

# Bayesian Nonparametric Models

Peter Orbanz, Cambridge University  
Yee Whye Teh, University College London

**Related keywords:** Bayesian Methods, Prior Probabilities, Dirichlet Process, Gaussian Processes.

## Definition

A Bayesian nonparametric model is a Bayesian model on an infinite-dimensional parameter space. The parameter space is typically chosen as the set of all possible solutions for a given learning problem. For example, in a regression problem the parameter space can be the set of continuous functions, and in a density estimation problem the space can consist of all densities. A Bayesian nonparametric model uses only a finite subset of the available parameter dimensions to explain a finite sample of observations, with the set of dimensions chosen depending on the sample, such that the effective complexity of the model (as measured by the number of dimensions used) adapts to the data. Classical adaptive problems, such as nonparametric estimation and model selection, can thus be formulated as Bayesian inference problems. Popular examples of Bayesian nonparametric models include Gaussian process regression, in which the correlation structure is refined with growing sample size, and Dirichlet process mixture models for clustering, which adapt the number of clusters to the complexity of the data. Bayesian nonparametric models have recently been applied to a variety of machine learning problems, including regression, classification, clustering, latent variable modeling, sequential modeling, image segmentation, source separation and grammar induction.

## Motivation and Background

Most of machine learning is concerned with learning an appropriate set of parameters within a model class from training data. The meta level problems of determining appropriate model classes are referred to as model selection or model adaptation. These constitute important concerns for machine learning practitioners, chiefly for avoidance of over-fitting and under-fitting, but also for discovery of the causes and structures underlying data. Examples of model selection and adaptation include: selecting the number of clusters in a clustering problem, the number of hidden states in a hidden Markov model, the number

of latent variables in a latent variable model, or the complexity of features used in nonlinear regression.

*Nonparametric models* constitute an approach to model selection and adaptation, where the sizes of models are allowed to grow with data size. This is as opposed to *parametric models* which uses a fixed number of parameters. For example, a parametric approach to density estimation would be to fit a Gaussian or a mixture of a fixed number of Gaussians by maximum likelihood. A nonparametric approach would be a Parzen window estimator, which centers a Gaussian at each observation (and hence uses one mean parameter per observation). Another example is the support vector machine with a Gaussian kernel. The representer theorem shows that the decision function is a linear combination of Gaussian radial basis functions centered at every input vector, and thus has a complexity that grows with more observations. Nonparametric methods have long been popular in classical (non-Bayesian) statistics [1]. They often perform impressively in applications and, though theoretical results for such models are typically harder to prove than for parametric models, appealing theoretical properties have been established for a wide range of models.

Bayesian nonparametric methods provide a Bayesian framework for model selection and adaptation using nonparametric models. A Bayesian formulation of nonparametric problems is nontrivial, since a Bayesian model defines prior and posterior distributions on a single fixed parameter space, but the dimension of the parameter space in a nonparametric approach should change with sample size. The Bayesian nonparametric solution to this problem is to use an infinite-dimensional parameter space, and to invoke only a finite subset of the available parameters on any given finite data set. This subset generally grows with the data set. In the context of Bayesian nonparametric models, “infinite-dimensional” can therefore be interpreted as “of finite but unbounded dimension”. More precisely, a Bayesian nonparametric model is a model that (1) constitutes a Bayesian model on an infinite-dimensional parameter space and (2) can be evaluated on a finite sample in a manner that uses only a finite subset of the available parameters to explain the sample.

We make the above description more concrete in the next section when we describe a number of standard machine learning problems and the corresponding Bayesian nonparametric solutions. As we will see, the parameter space in (1) typically consists of functions or of measures, while (2) is usually achieved by marginalizing out surplus dimensions over the prior. Random functions and measures, and more generally probability distributions on infinite-dimensional random objects, are called *stochastic processes*; examples we will encounter include Gaussian processes, Dirichlet processes and beta processes. Bayesian nonparametric models are often named after the stochastic processes they contain. The examples are then followed by theoretical considerations, including formal constructions and representations of the stochastic processes used in Bayesian nonparametric models, exchangeability, and issues of consistency and convergence rate. We conclude this article with future directions and a reading list.

