, but the idea is basically the same
- Entropy condition is useful for the existence of tests
- Entropy condition can be replaced by a local entropy, which is more like a *dimension of* Θ_n

Interpretation

Assume that d and KL are equivalent

- We need $e^{n\epsilon_n^2}$ balls to cover Θ_n.
- If the prior spread evenly the mass on these balls, we have $e^{-Cn\epsilon_n^2}$ mass on each of these balls thus KL condition is satisfied
- If the spread is uneven, then KL condition might not be satisfied for some p_0.

General observations

- The previous theorem can be generalized to other models (like regression for instance)
- But we have to be careful with the metric we use, and the existence of test is not guaranteed!
- To be general we will have to assume that we can test away parameters

Existence of tests

Let d_n and e_n be two semi-metrics on Θ. For $\epsilon > 0$, and for all $\theta_1 \in \Theta$ such that $d_n(\theta_0, \theta_1) > \epsilon$ there exists ϕ_n

$$P^n_{\Theta_0} \phi_n \leq e^{-Kn\epsilon^2}, \quad \sup_{\theta, e_n(\theta,\theta_1) \leq \xi\epsilon} P^n_\theta (1 - \phi_n) \leq e^{-Kn\epsilon^2}$$

General observations

- The previous theorem can be generalized to other models (like regression for instance)
- But we have to be careful with the metric we use, and the existence of test is not guaranteed!
- To be general we will have to assume that we can test away parameters

Existence of tests

Let d_n and e_n be two semi-metrics on Θ. For $\epsilon > 0$, and for all $\theta_1 \in \Theta$ such that $d_n(\theta_0, \theta_1) > \epsilon$ there exists ϕ_n

$$P^n_{\Theta_0} \phi_n \leq e^{-Kn\epsilon^2}, \quad \sup_{\theta, e_n(\theta,\theta_1) \leq \xi\epsilon} P^n_\theta (1 - \phi_n) \leq e^{-Kn\epsilon^2}$$

General observations

- The previous theorem can be generalized to other models (like regression for instance)
- But we have to be careful with the metric we use, and the existence of test is not guaranteed!
- To be general we will have to assume that we can test away parameters

Existence of tests

Let d_n and e_n be two semi-metrics on Θ. For $\epsilon > 0$, and for all $\theta_1 \in \Theta$ such that $d_n(\theta_0, \theta_1) > \epsilon$ there exists ϕ_n

$$P^n_{\Theta_0}\phi_n \leq e^{-Kn\epsilon^2}, \quad \sup_{\theta, e_n(\theta,\theta_1)\leq \xi\epsilon} P^n_\theta(1-\phi_n) \leq e^{-Kn\epsilon^2}$$

General observations

- The previous theorem can be generalized to other models (like regression for instance)
- But we have to be careful with the metric we use, and the existence of test is not guaranteed!
- To be general we will have to assume that we can test away parameters

Existence of tests

Let d_n and e_n be two semi-metrics on Θ. For $\epsilon > 0$, and for all $\theta_1 \in \Theta$ such that $d_n(\theta_0, \theta_1) > \epsilon$ there exists ϕ_n

$$P^n_{\Theta_0}\phi_n \leq e^{-Kn\epsilon^2}, \quad \sup_{\theta, e_n(\theta,\theta_1)\leq \xi\epsilon} P^n_\theta(1-\phi_n) \leq e^{-Kn\epsilon^2}$$

General observations

- The previous theorem can be generalized to other models (like regression for instance)
- But we have to be careful with the metric we use, and the existence of test is not guaranteed!
- To be general we will have to assume that we can test away parameters

Existence of tests

Let d_n and e_n be two semi-metrics on Θ. For $\epsilon > 0$, and for all $\theta_1 \in \Theta$ such that $d_n(\theta_0, \theta_1) > \epsilon$ there exists ϕ_n

$$P^n_{\Theta_0} \phi_n \leq e^{-Kn\epsilon^2}, \quad \sup_{\theta, e_n(\theta,\theta_1) \leq \xi\epsilon} P^n_\theta (1 - \phi_n) \leq e^{-Kn\epsilon^2}$$

