STAT 576 Bayesian Analysis

Lecture 13: Nonparametric Models

Chencheng Cai

Washington State University

Prior Assumptions

Given a set of paired data $(x_1, y_1), \ldots, (x_n, y_n)$, we often assume that the expected value of y is a function of x:

$$y_i = \mu(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

for some unknown function f. The ϵ_i's are assumed to have mean zero.

Prior Assumptions

Given a set of paired data $(x_1, y_1), \ldots, (x_n, y_n)$, we often assume that the expected value of y is a function of x:

$$y_i = \mu(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

for some unknown function f. The ϵ_i's are assumed to have mean zero.

- In parametric models (e.g. linear regressions), we assume that μ belongs to a parametric family, e.g. $\mu(x) = \beta_0 + \beta_1 x$. A prior is often placed on the parameters β_0 and β_1.

Prior Assumptions

Given a set of paired data $(x_1, y_1), \ldots, (x_n, y_n)$, we often assume that the expected value of y is a function of x:

$$y_i = \mu(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

for some unknown function f. The ϵ_i's are assumed to have mean zero.

- In parametric models (e.g. linear regressions), we assume that μ belongs to a parametric family, e.g. $\mu(x) = \beta_0 + \beta_1 x$. A prior is often placed on the parameters β_0 and β_1.
- In state-space models, the prior is placed on the state variables $(x_1, \ldots, x_n)$ through a transition model.

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}).$$

Prior Assumptions

Given a set of paired data $(x_1, y_1), \ldots, (x_n, y_n)$, we often assume that the expected value of y is a function of x:

$$y_i = \mu(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

for some unknown function f. The ϵ_i's are assumed to have mean zero.

- In parametric models (e.g. linear regressions), we assume that μ belongs to a parametric family, e.g. $\mu(x) = \beta_0 + \beta_1 x$. A prior is often placed on the parameters β_0 and β_1.
- In state-space models, the prior is placed on the state variables $(x_1, \ldots, x_n)$ through a transition model.

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}).$$

- In this lecture, we will discuss nonparametric models, where the prior is placed directly on the function μ within a function space.

Gaussian Process

- A **stochastic process** is a collection of random variables indexed by some set, e.g. time or space.
  - Random walk: r.v.s. indexed by time.
  - Brownian motion: r.v.s. indexed by time.
  - Random field: r.v.s. indexed by space.

Gaussian Process

- A **stochastic process** is a collection of random variables indexed by some set, e.g. time or space.
  - Random walk: r.v.s. indexed by time.
  - Brownian motion: r.v.s. indexed by time.
  - Random field: r.v.s. indexed by space.
- A **Gaussian process** is a stochastic process such that any finite collection of r.v.s. has a multivariate normal distribution.

Gaussian Process

- A **stochastic process** is a collection of random variables indexed by some set, e.g. time or space.
  - Random walk: r.v.s. indexed by time.
  - Brownian motion: r.v.s. indexed by time.
  - Random field: r.v.s. indexed by space.
- A **Gaussian process** is a stochastic process such that any finite collection of r.v.s. has a multivariate normal distribution.
- Specifically, if $\{\mu(x) : x \in \mathcal{X}\}$ is a Gaussian process, then for any finite set of indices $x_1, \ldots, x_n \in \mathcal{X}$, the random vector $(\mu(x_1), \ldots, \mu(x_n))$ has a multivariate normal distribution.
- As a special case, for any $x \in \mathcal{X}$, $\mu(x)$ is a normal random variable.

Gaussian Process

- A Gaussian process is completely specified by its mean function $m(x) = E[\mu(x)]$ and covariance function $k(x, x') = \text{Cov}(\mu(x), \mu(x'))$.
- The process is denoted by $\mu(x) \sim \mathcal{GP}(m, k)$.

Gaussian Process

- A Gaussian process is completely specified by its mean function $m(x) = E[\mu(x)]$ and covariance function $k(x, x') = \text{Cov}(\mu(x), \mu(x'))$.
- The process is denoted by $\mu(x) \sim \mathcal{GP}(m, k)$.
- The joint distribution of $\mu(x_1), \ldots, \mu(x_n)$ is given by

$$\begin{pmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix}, K(x_1, \ldots, x_n) \right).$$

where $K(x_1, \ldots, x_n)$ is the covariance matrix with (i, j)-th element $k(x_i, x_j)$.

Gaussian Process

- A Gaussian process is completely specified by its mean function $m(x) = E[\mu(x)]$ and covariance function $k(x, x') = \text{Cov}(\mu(x), \mu(x'))$.
- The process is denoted by $\mu(x) \sim \mathcal{GP}(m, k)$.
- The joint distribution of $\mu(x_1), \ldots, \mu(x_n)$ is given by

$$\begin{pmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix}, K(x_1, \ldots, x_n) \right).$$

  where $K(x_1, \ldots, x_n)$ is the covariance matrix with (i, j)-th element $k(x_i, x_j)$.
- Consistent defintion: the distribution of $\mu(x_1), \ldots, \mu(x_m)$ derived from the joint distribution of $\mu(x_1), \ldots, \mu(x_n)$ is the same for any choice of $x_{m+1}, \ldots, x_n$.
- $K(x_1, \ldots, x_n)$ is positive definite for any choice of $x_1, \ldots, x_n$.