## Examples

**Clustering with mixture models.** Bayesian nonparametric generalizations of finite mixture models provide an approach for estimating both the number of components in a mixture model and the parameters of the individual mixture components simultaneously from data. Finite mixture models define a density function over data items  $x$  of the form  $p(x) = \sum_{k=1}^K \pi_k p(x|\theta_k)$ , where  $\pi_k$  is the mixing proportion and  $\theta_k$  are parameters associated with component  $k$ . The density can be written in a non-standard manner as an integral:  $p(x) = \int p(x|\theta) G(\theta) d\theta$ , where  $G = \sum_{k=1}^K \pi_k \delta_{\theta_k}$  is a discrete mixing distribution encapsulating all the parameters of the mixture model and  $\delta_\theta$  is a Dirac distribution (atom) centered at  $\theta$ . Bayesian nonparametric mixtures use mixing distributions consisting of a *countably infinite* number of atoms instead:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}. \quad (1)$$

This gives rise to mixture models with an infinite number of components. When applied to a finite training set, only a finite (but varying) number of components will be used to model the data, since each data item is associated with exactly one component but each component can be associated with multiple data items. Inference in the model then automatically recovers both the number of components to use and the parameters of the components. Being Bayesian, we need a prior over the mixing distribution  $G$ , and the most common prior to use is a *Dirichlet process* (DP). The resulting mixture model is called a DP mixture.

Formally, a Dirichlet process  $DP(\alpha, H)$  parametrized by a concentration parameter  $\alpha > 0$  and a base distribution  $H$  is a prior over distributions (probability measures)  $G$  such that, for any finite partition  $A_1, \dots, A_m$  of the parameter space, the induced random vector  $(G(A_1), \dots, G(A_m))$  is Dirichlet distributed with parameters  $(\alpha H(A_1), \dots, \alpha H(A_m))$  (see the theory section for a discussion of subtleties involved in this definition). It can be shown that draws from a DP will be discrete distributions as given in (1). The DP also induces a distribution over partitions of integers called the *Chinese restaurant process* (CRP), which directly describes the prior over how data items are clustered under the DP mixture. For more details on the DP and the CRP, see DP entry [?].

**Nonlinear regression.** The aim of regression is to infer a continuous function from a training set consisting of input-output pairs  $\{(t_i, x_i)\}_{i=1}^n$ . Parametric approaches parametrize the function using a finite number of parameters and attempt to infer these parameters from data. The prototypical Bayesian nonparametric approach to this problem is to define a prior distribution over continuous functions directly by means of a *Gaussian process* (GP). As explained in GP entry [?], a GP is a distribution on an infinite collection of random variables  $X_t$ , such that the joint distribution of each finite subset  $X_{t_1}, \dots, X_{t_m}$  is a multivariate Gaussian. A value  $x_t$  taken by the variable  $X_t$  can be regarded as the value of a continuous function  $f$  at  $t$ , that is,  $f(t) = x_t$ . Given the training set, the Gaussian process posterior is again a distribution on functions, conditional

on these functions taking values  $f(t_1) = x_1, \dots, f(t_n) = x_n$ .

**Latent feature models.** Latent feature models represent a set of objects in terms of a set of latent features, each of which represents an independent degree of variation exhibited by the data. Such a representation of data is sometimes referred to as a distributed representation. In analogy to nonparametric mixture models with an unknown number of clusters, a Bayesian nonparametric approach to latent feature modeling allows for an unknown number of latent features. The stochastic processes involved here are known as the *Indian buffet process* (IBP) and the *beta process* (BP). Draws from BPs are random discrete measures, where each of an infinite number of atoms has a mass in  $(0, 1)$  but the masses of atoms need not sum to 1. Each atom corresponds to a feature, with the mass corresponding to the probability that the feature is present for an object. We can visualize the occurrences of features among objects using a binary matrix, where the  $(i, k)$  entry is 1 if object  $i$  has feature  $k$  and 0 otherwise. The distribution over binary matrices induced by the BP is called the IBP.