General observations

- The previous theorem can be generalized to other models (like regression for instance)
- But we have to be careful with the metric we use, and the existence of test is not guaranteed!
- To be general we will have to assume that we can test away parameters

Existence of tests

Let d_n and e_n be two semi-metrics on Θ. For $\epsilon > 0$, and for all $\theta_1 \in \Theta$ such that $d_n(\theta_0, \theta_1) > \epsilon$ there exists ϕ_n

$$P^n_{\Theta_0} \phi_n \leq e^{-Kn\epsilon^2}, \quad \sup_{\theta, e_n(\theta,\theta_1) \leq \xi\epsilon} P^n_\theta (1 - \phi_n) \leq e^{-Kn\epsilon^2}$$

General theorem

Define the following KL-neighbourhood

$$V_{k,0}(f,g) = \int f |\log(f/g) - \mathrm{KL}(f,g)|^k d\mu$$

$$B_n(\theta_0, \epsilon, k) = \left\{ \theta \in \Theta \mathrm{KL}(p^n_{\theta_0}, p^n_\theta) \leq n\epsilon^2, V_{k,0}(p^n_{\theta_0}, p^n_\theta) \leq n^{k/2}\epsilon^k \right\}$$

General theorem

Theorem

Let d_n and e_n be two semi-metrics on Θ, such that tests exists, $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$, $k > 1$, $\Theta_n \subset \Theta$ such that for sufficiently large $j \in \mathbb{N}$

$$\sup_{\epsilon \geq \epsilon_n} \log N\left(\frac{1}{2}\xi\epsilon, \{\theta \in \Theta_n d_n(\theta_0, \theta) \leq \epsilon\}, e_n\right) \leq n\epsilon_n^2$$

$$\frac{\Pi_n(\theta \in \Theta_n, j\epsilon_n \leq d_n(\theta, \theta_0) \leq 2j\epsilon_n)}{\Pi_n(B_n(\theta_0, \epsilon_n, k))} \leq e^{Kn\epsilon_n^2 j^2/2}$$

$$\frac{\Pi_n(\Theta_n^c)}{\Pi_n(B_n(\theta_0, \epsilon_n, k))} \leq e^{-2n\epsilon_n}$$

then $P_{\theta_0}^n \Pi_n(d_n(\theta_0, \theta) \geq M_n\epsilon_n) = o(1)$

Independent observations

- Assume that the measure $P_\theta^n = \bigotimes_{i=1}^n P_{i,\theta}$ on some product space $\bigotimes_{i=1}^n \{\mathcal{X}_i, \mathcal{A}_i\}$.
- Assume that each measures $P_{i,\theta}$ are absolutely continuous w.r.t μ_i
- Define the Root average Hellinger distance

$$d_{n,H}(\theta, \theta') = \left(\frac{1}{n} \sum_{i=1}^n \int (\sqrt{dP_{i,\theta}} - \sqrt{dP_{i,\theta'}})^2 \right)^{1/2}$$

Lemma

For all here exists tests ϕ_n such that

$$P_{\theta_0}^n \phi_n \leq e^{-n d_{n,H}(\theta_0, \theta_1)}, \; P_\theta^n \leq e^{-n d_{n,H}(\theta_0, \theta_1)}$$

for all θ such that $d_{n,H}(\theta, \theta_1) \leq \frac{1}{18} d_{n,H}(\theta_0, \theta_1)$

Independent observations

- Assume that the measure $P_\theta^n = \bigotimes_{i=1}^n P_{i,\theta}$ on some product space $\bigotimes_{i=1}^n \{\mathcal{X}_i, \mathcal{A}_i\}$.
- Assume that each measures $P_{i,\theta}$ are absolutely continuous w.r.t μ_i
- Define the Root average Hellinger distance

$$d_{n,H}(\theta, \theta') = \left(\frac{1}{n} \sum_{i=1}^n \int (\sqrt{dP_{i,\theta}} - \sqrt{dP_{i,\theta'}})^2 \right)^{1/2}$$