Gaussian Process — Covariance

A common choice for $k(x, x')$ is

$$k(x, x') = \tau^2 \exp\left(-\frac{|x - x'|^2}{2l^2}\right),$$

Gaussian Process — Covariance

A common choice for $k(x, x')$ is

$$k(x, x') = \tau^2 \exp\left(-\frac{|x - x'|^2}{2l^2}\right),$$

Show that the covariance matrix $K(x_1, \ldots, x_n)$ is positive definite.

Gaussian Process — Covariance

A common choice for $k(x, x')$ is

$$k(x, x') = \tau^2 \exp\left(-\frac{|x - x'|^2}{2l^2}\right),$$

Show that the covariance matrix $K(x_1, \ldots, x_n)$ is positive definite.

WLOG, we assume $\tau^2 = l^2 = 1$. To show K is positive definite, we need to show that for any vector $u = (u_1, \ldots, u_n)$, $u^T K u \geq 0$.

- Notice that
$$k(x_i, x_j) = \exp\left(-\frac{1}{2}|x_i - x_j|^2\right) = \mathbb{E}\left[e^{i|x_i - x_j|Z}\right]$$
for $Z \sim \mathcal{N}(0, 1)$.
- Therefore,

$$u^T K u = \sum_{i,j} u_i u_j k(x_i, x_j) = \sum_{i,j} u_i u_j \mathbb{E}\left[e^{i|x_i - x_j|Z}\right] = \mathbb{E}\left[\left(\sum_i u_i e^{ix_i Z}\right)^2\right] \geq 0.$$

Gaussian Process — Basis Functions

The Gaussian Process can also be constructed by basis functions:

$$\mu(x) = \sum_{h=1}^{H} \beta_h b_h(x), \quad \beta = (\beta_1, \ldots, \beta_H) \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta)$$

Gaussian Process — Basis Functions

The Gaussian Process can also be constructed by basis functions:

$$\mu(x) = \sum_{h=1}^{H} \beta_h b_h(x), \quad \beta = (\beta_1, \ldots, \beta_H) \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta)$$

Then μ is a Gaussian process with

$$m(x) = \boldsymbol{b}(x)^T \boldsymbol{\beta}_0, \quad k(x, x') = \boldsymbol{b}(x)^T \boldsymbol{\Sigma}_\beta \boldsymbol{b}(x').$$

Gaussian Process — Inference

Suppose, we have observed the data $(x_1, y_1), \ldots, (x_n, y_n)$, and we assume that

$$y_i = \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Given a new point $\tilde{x}$, we want to estimate the expected value of y at $\tilde{x}$, i.e. $\mu(\tilde{x})$.

Gaussian Process — Inference

Suppose, we have observed the data $(x_1, y_1), \ldots, (x_n, y_n)$, and we assume that

$$y_i = \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Given a new point $\tilde{x}$, we want to estimate the expected value of y at $\tilde{x}$, i.e. $\mu(\tilde{x})$.

- We assume that μ is a Gaussian process with mean function $m(x) = 0$ and covariance function $k(x, x') = \tau^2 \exp\left(-|x - x'|^2/(2l^2)\right)$.

Gaussian Process — Inference

Suppose, we have observed the data $(x_1, y_1), \ldots, (x_n, y_n)$, and we assume that

$$y_i = \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Given a new point $\tilde{x}$, we want to estimate the expected value of y at $\tilde{x}$, i.e. $\mu(\tilde{x})$.

- We assume that μ is a Gaussian process with mean function $m(x) = 0$ and covariance function $k(x, x') = \tau^2 \exp\left(-|x - x'|^2/(2l^2)\right)$.
- The joint distribution of $y = (y_1, \ldots, y_n)$ and $\mu(\tilde{x})$ is given by

$$\begin{pmatrix} y \\ \mu(\tilde{x}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(x,x) + \sigma^2 I & K(x, \tilde{x}) \\ K(\tilde{x}, x) & K(\tilde{x}, \tilde{x}) \end{pmatrix}\right).$$

Gaussian Process — Inference

Suppose, we have observed the data $(x_1, y_1), \ldots, (x_n, y_n)$, and we assume that

$$y_i = \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Given a new point $\tilde{x}$, we want to estimate the expected value of y at $\tilde{x}$, i.e. $\mu(\tilde{x})$.