**Hidden Markov models.** Hidden Markov models (HMMs) are popular models for sequential or temporal data, where each time step is associated with a state, with state transitions dependent on the previous state. An *infinite HMM* is a Bayesian nonparametric approach to HMMs, where the number of states is unbounded and allowed to grow with the sequence length. It is defined using one DP prior for the transition probabilities going out from each state. To ensure that the set of states reachable from each outgoing state is the same, the base distributions of the DPs are shared and given a DP prior recursively. The construction is called a *hierarchical Dirichlet process* (HDP); see below.

**Density estimation.** A nonparametric Bayesian approach to density estimation requires a prior on densities or distributions. However, the DP is not useful in this context, since it generates discrete distributions. A useful density estimator should smooth the empirical density (such as a Parzen window estimator), which requires a prior that can generate smooth distributions. Priors applicable in density estimation problems include DP mixture models and Pólya trees.

DP mixture models: Since the mixing distribution in the DP mixture is random, the induced density  $p(x)$  is random thus the DP mixture can be used as a prior over densities. Despite the fact that these are now primarily used in machine learning as clustering models, they were in fact originally proposed for density estimation.

*Pólya Trees* are priors on probability distributions that can generate both discrete and piecewise continuous distributions, depending on the choice of parameters. Pólya trees are defined by a recursive infinitely deep binary subdivision of the domain of the generated random measure. Each subdivision is associated with a beta random variable which describes the relative amount of mass on each side of the subdivision. The DP is a special case of a Pólya tree corresponding to a particular parametrization. For other parametrizations the resulting random distribution can be smooth so is suitable for density estimation.

**Power-law Phenomena.** Many naturally occurring phenomena exhibit power-

law behavior. Examples include natural languages, images and social and genetic networks. An interesting generalization of the DP, called the *Pitman-Yor process*,  $\text{PYP}(\alpha, d, H)$ , has recently been successfully used as models of such power-law data. The Pitman-Yor process augments the DP by a third parameter  $d \in [0, 1)$ . When  $d = 0$  the PYP is a  $\text{DP}(\alpha, H)$ , while when  $\alpha = 0$  it is a so called *normalized stable process*.

**Sequential modeling.** HMMs model sequential data using latent variables representing the underlying state of the system, and assuming that each state only depends on the previous state (the so called Markov property). In some applications, for example language modeling and text compression, we are interested in directly modeling sequences without using latent variables, and without making any Markov assumptions, i.e. modeling each observation conditional on all previous observations in the sequence. Since the set of potential sequences of previous observations is unbounded, this calls for nonparametric models. A *hierarchical Pitman-Yor process* can be used to construct a Bayesian nonparametric solution whereby the conditional probabilities are coupled hierarchically. **Dependent and hierarchical models.** Most of the Bayesian nonparametric models described above are applied in settings where observations are homogeneous or exchangeable. In many real world settings observations are often not homogeneous, in fact they are often structured in interesting ways. For example, the data generating process might change over time thus observations at different times are not exchangeable, or observations might come in distinct groups with those in the same group being more similar than across groups.

Significant recent efforts in Bayesian nonparametrics research have been placed in developing extensions that can handle these non-homogeneous settings. Dependent Dirichlet processes are stochastic processes, typically over a spatial or temporal domain, which define a Dirichlet process (or a related random measure) at each point with neighboring DPs being more dependent. These are used for spatial modeling, nonparametric regression, as well as for modeling temporal changes. Alternatively, hierarchical Bayesian nonparametric models like the hierarchical DP aim to couple multiple Bayesian nonparametric models within a hierarchical Bayesian framework. The idea is to allow sharing of statistical strength across multiple groups of observations. Among other applications, these have been used in the infinite HMM, topic modeling, language modeling, word segmentation, image segmentation and grammar induction. For an overview of various dependent Bayesian nonparametric models and their applications in biostatistics please consult [2]. [3] is an overview of hierarchical Bayesian nonparametric models as well as a variety of applications in machine learning.