Lemma

For all here exists tests ϕ_n such that

$$P_{\theta_0}^n \phi_n \leq e^{-n d_{n,H}(\theta_0, \theta_1)}, \quad P_\theta^n \leq e^{-n d_{n,H}(\theta_0, \theta_1)}$$

for all θ such that $d_{n,H}(\theta, \theta_1) \leq \frac{1}{18} d_{n,H}(\theta_0, \theta_1)$

Independent observations

- Assume that the measure $P_\theta^n = \bigotimes_{i=1}^n P_{i,\theta}$ on some product space $\bigotimes_{i=1}^n \{\mathcal{X}_i, \mathcal{A}_i\}$.
- Assume that each measures $P_{i,\theta}$ are absolutely continuous w.r.t μ_i
- Define the Root average Hellinger distance

$$d_{n,H}(\theta, \theta') = \left(\frac{1}{n} \sum_{i=1}^n \int (\sqrt{dP_{i,\theta}} - \sqrt{dP_{i,\theta'}})^2 \right)^{1/2}$$

Lemma

For all here exists tests ϕ_n such that

$$P_{\theta_0}^n \phi_n \leq e^{-n d_{n,H}(\theta_0, \theta_1)}, \quad P_\theta^n \leq e^{-n d_{n,H}(\theta_0, \theta_1)}$$

for all θ such that $d_{n,H}(\theta, \theta_1) \leq \frac{1}{18} d_{n,H}(\theta_0, \theta_1)$

Independent observations

- Assume that the measure $P_\theta^n = \bigotimes_{i=1}^n P_{i,\theta}$ on some product space $\bigotimes_{i=1}^n \{\mathcal{X}_i, \mathcal{A}_i\}$.
- Assume that each measures $P_{i,\theta}$ are absolutely continuous w.r.t μ_i
- Define the Root average Hellinger distance

$$d_{n,H}(\theta, \theta') = \left(\frac{1}{n} \sum_{i=1}^n \int (\sqrt{dP_{i,\theta}} - \sqrt{dP_{i,\theta'}})^2 \right)^{1/2}$$

Lemma

For all here exists tests ϕ_n such that

$$P_{\theta_0}^n \phi_n \leq e^{-n d_{n,H}(\theta_0, \theta_1)}, \quad P_\theta^n \leq e^{-n d_{n,H}(\theta_0, \theta_1)}$$

for all θ such that $d_{n,H}(\theta, \theta_1) \leq \frac{1}{18} d_{n,H}(\theta_0, \theta_1)$

Independent observations

- Assume that the measure $P_\theta^n = \bigotimes_{i=1}^n P_{i,\theta}$ on some product space $\bigotimes_{i=1}^n \{\mathcal{X}_i, \mathcal{A}_i\}$.
- Assume that each measures $P_{i,\theta}$ are absolutely continuous w.r.t μ_i
- Define the Root average Hellinger distance

$$d_{n,H}(\theta, \theta') = \left(\frac{1}{n} \sum_{i=1}^n \int (\sqrt{dP_{i,\theta}} - \sqrt{dP_{i,\theta'}})^2 \right)^{1/2}$$

Lemma

For all here exists tests ϕ_n such that

$$P_{\theta_0}^n \phi_n \leq e^{-n d_{n,H}(\theta_0, \theta_1)}, \quad P_\theta^n \leq e^{-n d_{n,H}(\theta_0, \theta_1)}$$

for all θ such that $d_{n,H}(\theta, \theta_1) \leq \frac{1}{18} d_{n,H}(\theta_0, \theta_1)$

Independent observations

We can also simplify the KL condition in this case. Note that

$$KL(p^n_{\theta_0}), p^n_\theta) = \sum_{i=1}^{n} \mathrm{KL}(p_{i,\theta_0}, p_{i,\theta})$$

Furthermore for the KL-variation term we have that

$$V_{k,0}(p^n_{\theta_0}, p^n_\theta) \leq n^{k/2} C_k \frac{1}{n} \sum_{i=1}^{n} V_{k,0}(p_{i,\theta_0}, p_{i,\theta})$$