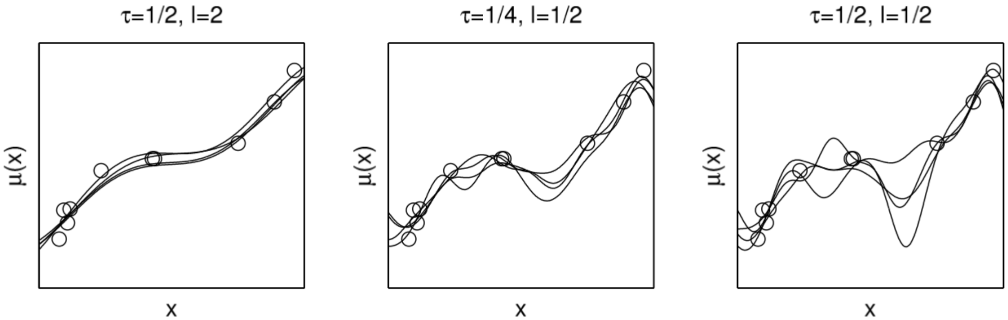
- We assume that μ is a Gaussian process with mean function $m(x) = 0$ and covariance function $k(x, x') = \tau^2 \exp\left(-|x - x'|^2/(2l^2)\right)$.
- The joint distribution of $y = (y_1, \ldots, y_n)$ and $\mu(\tilde{x})$ is given by

$$\begin{pmatrix} y \\ \mu(\tilde{x}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(x,x) + \sigma^2 I & K(x,\tilde{x}) \\ K(\tilde{x},x) & K(\tilde{x},\tilde{x}) \end{pmatrix}\right).$$

- With the properties of conditional distribution of multivariate normal, we can derive the posterior distribution of $\mu(\tilde{x})$ given y.

$$\begin{aligned} &\mu(\tilde{x}) \mid x, y, \tau^2, \sigma^2, l^2 \\ &\sim \mathcal{N}\left(K(\tilde{x},x)(K(x,x)+\sigma^2 I)^{-1} y, K(\tilde{x},\tilde{x}) - K(\tilde{x},x)(K(x,x)+\sigma^2 I)^{-1} K(x,\tilde{x})\right) \end{aligned}$$

Gaussian Process — Example



Gaussian Process — Inference

For a Bayesian procedure, we need to specify the prior distributions for the hyperparameters τ^2, σ^2, and l^2.

Gaussian Process — Inference

For a Bayesian procedure, we need to specify the prior distributions for the hyperparameters τ^2, σ^2, and l^2.
A common choice is

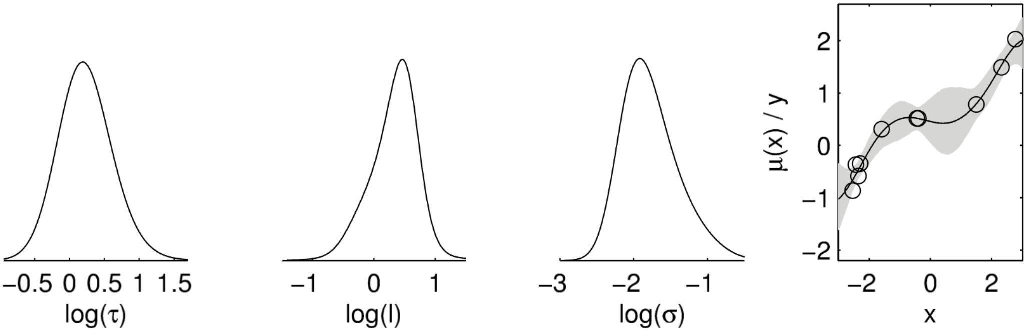
$$p(\log \tau) \propto 1, \quad p(\log \sigma) \propto 1, \quad p(\log l) \propto 1.$$

The log-likelihood is

$$\log p(y \mid x, \tau^2, \sigma^2, l^2) = -\frac{1}{2} y^T (K(x,x) + \sigma^2 I)^{-1} y - \frac{1}{2} \log |K(x,x) + \sigma^2 I| - \frac{n}{2} \log(2\pi).$$

The posterior is now straightforward.

Gaussian Process — Inference



Example — Birth Dates

In this example, we analyze the patterns of birthdays in the United States. The data is the number of births on each day of the year from 1969 to 1988.

- This is a time series data, where the index is the number of days from 1969-01-01.
- The series contains periodic patterns, e.g. yearly and weekly patterns.
- The series also contains long term trends.

Example — Birth Dates

We model the time series as an additive model:

$$y(t) = f_1(t) + f_2(t) + f_3(t) + f_4(t) + f_5(t) + \epsilon_t,$$

- Long-term trend:

$$f_1(t) \sim \mathcal{GP}(0, k_1), \quad k_1(t, t') = \sigma_1^2 \exp\left(-\frac{|t - t'|^2}{2l_1^2}\right)$$

- Short-term trend:

$$f_2(t) \sim \mathcal{GP}(0, k_2), \quad k_2(t, t') = \sigma_2^2 \exp\left(-\frac{|t - t'|^2}{2l_2^2}\right)$$

Example — Birth Dates

We model the time series as an additive model:

$$y(t) = f_1(t) + f_2(t) + f_3(t) + f_4(t) + f_5(t) + \epsilon_t,$$

- Weekly pattern:

$$f_3(t) \sim \mathcal{GP}(0, k_3), \quad k_3(t, t') = \sigma_3^2 \exp\left(-\frac{2\sin^2(\pi(t-t')/7)}{l_{3,1}^2}\right) \exp\left(-\frac{|t-t'|^2}{2l_{3,2}^2}\right)$$