## Theory

As we saw in the preceding examples, Bayesian nonparametric models often make use of priors over functions and measures. Because these spaces typically have an uncountable number of dimensions, extra care has to be taken to define

the priors properly and to study the asymptotic properties of estimation in the resulting models. In this section we give an overview of the basic concepts involved in the theory of Bayesian nonparametric models. We start with a discussion of the importance of exchangeability in Bayesian parametric and nonparametric statistics. This is followed by representations of the priors and issues of convergence.

### Exchangeability

The underlying assumption of all Bayesian methods is that the parameter specifying the observation model is a random variable. This assumption is subject to much criticism, and at the heart of the Bayesian versus non-Bayesian debate that has long divided the statistics community. However, there is a very general type of observations for which the existence of such a random variable can be derived mathematically: For so-called *exchangeable* observations, the Bayesian assumption that a randomly distributed parameter exists is not a modeling assumption, but a mathematical consequence of the data's properties.

Formally, a sequence of variables  $X_1, X_2, \dots, X_n$  over the same probability space  $(\mathcal{X}, \Omega)$  is *exchangeable* if their joint distribution is invariant to permuting the variables. That is, if  $P$  is the joint distribution and  $\sigma$  any permutation of  $\{1, \dots, n\}$ , then

$$P(X_1 = x_1, X_2 = x_2 \dots X_n = x_n) = P(X_1 = x_{\sigma(1)}, X_2 = x_{\sigma(2)} \dots X_n = x_{\sigma(n)}) \quad (2)$$

An infinite sequence  $X_1, X_2, \dots$  is *infinitely exchangeable* if  $X_1, \dots, X_n$  is exchangeable for *every*  $n \geq 1$ . In this paper we will mean infinite exchangeability whenever we write exchangeability. Exchangeability reflects the assumption that the variables do not depend on their indices although they may be dependent among themselves. This is typically a reasonable assumption in machine learning and statistical applications, even if the variables are not themselves iid (independently and identically distributed). Exchangeability is a much weaker assumption than iid; iid variables are automatically exchangeable.

If  $\theta$  parametrizes the underlying distribution, and one assumes a prior distribution over  $\theta$ , then the resulting marginal distribution over  $X_1, X_2, \dots$  with  $\theta$  marginalized out will still be exchangeable. A fundamental result credited to de Finetti [4] states that the converse is also true. That is, if  $X_1, X_2, \dots$  is (infinitely) exchangeable, then there is a random  $\theta$  such that:

$$P(X_1, \dots, X_n) = \int P(\theta) \prod_{i=1}^n P(X_i | \theta) d\theta \quad (3)$$

for every  $n \geq 1$ . In other words, the seemingly innocuous assumption of exchangeability automatically implies the existence of a hierarchical Bayesian model with  $\theta$  being the random latent parameter. This the crux of the fundamental importance of exchangeability to Bayesian statistics.

In de Finetti's Theorem it is important to stress that  $\theta$  can be infinite dimensional (it is typically a random measure), thus the hierarchical Bayesian model

(3) is typically a nonparametric one. For example, the Blackwell-MacQueen urn scheme (related to the CRP) is exchangeable thus implicitly defines a random measure, namely the DP (see the DP entry [?] for more details). In this sense, we will see below that de Finetti's Theorem is an alternative route to Kolmogorov's Extension Theorem, which implicitly defines the stochastic processes underlying Bayesian nonparametric models.

## Model Representations

In finite dimensions, a probability model is usually defined by a density function or probability mass function. In infinite-dimensional spaces, this approach is not generally feasible, for reasons explained below. To define or work with a Bayesian nonparametric model, we have to choose alternative mathematical representations.

**Weak Distributions.** A weak distribution is a representation for the distribution of a stochastic process, that is, for a probability distribution on an infinite-dimensional sample space. If we assume that the dimensions of the space are indexed by  $t \in T$ , the stochastic process can be regarded as the joint distribution  $P$  of an infinite set of random variables  $\{X_t\}_{t \in T}$ . For any finite subset  $S \subset T$  of dimensions, the joint distribution  $P_S$  of the corresponding subset  $\{X_t\}_{t \in S}$  of random variables is a finite-dimensional marginal of  $P$ . The *weak distribution* of a stochastic process is the set of all its finite-dimensional marginals, that is, the set  $\{P_S : S \subset T, |S| < \infty\}$ . For example, the customary definition of the Gaussian process as an infinite collection of random variables, each finite subset of which has a joint Gaussian distribution, is an example of a weak distribution representation. In contrast to the explicit representations to be described below, this representation is generally not generative, because it represents the distribution rather than a random draw, but is more widely applicable.