Thus the KL condition can be re-written

$$B^*_n(\theta_0, \epsilon, k) = \left\{ \theta, \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}(p_{i,\theta_0}, p_{i,\theta}) \leq \epsilon^2, \frac{1}{n} \sum_{i=1}^{n} V_{k,0}(p_{i,\theta_0}, p_{i,\theta}) \leq C_k \epsilon^2 \right\}$$

Independent observations

We can also simplify the KL condition in this case. Note that

$$KL(p^n_{\theta_0}), p^n_\theta) = \sum_{i=1}^{n} \mathrm{KL}(p_{i,\theta_0}, p_{i,\theta})$$

Furthermore for the KL-variation term we have that

$$V_{k,0}(p^n_{\theta_0}, p^n_\theta) \leq n^{k/2} C_k \frac{1}{n} \sum_{i=1}^{n} V_{k,0}(p_{i,\theta_0}, p_{i,\theta})$$

Thus the KL condition can be re-written

$$B^*_n(\theta_0, \epsilon, k) = \left\{\theta, \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}(p_{i,\theta_0}, p_{i,\theta}) \leq \epsilon^2, \frac{1}{n}\sum_{i=1}^{n} V_{k,0}(p_{i,\theta_0}, p_{i,\theta}) \leq C_k \epsilon^2\right\}$$

Independent observations

We can also simplify the KL condition in this case. Note that

$$KL(p^n_{\theta_0}), p^n_\theta) = \sum_{i=1}^{n} \mathrm{KL}(p_{i,\theta_0}, p_{i,\theta})$$

Furthermore for the KL-variation term we have that

$$V_{k,0}(p^n_{\theta_0}, p^n_\theta) \leq n^{k/2} C_k \frac{1}{n} \sum_{i=1}^{n} V_{k,0}(p_{i,\theta_0}, p_{i,\theta})$$

Thus the KL condition can be re-written

$$B^*_n(\theta_0, \epsilon, k) = \left\{\theta, \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}(p_{i,\theta_0}, p_{i,\theta}) \leq \epsilon^2, \frac{1}{n}\sum_{i=1}^{n} V_{k,0}(p_{i,\theta_0}, p_{i,\theta}) \leq C_k \epsilon^2\right\}$$

Independent observations

We can also simplify the KL condition in this case. Note that

$$KL(p^n_{\theta_0}), p^n_\theta) = \sum_{i=1}^{n} \mathrm{KL}(p_{i,\theta_0}, p_{i,\theta})$$

Furthermore for the KL-variation term we have that

$$V_{k,0}(p^n_{\theta_0}, p^n_\theta) \leq n^{k/2} C_k \frac{1}{n} \sum_{i=1}^{n} V_{k,0}(p_{i,\theta_0}, p_{i,\theta})$$

Thus the KL condition can be re-written

$$B^*_n(\theta_0, \epsilon, k) = \left\{\theta, \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}(p_{i,\theta_0}, p_{i,\theta}) \leq \epsilon^2, \frac{1}{n}\sum_{i=1}^{n} V_{k,0}(p_{i,\theta_0}, p_{i,\theta}) \leq C_k \epsilon^2\right\}$$

Consider the model

$$X_i = f(z_i) + \epsilon_i, \ \ i = 1, \ldots, n$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and the $z_i \in \mathbb{L}$ are known fixed covariates. For simplicity σ^2 is also assume to be known. Let $\mathbb{P}_n^z = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and $\|\cdot\|_n$ the $L_2(\mathbb{P}_n^z)$ norm

Lemma

We have the following results

$$KL(P_{f,i}, P_{g,i}) = \frac{1}{2\sigma^2}(f(z_i) - g(z_i))^2$$

$$V_{0,2}(P_{f,i}, P_{g,i}) = \frac{1}{\sigma^2}(f(z_i) - g(z_i))^2$$

Consider the model

$$X_i = f(z_i) + \epsilon_i, \ \ i = 1, \ldots, n$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and the $z_i \in \mathbb{L}$ are known fixed covariates. For simplicity σ^2 is also assume to be known. Let $\mathbb{P}_n^z = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and $\|\cdot\|_n$ the $L_2(\mathbb{P}_n^z)$ norm