- Yearly pattern:

$$f_4(t) \sim \mathcal{GP}(0, k_4), \quad k_4(t, t') = \sigma_4^2 \exp\left(-\frac{2\sin^2(\pi(t-t')/365.25)}{l_{4,1}^2}\right) \exp\left(-\frac{|t-t'|^2}{2l_{4,2}^2}\right)$$

Example — Birth Dates

We model the time series as an additive model:

$$y(t) = f_1(t) + f_2(t) + f_3(t) + f_4(t) + f_5(t) + \epsilon_t,$$

- Special days and its interaction with weekends:

  $$f_5(t) = I_{s.d.}(t)\beta_a + I_{s.d.}(t)I_{w.e.}(t)\beta_b$$

  where $I_{s.d.}(t)$ is an indicator function for special days (13 holidays), and $I_{w.e.}(t)$ is an indicator function for weekends.
- $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is the residual.

Example — Birth Dates

- Sum of Gaussian processes is still a Gaussian process.
- The model can be fit through a standard GP inference.

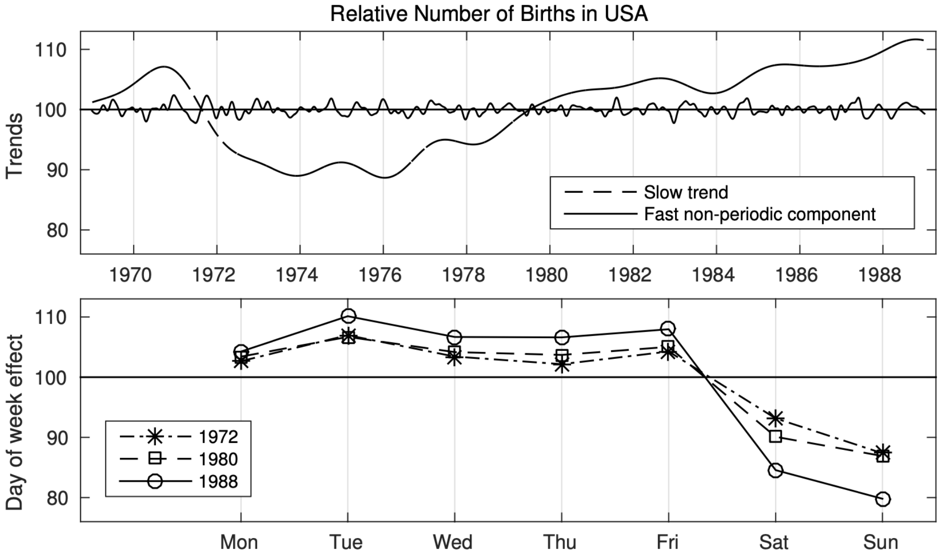
Example — Birth Dates

- Sum of Gaussian processes is still a Gaussian process.
- The model can be fit through a standard GP inference.
- log-t prior for time scales l.
- log-unifom prior for other parameters.

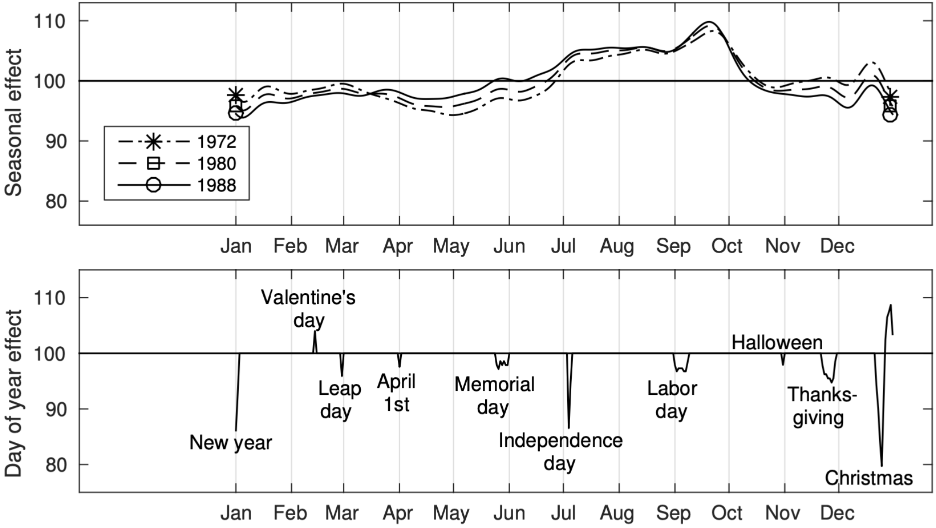
Example — Birth Dates

- Sum of Gaussian processes is still a Gaussian process.
- The model can be fit through a standard GP inference.
- log-t prior for time scales l.
- log-unifom prior for other parameters.
- The model can be further extended by considering weekdays v.s. weekends. See textbook Ch. 21.2.

Example — Birth Dates



Example — Birth Dates



Dirichlet Process

- The Gaussian process gives a prior on the function space.
- A special type of the function space is the distribution space — all nonnegative integrable functions with integral 1.

Dirichlet Process

- The Gaussian process gives a prior on the function space.
- A special type of the function space is the distribution space — all nonnegative integrable functions with integral 1.
- In defining Gaussian process, we considered the joint distribution of the function values at any finite number of points.