Apparently, just defining a weak distribution in this manner need not imply that it is a valid representation of a stochastic process. A given collection of finite-dimensional distributions represents a stochastic process only (1) if a process with these distributions as its marginals actually exists, and (2) if it is uniquely defined by the marginals. The mathematical result which guarantees that weak distribution representations are valid is the *Kolmogorov Extension Theorem* (also known as the Daniell-Kolmogorov theorem or the Kolmogorov Consistency Theorem). Suppose that a collection  $\{P_S : S \subset T, |S| < \infty\}$  of distributions is given. If all distributions in the collection are marginals of each other, that is, if  $P_{S_1}$  is a marginal of  $P_{S_2}$  whenever  $S_1 \subset S_2$ , the set of distributions is called a *projective family*. The Kolmogorov Extension Theorem states that, if the set  $T$  is countable, and if the distributions  $P_S$  form a projective family, then there exists a uniquely defined stochastic process with the collection  $\{P_S\}$  as its marginal distributions. In other words, any projective family for a countable set  $T$  of dimensions is the weak distribution of a stochastic process. Conversely, any stochastic process can be represented in this manner, by computing its set of finite-dimensional marginals.

The weak distribution representation assumes that all individual random variable  $X_t$  of the stochastic process take values in the same sample space  $\Omega$ . The stochastic process  $P$  defined by the weak distribution is then a probability distribution on the sample space  $\Omega^T$ , which can be interpreted as the set of all function  $f : T \rightarrow \Omega$ . For example, to construct a GP we might choose  $T = \mathbb{Q}$  and  $\Omega = \mathbb{R}$  to obtain real-valued functions on the countable space of rational numbers. Since  $\mathbb{Q}$  is dense in  $\mathbb{R}$ , the function  $f$  can then be extended to all of  $\mathbb{R}$  by continuity. To define the DP as a distribution over probability measures on  $\mathbb{R}$ , we note that a probability measure is a set function that maps “random events”, i.e. elements of the Borel  $\sigma$ -algebra  $\mathcal{B}(\mathbb{R})$  of  $\mathbb{R}$ , into probabilities in  $[0, 1]$ . We could therefore choose a weak distribution consisting of Dirichlet distributions, and set  $T = \mathcal{B}(\mathbb{R})$  and  $\Omega = [0, 1]$ . However, this approach raises a new problem because the set  $\mathcal{B}(\mathbb{R})$  is not countable. As in the GP, we can first define the DP on a countable “base” for  $\mathcal{B}(\mathbb{R})$  then extend to all random events by continuity of measures. More precise descriptions are unfortunately beyond the scope of this entry.

**Explicit Representations.** Explicit representations directly describe a random draw from a stochastic process, rather than describing its distribution. A prominent example of an explicit representation is the so-called *stick-breaking representation* of the Dirichlet process. The discrete random measure  $G$  in (1) is completely determined by the two infinite sequences  $\{\pi_k\}_{k \in \mathbb{N}}$  and  $\{\theta_k\}_{k \in \mathbb{N}}$ . The stick-breaking representation of the DP generates these two sequences by drawing  $\theta_k \sim H$  iid and  $v_k \sim \text{Beta}(1, \alpha)$  for  $k = 1, 2, \dots$ . The coefficients  $\pi_k$  are then computed as  $\pi_k = v_k \prod_{j=1}^{k-1} (1 - v_j)$ . The measure  $G$  so obtained can be shown to be distributed according to a  $\text{DP}(\alpha, G_0)$ . Similar representations can be derived for the Pitman-Yor process and the beta process as well. Explicit representations, if they exist for a given model, are typically of great practical importance for the derivation of algorithms.