Lemma

We have the following results

$$KL(P_{f,i}, P_{g,i}) = \frac{1}{2\sigma^2}(f(z_i) - g(z_i))^2$$

$$V_{0,2}(P_{f,i}, P_{g,i}) = \frac{1}{\sigma^2}(f(z_i) - g(z_i))^2$$

Consider the model

$$X_i = f(z_i) + \epsilon_i, \ \ i = 1, \ldots, n$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and the $z_i \in \mathbb{L}$ are known fixed covariates. For simplicity σ^2 is also assume to be known. Let $\mathbb{P}_n^z = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and $\|\cdot\|_n$ the $L_2(\mathbb{P}_n^z)$ norm

Lemma

We have the following results

$$KL(P_{f,i}, P_{g,i}) = \frac{1}{2\sigma^2}(f(z_i) - g(z_i))^2$$

$$V_{0,2}(P_{f,i}, P_{g,i}) = \frac{1}{\sigma^2}(f(z_i) - g(z_i))^2$$

NP Regression with splines

Assume that $f_0 \in \mathcal{H}(\alpha, L)$ such that $\|f_0\|_\infty \leq H$, then the $d^2_{n,H}$ and $\|\cdot\|^2_n$ are equivalent.

Spline prior

Consider $(B_j)_{j=1}^J$ the B-splines basis with J equally spaced nodes, and consider

$$f_\beta(\cdot) = \sum_{j=1}^{J} \beta_j B_j(\cdot)$$

and induce a prior on f by choosing a prior on β, $\beta_j \overset{iid}{\sim} g$.

Approximation techniques with splines gives us that for $\beta^* \in \mathbb{L}^J$ the coefficient of the projection of f_0 in $\mathrm{Span}(B_j)$,

$$\|f_{\beta^*} - f_0\|_\infty \leq J^{-\alpha}\|f_0\|_\alpha$$

Assume that $f_0 \in \mathcal{H}(\alpha, L)$ such that $\|f_0\|_\infty \leq H$, then the $d^2_{n,H}$ and $\|\cdot\|^2_n$ are equivalent.

Spline prior

Consider $(B_j)_{j=1}^J$ the B-splines basis with J equally spaced nodes, and consider

$$f_\beta(\cdot) = \sum_{j=1}^{J} \beta_j B_j(\cdot)$$

and induce a prior on f by choosing a prior on β, $\beta_j \overset{iid}{\sim} g$.

Approximation techniques with splines gives us that for $\beta^* \in \mathbb{L}^J$ the coefficient of the projection of f_0 in $\mathrm{Span}(B_j)$,

$$\|f_{\beta^*} - f_0\|_\infty \leq J^{-\alpha}\|f_0\|_\alpha$$

Assume that $f_0 \in \mathcal{H}(\alpha, L)$ such that $\|f_0\|_\infty \leq H$, then the $d^2_{n,H}$ and $\|\cdot\|^2_n$ are equivalent.

Spline prior

Consider $(B_j)_{j=1}^J$ the B-splines basis with J equally spaced nodes, and consider

$$f_\beta(\cdot) = \sum_{j=1}^{J} \beta_j B_j(\cdot)$$

and induce a prior on f by choosing a prior on β, $\beta_j \overset{iid}{\sim} g$.

Approximation techniques with splines gives us that for $\beta^* \in \mathbb{L}^J$ the coefficient of the projection of f_0 in $\mathrm{Span}(B_j)$,

$$\|f_{\beta^*} - f_0\|_\infty \leq J^{-\alpha} \|f_0\|_\alpha$$

NP Regression with splines

We also need to impose conditions on the design. Let Σ_n be such that $\Sigma_{n,i,j} = \int B_i B_j d\mathbb{P}_n^z$. We assume that

$$J^{-1}\|\beta\|^2 \asymp \beta' \Sigma_n \beta$$

so that

$$\|f_{\beta_1} - f_{\beta_2}\|_n \asymp \sqrt{J}\|\beta_1 - \beta_2\|$$

We can thus perform calculations in terms of the Euclidean norm of the coefficients.