Dirichlet Process

- The Gaussian process gives a prior on the function space.
- A special type of the function space is the distribution space — all nonnegative integrable functions with integral 1.
- In defining Gaussian process, we considered the joint distribution of the function values at any finite number of points.
- For the distribution space, we consider probabilites of any finite partitions of the sample space.

Dirichlet Process

Consider a finite partition of the sample space:

$$\Omega = B_1 \cup B_2 \cup \cdots \cup B_k, \text{ and } B_i \cap B_j = \varnothing \; \forall \; i \neq j$$

Dirichlet Process

Consider a finite partition of the sample space:

$$\Omega = B_1 \cup B_2 \cup \cdots \cup B_k, \text{ and } B_i \cap B_j = \varnothing \ \forall \ i \neq j$$

Let P be a probaility measure on Ω with density function f. The probability measures of the partitions are

$$(P(B_1), \dots, P(B_k)) = \left(\int_{B_1} f(y)dy, \dots, \int_{B_k} f(y)dy \right)$$

Dirichlet Process

Consider a finite partition of the sample space:

$$\Omega = B_1 \cup B_2 \cup \cdots \cup B_k, \text{ and } B_i \cap B_j = \varnothing \ \forall \ i \neq j$$

Let P be a probaility measure on Ω with density function f. The probability measures of the partitions are

$$(P(B_1), \ldots, P(B_k)) = \left(\int_{B_1} f(y)dy, \ldots, \int_{B_k} f(y)dy \right)$$

with

$$\sum_{i=1}^{k} P(B_i) = 1.$$

Dirichlet Process

A natural probability measure on $(P(B_1), \dots, P(B_k))$ is the Dirichlet distribution:

$$(P(B_1), \dots, P(B_k)) \sim \text{Dirichlet}\,(\alpha P_0(B_1), \dots, \alpha P_0(B_k))$$

where P_0 is some baseline measure on Ω.

Dirichlet Process

A natural probability measure on $(P(B_1), \dots, P(B_k))$ is the Dirichlet distribution:

$$(P(B_1), \dots, P(B_k)) \sim \text{Dirichlet}\,(\alpha P_0(B_1), \dots, \alpha P_0(B_k))$$

where P_0 is some baseline measure on Ω.

Now we assume P is a random measure, and the distribution of P is the **Dirichlet process**, denoted by $\mathcal{DP}(\alpha P_0)$ if:

Dirichlet Process

A natural probability measure on $(P(B_1), \ldots, P(B_k))$ is the Dirichlet distribution:

$$(P(B_1), \ldots, P(B_k)) \sim \text{Dirichlet}\left(\alpha P_0(B_1), \ldots, \alpha P_0(B_k)\right)$$

where P_0 is some baseline measure on Ω.

Now we assume P is a random measure, and the distribution of P is the **Dirichlet process**, denoted by $\mathcal{DP}(\alpha P_0)$ if:

for any finite partition $B_1, \ldots, B_k$ of Ω, the probability measures $(P(B_1), \ldots, P(B_k))$ follows the Dirichlet distribution with parameters $(\alpha P_0(B_1), \ldots, \alpha P_0(B_k))$.

Dirichlet Process

To check the consistency of the defition of a Dirichlet process, consider two partitions $B_1, \ldots, B_k$ and $B'_1, \ldots, B'_{k-1}$ with

$$B_i = B'_i \text{ for } 1 \leq i \leq k-2, \text{ and } B'_{k-1} = B_{k-1} \cup B_k$$

Dirichlet Process

To check the consistency of the defition of a Dirichlet process, consider two partitions $B_1, \ldots, B_k$ and $B'_1, \ldots, B'_{k-1}$ with

$$B_i = B'_i \text{ for } 1 \le i \le k-2, \text{ and } B'_{k-1} = B_{k-1} \cup B_k$$

- On the one hand, the distribution of $(P(B_1), \ldots, P(B_k))$ is Dirichlet with parameters $(\alpha P_0(B_1), \ldots, \alpha P_0(B_k))$. Therefore, the distribution of $(P(B_1)', \ldots, P(B'_{k-1}))$ is Dirichlet with parameters

$$(\alpha P_0(B_1), \ldots, \alpha P_0(B_{k-2}), \alpha P_0(B'_{k-1}) + \alpha P_0(B'_k)).$$

Dirichlet Process

To check the consistency of the defition of a Dirichlet process, consider two partitions $B_1, \ldots, B_k$ and $B'_1, \ldots, B'_{k-1}$ with

$$B_i = B'_i \text{ for } 1 \leq i \leq k-2, \text{ and } B'_{k-1} = B_{k-1} \cup B_k$$

- On the one hand, the distribution of $(P(B_1), \ldots, P(B_k))$ is Dirichlet with parameters $(\alpha P_0(B_1), \ldots, \alpha P_0(B_k))$. Therefore, the distribution of $(P(B_1)', \ldots, P(B'_{k-1}))$ is Dirichlet with parameters

$$(\alpha P_0(B_1), \ldots, \alpha P_0(B_{k-2}), \alpha P_0(B'_{k-1}) + \alpha P_0(B'_k)).$$