**Implicit Representations.** A third representation of infinite dimensional models is based on de Finetti’s Theorem. Any exchangeable sequence  $X_1, \dots, X_n$  uniquely defines a stochastic process  $\theta$ , called the de Finetti measure, making the  $X_i$ ’s iid. If the  $X_i$ ’s are sufficient to define the rest of the model and their conditional distributions are easily specified, then it is sufficient to work directly with the  $X_i$ ’s and have the underlying stochastic process implicitly defined. Examples include the Chinese restaurant process (an exchangeable distribution over partitions) with the DP as the de Finetti measure, and the Indian buffet process (an exchangeable distribution over binary matrices) with the BP being the corresponding de Finetti measure. These implicit representations are useful in practice as they can lead to simple and efficient inference algorithms.

**Finite Representations.** A fourth representation of Bayesian nonparametric models is as the infinite limit of finite (parametric) Bayesian models. For example, DP mixtures can be derived as the infinite limit of finite mixture models with particular Dirichlet priors on mixing proportions, GPs can be derived as the infinite limit of particular Bayesian regression models with Gaussian priors, while BPs can be derived as from the limit of an infinite number of indepen-



dent beta variables. These representations are sometimes more intuitive for practitioners familiar with parametric models. However not all Bayesian nonparametric models can be expressed in this fashion, and they do not necessarily make clear the mathematical subtleties involved.

### Consistency and Convergence Rates

Recent work in mathematical statistics examines the convergence properties of Bayesian nonparametric models, and in particular the questions of *consistency* and *convergence rates*.

Intuitively, a *consistent* estimation procedure is a method that will result in a correct estimate if it has access to an infinite amount of data, that is, a procedure that will be accurate unless it is hampered by insufficient sample size. In Bayesian statistics, an estimate is a distribution over possible parameter values (the posterior), and hence consistency in Bayesian models is defined by demanding that the posterior converges to a delta peak at the true parameter. A classic theorem by J. L. Doob shows that for any Bayesian model, the set of all possible true parameter values for which the model will be consistent has probability one under the prior. In other words, if the true parameter (and hence the data distribution) is chosen at random from the prior, we will always end up with a consistent model, and a Bayesian who is certain about the prior need not worry about inconsistency. This result does not hold anymore if the true parameter value is not in the domain of the prior. In nonparametric models, which have to spread out their probability mass over an infinite-dimensional space, this can result in seemingly unreasonable behavior of the model. For example, the Dirichlet process may be used as a prior in density estimation. Its support consists of the discrete distributions, which means that if the data is drawn from any smooth distribution, the true model is not in the support of the prior (so Doob's theorem does not apply). As shown by Diaconis and Freedman [5], the posterior will not necessarily concentrate in a region close to the true model. Intuitively, the implication is that we cannot generally assume in nonparametric models that the effect of the prior will eventually become negligible if only we see enough data. However, this does not mean that nonparametric Bayesian models are generally inconsistent: A large and growing literature in mathematical statistics shows that consistency can be guaranteed by proper choice of an adequate nonparametric model [6].

Recent results, notably by van der Vaart and Ghosal, apply modern methods of mathematical statistics to study the convergence properties of Bayesian nonparametric models (see [6] for further references). Consistency has been established for a number of models, including Gaussian processes and Dirichlet process mixtures. A particularly interesting aspect of this line of work are results on convergence rates, which specify how rapidly the posterior concentrates with growing sample size, depending on the complexity of the model and on how much probability mass the prior places around the true solution. To make such results quantitative requires a measure for the complexity of a Bayesian nonparametric model. This is done by means of complexity measures developed in

empirical process theory and statistical learning theory, such as metric entropies, covering numbers and bracketing, some of which are well-known in theoretical machine learning. Examples of such results include the consistency of Dirichlet process mixture models for density estimation if both the target density and the parametric mixture components are smooth, and a range of consistency results for regression and density estimation with Gaussian processes. For all of these results, convergence rates can be specified as well (references are given in [6]). A large class of infinite-dimensional models which do behave well even if the true parameter is not in the domain of the prior is identified in [7]. In this case, the posterior will concentrate in the region of the prior support which is closest to the true parameter in a Kullback-Leibler sense.