We also need to impose conditions on the design.Let Σ_n be such that $\Sigma_{n,i,j} = \int B_i B_j d\mathbb{P}_n^z$. We assume that

$$J^{-1}\|\beta\|^2 \asymp \beta'\Sigma_n\beta$$

so that

$$\|f_{\beta_1} - f_{\beta_2}\|_n \asymp \sqrt{J}\|\beta_1 - \beta_2\|$$

We can thus perform calculations in terms of the Euclidean norm of the coefficients.

We also need to impose conditions on the design.Let Σ_n be such that $\Sigma_{n,i,j} = \int B_i B_j d\mathbb{P}_n^z$. We assume that

$$J^{-1}\|\beta\|^2 \asymp \beta' \Sigma_n \beta$$

so that

$$\|f_{\beta_1} - f_{\beta_2}\|_n \asymp \sqrt{J}\|\beta_1 - \beta_2\|$$

We can thus perform calculations in terms of the Euclidean norm of the coefficients.

Theorem

Assume that g is a standard Gaussian distribution, and assume that $J = J_n \asymp n^{1/(2\alpha+1)}$, then the posterior contracts at a rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}$.

- This is the minimax rate, in addition this rate is uniform over all bounded $\mathcal{H}(\alpha, L)$ functions.
- Some condition can be relaxed, in particular, g could be any distribution such that for every β^* such that $\|\beta^*\|_\infty \leq C$ $\Pi(\|\beta - \beta^*\| \leq \epsilon) \geq e^{-cJ\log(1/\epsilon)}$. Some log factor may appear in the rate.
- The boundedness condition could also be dropped by considering likelihood ratio tests for $\|\cdot\|_2$ norm.

Theorem

Assume that g is a standard Gaussian distribution, and assume that $J = J_n \asymp n^{1/(2\alpha+1)}$, then the posterior contracts at a rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}$.

- This is the minimax rate, in addition this rate is uniform over all bounded $\mathcal{H}(\alpha, L)$ functions.
- Some condition can be relaxed, in particular, g could be any distribution such that for every β^* such that $\|\beta^*\|_\infty \leq C$ $\Pi(\|\beta - \beta^*\| \leq \epsilon) \geq e^{-cJ\log(1/\epsilon)}$. Some log factor may appear in the rate.
- The boundedness condition could also be dropped by considering likelihood ratio tests for $\|\cdot\|_n$ norm.

Theorem

Assume that g is a standard Gaussian distribution, and assume that $J = J_n \asymp n^{1/(2\alpha+1)}$, then the posterior contracts at a rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}$.

- This is the minimax rate, in addition this rate is uniform over all bounded $\mathcal{H}(\alpha, L)$ functions.
- Some condition can be relaxed, in particular, g could be any distribution such that for every β^* such that $\|\beta^*\|_\infty \leq C$ $\Pi(\|\beta - \beta^*\| \leq \epsilon) \geq e^{-cJ \log(1/\epsilon)}$. Some log factor may appear in the rate.
- The boundedness condition could also be dropped by considering likelihood ratio tests for $\|\cdot\|_n$ norm.

Theorem

Assume that g is a standard Gaussian distribution, and assume that $J = J_n \asymp n^{1/(2\alpha+1)}$, then the posterior contracts at a rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}$.

- This is the minimax rate, in addition this rate is uniform over all bounded $\mathcal{H}(\alpha, L)$ functions.
- Some condition can be relaxed, in particular, g could be any distribution such that for every β^* such that $\|\beta^*\|_\infty \leq C$ $\Pi(\|\beta - \beta^*\| \leq \epsilon) \geq e^{-cJ \log(1/\epsilon)}$. Some log factor may appear in the rate.
- The boundedness condition could also be dropped by considering likelihood ratio tests for $\|\cdot\|_n$ norm.

Acknowledgements

I would like to thank Jean-Bernard Salomond and Botond Szabo for sharing his expertise and slides on asymptotic aspects of Bayesian nonparametric procedures.