- On the other hand, the distribution of $(P(B_1)', \ldots, P(B'_{k-1}))$ is Dirichlet with parameters

$$(\alpha P_0(B'_1), \ldots, \alpha P_0(B'_{k-2}), \alpha P_0(B'_{k-1})).$$

Dirichlet Process

To check the consistency of the defition of a Dirichlet process, consider two partitions $B_1, \ldots, B_k$ and $B'_1, \ldots, B'_{k-1}$ with

$$B_i = B'_i \text{ for } 1 \leq i \leq k-2, \text{ and } B'_{k-1} = B_{k-1} \cup B_k$$

- On the one hand, the distribution of $(P(B_1), \ldots, P(B_k))$ is Dirichlet with parameters $(\alpha P_0(B_1), \ldots, \alpha P_0(B_k))$. Therefore, the distribution of $(P(B_1)', \ldots, P(B'_{k-1}))$ is Dirichlet with parameters
$$(\alpha P_0(B_1), \ldots, \alpha P_0(B_{k-2}), \alpha P_0(B'_{k-1}) + \alpha P_0(B'_k)).$$
- On the other hand, the distribution of $(P(B_1)', \ldots, P(B'_{k-1}))$ is Dirichlet with parameters
$$(\alpha P_0(B'_1), \ldots, \alpha P_0(B'_{k-2}), \alpha P_0(B'_{k-1})).$$
- They are equal because
$$B_i = B'_i \text{ for } 1 \leq i \leq k-2, \text{ and } P_0(B'_{k-1}) = P_0(B_{k-1}) + P_0(B_k)$$

Dirichlet Process

Let B be a measurable subset of Ω. Then its probability measure follows a Dirichlet process:

$$(P(B), P(\Omega \setminus B)) \sim \mathrm{Dirichlet}(\alpha P_0(B), \alpha(1 - P_0(B))) \sim \mathrm{Beta}(\alpha P_0(B), \alpha(1 - P_0(B)))$$

Dirichlet Process

Let B be a measurable subset of Ω. Then its probability measure follows a Dirichlet process:

$$(P(B), P(\Omega \setminus B)) \sim \mathrm{Dirichlet}(\alpha P_0(B), \alpha(1 - P_0(B))) \sim \mathrm{Beta}(\alpha P_0(B), \alpha(1 - P_0(B)))$$

Therefore, the expectation of $P(B)$ is

$$E[P(B)] = \frac{\alpha P_0(B)}{\alpha P_0(B) + \alpha(1 - P_0(B))} = P_0(B)$$

Dirichlet Process

Let B be a measurable subset of Ω. Then its probability measure follows a Dirichlet process:

$$(P(B), P(\Omega \setminus B)) \sim \text{Dirichlet}(\alpha P_0(B), \alpha(1 - P_0(B))) \sim \text{Beta}(\alpha P_0(B), \alpha(1 - P_0(B)))$$

Therefore, the expectation of $P(B)$ is

$$E[P(B)] = \frac{\alpha P_0(B)}{\alpha P_0(B) + \alpha(1 - P_0(B))} = P_0(B)$$

and the variance is

$$Var[P(B)] = \frac{\alpha P_0(B)\alpha(1 - P_0(B))}{(\alpha P_0(B) + \alpha(1 - P_0(B))^2)(\alpha P_0(B) + \alpha(1 - P_0(B)))} = \frac{P_0(B)(1 - P_0(B))}{\alpha + 1}$$

Dirichlet Process

Let B be a measurable subset of Ω. Then its probability measure follows a Dirichlet process:

$$(P(B), P(\Omega \setminus B)) \sim \text{Dirichlet}(\alpha P_0(B), \alpha(1 - P_0(B))) \sim \text{Beta}(\alpha P_0(B), \alpha(1 - P_0(B)))$$

Therefore, the expectation of $P(B)$ is

$$E[P(B)] = \frac{\alpha P_0(B)}{\alpha P_0(B) + \alpha(1 - P_0(B))} = P_0(B)$$

and the variance is

$$Var[P(B)] = \frac{\alpha P_0(B)\alpha(1 - P_0(B))}{(\alpha P_0(B) + \alpha(1 - P_0(B))^2)(\alpha P_0(B) + \alpha(1 - P_0(B)))} = \frac{P_0(B)(1 - P_0(B))}{\alpha + 1}$$

Therefore, in the Dirichlet process, P_0 controls the mean measure and α controls the variance.

Dirichlet Process — Inference

Now suppose we observed $y_1, \ldots, y_n$ from some unknown distribution P.

- We choose the prior for P to be a Dirichlet process $\mathcal{DP}(\alpha P_0)$.

Dirichlet Process — Inference

Now suppose we observed $y_1, \ldots, y_n$ from some unknown distribution P.

- We choose the prior for P to be a Dirichlet process $\mathcal{DP}(\alpha P_0)$.
- For any finite partition $B_1, \ldots, B_k$ of Ω, the prior distribution of $(P(B_1), \ldots, P(B_k))$ is Dirichlet with parameters $(\alpha P_0(B_1), \ldots, \alpha P_0(B_k))$.