## Inference

There are two aspects to inference in Bayesian nonparametric models: the analytic tractability of posteriors for the stochastic processes embedded in Bayesian nonparametric models, and practical inference algorithms for the overall models. Bayesian nonparametric models typically include stochastic processes such as the Gaussian process and the Dirichlet process. These processes have an infinite number of dimensions thus naïve algorithmic approaches to computing posteriors is generally infeasible. Fortunately, these processes typically have analytically tractable posteriors, so all but finitely many of the dimensions can be analytically integrated out efficiently. The remaining dimensions, along with the parametric parts of the models, can then be handled by the usual inference techniques employed in parametric Bayesian modeling, including Markov chain Monte Carlo, sequential Monte Carlo, variational inference, and message-passing algorithms like expectation propagation. The precise choice of approximations to use will depend on the specific models under consideration, with speed/accuracy trade-offs between different techniques generally following those for parametric models. In the following, we will give two examples to illustrate the above points, and discuss a few theoretical issues associated with the analytic tractability of stochastic processes.

### Examples

In Gaussian process regression, we model the relationship between an input  $x$  and an output  $y$  using a function  $f$ , so that  $y \sim f(x) + \epsilon$  where  $\epsilon$  is iid Gaussian noise. Given a GP prior over  $f$  and a finite training data set  $\{(x_i, y_i)\}_{i=1}^n$  we wish to compute the posterior over  $f$ . Here we can use the weak representation of  $f$  and note that  $\{f(x_i)\}_{i=1}^n$  is simply a finite-dimensional Gaussian with mean and covariance given by the mean and covariance functions of the GP. Inference for  $\{f(x_i)\}_{i=1}^n$  is then straightforward. The approach can be thought of equivalently as marginalizing out the whole function except its values on the training inputs. Note that although we only have the posterior over  $\{f(x_i)\}_{i=1}^n$ , this is sufficient to reconstruct the function evaluated at any other point  $x_0$  (say

the test input), since  $f(x_0)$  is Gaussian and independent of the training data  $\{(x_i, y_i)\}_{i=1}^n$  given  $\{f(x_i)\}_{i=1}^n$ . In GP regression the posterior over  $\{f(x_i)\}_{i=1}^n$  can be computed exactly. In GP classification or other regression settings with nonlinear likelihood functions, the typical approach is to use sparse methods based on variational approximations or expectation propagation; see GP entry [?] for details.

Our second example involves Dirichlet process mixture models. Recall that the DP induces a clustering structure on the data items. If our training set consists of  $n$  data items, since each item can only belong to one cluster, there are at most  $n$  clusters represented in the training set. Even though the DP mixture itself has an infinite number of potential clusters, all but finitely many of these are not associated with data, thus the associated variables need not be explicitly represented at all. This can be understood either as marginalizing out these variables, or as an implicit representation which can be made explicit whenever required by sampling from the prior. This idea is applicable for DP mixtures using both the Chinese restaurant process and the stick-breaking representations. In the CRP representation, each data item  $x_i$  is associated with a cluster index  $z_i$ , and each cluster  $k$  with a parameter  $\theta_k^*$  (these parameters can be marginalized out if  $H$  is conjugate to  $F$ ), and these are the only latent variables that need be represented in memory. In the stick-breaking representation, clusters are ordered by decreasing prior expected size, with cluster  $k$  associated with a parameter  $\theta_k^*$  and a size  $\pi_k$ . Each data item is again associated with a cluster index  $z_i$ , and only the clusters up to  $K = \max(z_1, \dots, z_n)$  need be represented. All clusters with index  $> K$  need not be represented since their posterior conditioning on  $\{(x_i, z_i)\}_{i=1}^n$  is just the prior.