References I

[1] Charles E Antoniak. "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems". In: *The Annals of Statistics* (1974), pp. 1152–1174.

[2] Julyan Arbel et al. "BNPdensity: Bayesian nonparametric mixture modeling in R". In: *Australian & New Zealand Journal of Statistics* 63 (3 2021), pp. 542–564. DOI: 10.1111/anzs.12342. eprint: 2110.10019.

[3] David M Blei, Michael I Jordan, et al. "Variational inference for Dirichlet process mixtures". In: *Bayesian analysis* 1.1 (2006), pp. 121–144.

[4] David M Blei et al. "Latent Dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[5] Anders Brix. "Generalized gamma measures and shot-noise Cox processes". In: *Advances in Applied Probability* (1999), pp. 929–953.

[6] Aaron Clauset et al. "Power-law distributions in empirical data". In: *SIAM review* 51.4 (2009), pp. 661–703.

[7] David B Dahl. "Model-based clustering for expression data via a Dirichlet process mixture model". In: *Bayesian inference for gene expression and proteomics* (2006), pp. 201–218.

References II

[8] Pierpaolo De Blasi et al. "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2 (2015), pp. 212–229.

[9] Warren J Ewens. "The sampling theory of selectively neutral alleles". In: *Theoretical population biology* 3.1 (1972), pp. 87–112.

[10] T.S. Ferguson. "A Bayesian analysis of some nonparametric problems". In: *The Annals of Statistics* 1.2 (1973), pp. 209–230. ISSN: 0090-5364.

[11] Zoubin Ghahramani and Thomas L Griffiths. "Infinite latent feature models and the Indian buffet process". In: *Advances in neural information processing systems*. 2006, pp. 475–482.

[12] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press, 2017.

[13] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. New York: Springer, 2003.

References III

[14] Nils Lid Hjort et al. *Bayesian nonparametrics*. Vol. 28. Cambridge University Press, Apr. 2010. URL: http://www.cambridge.org/us/academic/subjects/statistics-probability/statistical-theory-and-methods/bayesian-nonparametrics.

[15] H. Ishwaran and L.F. James. "Gibbs sampling methods for stick-breaking priors". In: *Journal of the American Statistical Association* 96.453 (2001), pp. 161–173. ISSN: 0162-1459.

[16] Stephan Mandt et al. "Stochastic Gradient Descent as Approximate Bayesian Inference". In: *J. Mach. Learn. Res.* 18.1 (Jan. 2017), pp. 4873–4907. ISSN: 1532-4435.

[17] Jeffrey W Miller and Matthew T Harrison. "A simple example of Dirichlet process mixture inconsistency for the number of components". In: *Advances in neural information processing systems*. 2013, pp. 199–206.

[18] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL: http://probml.github.io/book2.

[19] Radford M Neal. "Markov chain sampling methods for Dirichlet process mixture models". In: *Journal of computational and graphical statistics* 9.2 (2000), pp. 249–265.

References IV

[20] Mark EJ Newman. "Power laws, Pareto distributions and Zipf's law". In: *Contemporary physics* 46.5 (2005), pp. 323–351.

[21] Jim Pitman. "Poisson-Kingman partitions". In: *Lecture Notes-Monograph Series* (2003), pp. 1–34.

[22] Jim Pitman and Marc Yor. "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator". In: *The Annals of Probability* 25.2 (1997), pp. 855–900.

[23] Łukasz Rajkowski. "Analysis of the maximal a posteriori partition in the Gaussian Dirichlet Process Mixture Model". In: *Bayesian Analysis* 14.2 (2019), pp. 477–494.

[24] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. DOI: 10.1.1.86.3414.

[25] Jayaram Sethuraman. "A constructive definition of Dirichlet priors". In: *Statistica Sinica* 4 (1994), pp. 639–650.

[26] Y.W. Teh et al. "Hierarchical Dirichlet processes". In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1566–1581. ISSN: 0162-1459.

References V

[27] Sara Wade and Zoubin Ghahramani. "Bayesian cluster analysis: Point estimation and credible balls (with discussion)". In: *Bayesian Analysis* 13.2 (2018), pp. 559–626.