Dirichlet Process — Inference

Now suppose we observed $y_1, \ldots, y_n$ from some unknown distribution P.

- We choose the prior for P to be a Dirichlet process $\mathcal{DP}(\alpha P_0)$.
- For any finite partition $B_1, \ldots, B_k$ of Ω, the prior distribution of $(P(B_1), \ldots, P(B_k))$ is Dirichlet with parameters $(\alpha P_0(B_1), \ldots, \alpha P_0(B_k))$.
- The likelihood of y given P is

$$p(y \mid P(B_1), \ldots, P(B_k)) = \prod_{j=1}^{k} [P(B_j)]^{\sum_{i=1}^{n} \mathbb{I}\{y_i \in B_j\}}$$

Dirichlet Process — Inference

Now suppose we observed $y_1, \ldots, y_n$ from some unknown distribution P.

- We choose the prior for P to be a Dirichlet process $\mathcal{DP}(\alpha P_0)$.
- For any finite partition $B_1, \ldots, B_k$ of Ω, the prior distribution of $(P(B_1), \ldots, P(B_k))$ is Dirichlet with parameters $(\alpha P_0(B_1), \ldots, \alpha P_0(B_k))$.
- The likelihood of y given P is

$$p(y \mid P(B_1), \ldots, P(B_k)) = \prod_{j=1}^{k} [P(B_j)]^{\sum_{i=1}^{n} \mathbb{I}\{y_i \in B_j\}}$$

- The posterior distribution of $P(B_1), \ldots, P(B_k)$ is

$$\text{Dirichlet}\left(\alpha P_0(B_1) + \sum_i \mathbb{I}\{y_i \in B_1\}, \ldots, \alpha P_0(B_k) + \sum_i \mathbb{I}\{y_i \in B_k\}\right)$$

Dirichlet Process — Inference

The argument on the previous page holds for any finite partition of Ω. Therefore, the posterior distribution of P given y is still a Dirichlet process.

$$\mathcal{DP}\left(\alpha P_0 + \sum_i \delta_{y_i}\right)$$

where δ_{y_i} is the Dirac measure at y_i.

Dirichlet Process — Inference

The argument on the previous page holds for any finite partition of Ω. Therefore, the posterior distribution of P given y is still a Dirichlet process.

$$\mathcal{DP}\left(\alpha P_0 + \sum_i \delta_{y_i}\right)$$

where δ_{y_i} is the Dirac measure at y_i.

For any measurable set B, we have

$$\mathbb{E}[P(B) \mid y] = \frac{\alpha}{\alpha + n} P_0(B) + \frac{n}{\alpha + n} \sum_i \delta_{y_i}(B)$$

Dirichlet Process — Inference

The argument on the previous page holds for any finite partition of Ω. Therefore, the posterior distribution of P given y is still a Dirichlet process.

$$\mathcal{DP}\left(\alpha P_0 + \sum_i \delta_{y_i}\right)$$

where δ_{y_i} is the Dirac measure at y_i.

For any measurable set B, we have

$$\mathbb{E}[P(B) \mid y] = \frac{\alpha}{\alpha + n} P_0(B) + \frac{n}{\alpha + n} \sum_i \delta_{y_i}(B)$$

In the special case that $\alpha = 0$, we have

$$P \mid y \sim \mathcal{DP}\left(\sum_i \delta_{y_i}\right)$$

Dirichlet Process — Stick-breaking Construction

We can construct the Dirichlet process through a stick-breaking process:

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

with

Dirichlet Process — Stick-breaking Construction

We can construct the Dirichlet process through a stick-breaking process:

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

with

$$\pi_h = V_h \prod_{j=1}^{h-1} (1 - V_j), \quad V_h \sim \text{Beta}(1, \alpha)$$
$$\theta_h \sim P_0$$

Dirichlet Process — Stick-breaking Construction

We can construct the Dirichlet process through a stick-breaking process:

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

with

$$\pi_h = V_h \prod_{j=1}^{h-1} (1 - V_j), \quad V_h \sim \mathrm{Beta}(1, \alpha)$$
$$\theta_h \sim P_0$$

It is easy to verify that:

$$\sum_{h=1}^{\infty} \pi_h = 1$$

Dirichlet Process — Stick-breaking Construction

The process can be described as follows:

- Start with a stick of length 1.

Dirichlet Process — Stick-breaking Construction

The process can be described as follows:

- Start with a stick of length 1.
- Break the stick at a random point V_1 with $V_1 \sim \mathrm{Beta}(1, \alpha)$.
- The length of the remaining stick is $1 - V_1$.

Dirichlet Process — Stick-breaking Construction

The process can be described as follows:

- Start with a stick of length 1.
- Break the stick at a random point V_1 with $V_1 \sim \mathrm{Beta}(1, \alpha)$.
- The length of the remaining stick is $1 - V_1$.
- Break the remaining stick at a random point V_2 with $V_2 \sim \mathrm{Beta}(1, \alpha)$.
- The length of the remaining stick is $(1 - V_1)V_2$.

Dirichlet Process — Stick-breaking Construction

The process can be described as follows:

- Start with a stick of length 1.
- Break the stick at a random point V_1 with $V_1 \sim \text{Beta}(1, \alpha)$.
- The length of the remaining stick is $1 - V_1$.
- Break the remaining stick at a random point V_2 with $V_2 \sim \text{Beta}(1, \alpha)$.
- The length of the remaining stick is $(1 - V_1)V_2$.
- Repeat the process.

Dirichlet Process Mixtures

The major drawback of the Dirichlet process is that it is discrete.

Dirichlet Process Mixtures

The major drawback of the Dirichlet process is that it is discrete.

To overcome this, we can use the Dirichlet process as a prior for the mixing distribution in a mixture model.

$$p(y \mid P) = \int \mathcal{K}(y \mid \theta) dP(\theta)$$

where $\mathcal{K}$ is the kernel of the mixture model.

Dirichlet Process Mixtures

The major drawback of the Dirichlet process is that it is discrete.

To overcome this, we can use the Dirichlet process as a prior for the mixing distribution in a mixture model.

$$p(y \mid P) = \int \mathcal{K}(y \mid \theta) dP(\theta)$$

where $\mathcal{K}$ is the kernel of the mixture model.

The model can be written as

$$y_i \sim \mathcal{K}(\theta_i), \quad \theta_i \sim P, \quad P \sim \mathcal{DP}(\alpha P_0)$$

Dirichlet Process Mixtures

$$y_i \sim \mathcal{K}(\theta_i), \quad \theta_i \sim P, \quad P \sim \mathcal{DP}(\alpha P_0)$$

Dirichlet Process Mixtures

$$y_i \sim \mathcal{K}(\theta_i), \quad \theta_i \sim P, \quad P \sim \mathcal{DP}(\alpha P_0)$$

Suppose we have observed $\theta_1, \ldots, \theta_{i-1}$, then the predictive distribution of θ_i is

$$p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}) = \frac{\alpha}{\alpha + i - 1} P_0(\theta_i) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\theta_j}(\theta_i)$$

This is called "Polya urn predictive rule".

Dirichlet Process Mixtures

$$y_i \sim \mathcal{K}(\theta_i), \quad \theta_i \sim P, \quad P \sim \mathcal{DP}(\alpha P_0)$$

Suppose we have observed $\theta_1, \ldots, \theta_{i-1}$, then the predictive distribution of θ_i is

$$p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}) = \frac{\alpha}{\alpha + i - 1} P_0(\theta_i) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\theta_j}(\theta_i)$$

This is called "Polya urn predictive rule".

Polya urn model:

- We start with a urn with x red balls and y blue balls.
- At each step, we draw a ball from the urn and put it back with an additional ball of the same color.

Dirichlet Process Mixtures

It is also connect to the **Chinese restaurant process**.

Dirichlet Process Mixtures

It is also connect to the **Chinese restaurant process**.

Chinese restaurant process:

- There is a restaurant with infinite tables.

Dirichlet Process Mixtures

It is also connect to the **Chinese restaurant process**.

Chinese restaurant process:

- There is a restaurant with infinite tables.
- The first customer sits at the first table.

Dirichlet Process Mixtures

It is also connect to the **Chinese restaurant process**.

Chinese restaurant process:

- There is a restaurant with infinite tables.
- The first customer sits at the first table.
- The i-th customer sits at the j-th table with probability $\frac{n_j}{\alpha+i-1}$, where n_j is the number of customers at the j-th table.

Dirichlet Process Mixtures

It is also connect to the **Chinese restaurant process**.

Chinese restaurant process:

- There is a restaurant with infinite tables.
- The first customer sits at the first table.
- The i-th customer sits at the j-th table with probability $\frac{n_j}{\alpha+i-1}$, where n_j is the number of customers at the j-th table.
- The i-th customer sits at a new table with probability $\frac{\alpha}{\alpha+i-1}$.

Dirichlet Process Mixtures

It is also connect to the **Chinese restaurant process**.

Chinese restaurant process:

- There is a restaurant with infinite tables.
- The first customer sits at the first table.
- The i-th customer sits at the j-th table with probability $\frac{n_j}{\alpha+i-1}$, where n_j is the number of customers at the j-th table.
- The i-th customer sits at a new table with probability $\frac{\alpha}{\alpha+i-1}$.

- The number of customers at each table follows a Polya urn model.
- The process is exchangeable.

Dirichlet Process Mixtures

The hyperprior for α is often chosen to be a gamma distribution:

$$\alpha \sim \mathrm{Gamma}(a, b)$$

Dirichlet Process Mixtures

The hyperprior for α is often chosen to be a gamma distribution:

$$\alpha \sim \mathrm{Gamma}(a, b)$$

It is usually more difficult to choose the hyperprior for P_0.

Dirichlet Process Mixtures

The hyperprior for α is often chosen to be a gamma distribution:

$$\alpha \sim \text{Gamma}(a, b)$$

It is usually more difficult to choose the hyperprior for P_0.

Then the model is a hierarchical model. The posterior distribution of α can be derived through MCMC.

- See marginal Gibbs sampling and block Gibbs sampling in textbook Ch. 23.3.