## On Bayes Equations and Conjugacy

It is worth noting that the posterior of a Bayesian model is, in abstract terms, defined as the conditional distribution of the parameter given the data and the hyperparameters, and this definition does not require the existence of a Bayes equation. If a Bayes equation exists for the model, the posterior can equivalently be defined as the left-hand side of the Bayes equation. However for some stochastic processes, notably the DP on an uncountable space such as  $\mathbb{R}$ , it is not possible to define a Bayes equation even though the posterior is still a well-defined mathematical object. Technically speaking, existence of a Bayes equation requires the family of all possible posteriors to be dominated by the prior, but this is not the case for the DP. That posteriors of these stochastic processes can be evaluated at all is solely due to the fact that they admit an analytic representation.

The particular form of tractability exhibited by many stochastic processes in the literature is that of a *conjugate* posterior, that is, the posterior belongs to the same model family as the prior, and the posterior parameters can be computed as a function of the prior hyperparameters and the observed data. For example, the posterior of a  $\text{DP}(\alpha, G_0)$  under observations  $\theta_1, \dots, \theta_n$  is again a Dirichlet process,  $\text{DP}(\alpha + n, \frac{1}{\alpha + n}(\alpha G_0 + \sum \delta_{\theta_i}))$ . Similarly the posterior of a GP under

observations of  $f(x_1), \dots, f(x_n)$  is still a GP. It is this conjugacy that allows practical inference in the examples above. A Bayesian nonparametric model is conjugate if and only if the elements of its weak distribution, i.e. its finite-dimensional marginals, have a conjugate structure as well [8]. In particular, this characterizes a class of conjugate Bayesian nonparametric models whose weak distributions consist of exponential family models. Note however that lack of conjugacy do not imply intractable posteriors. An example is given by the Pitman-Yor process, where the posterior is given by a sum of a finite number of atoms and a Pitman-Yor process independent from the atoms.

## Future Directions

Since MCMC sampling algorithms for Dirichlet process mixtures became available in the 1990s and made latent variable models with nonparametric Bayesian components applicable to practical problems, the development of Bayesian nonparametrics has experienced explosive growth [9, 10]. Arguably, though, the results available so far have only scratched the surface. The repertoire of available models is still mostly limited to using the Gaussian process, the Dirichlet process, the beta process, and generalizations derived from those. In principle, Bayesian nonparametric models may be defined on any infinite-dimensional mathematical object of possible interest to machine learning and statistics. Possible examples are kernels, infinite graphs, special classes of functions (e.g. piecewise continuous or Sobolev functions), and permutations.

Aside from the obvious modeling questions, two major future directions are to make Bayesian nonparametric methods available to a larger audience of researchers and practitioners through the development of software packages, and to understand and quantify the theoretical properties of available methods.

### General-Purpose Software Package

There is currently significant growth in the application of Bayesian nonparametric models across a variety of application domains both in machine learning and in statistics. However significant hurdles still exist, especially the expense and expertise needed to develop computer programs for inference in these complex models. One future direction is thus the development of software packages that can compile efficient inference algorithms automatically given model specifications, thus allowing a much wider range of modeler to make use of these models. Current developments include the R DPpackage<sup>1</sup>, the hierarchical Bayesian compiler<sup>2</sup>, adaptor grammars<sup>3</sup>, the MIT-Church project<sup>4</sup>, as well as efforts to add Bayesian nonparametric models to the repertoire of current Bayesian modeling

<sup>1</sup><http://cran.r-project.org/web/packages/DPpackage>

<sup>2</sup><http://www.cs.utah.edu/~hal/HBC>

<sup>3</sup><http://www.cog.brown.edu/~mj/Software.htm>

<sup>4</sup><http://projects.csail.mit.edu/church/wiki/Church>

environments like OpenBugs<sup>5</sup> and infer.NET<sup>6</sup>.

### Statistical Properties of Models

Recent work in mathematical statistics provides some insight into the quantitative behavior of Bayesian nonparametric models (cf theory section). The elegant, methodical approach underlying these results, which quantifies model complexity by means of empirical process theory and then derives convergence rates as a function of the complexity, should be applicable to a wide range of models. So far, however, only results for Gaussian processes and Dirichlet process mixtures have been proven, and it will be of great interest to establish properties for other priors. Some models developed in machine learning, such as the infinite HMM, may pose new challenges to theoretical methodology, since their study will probably have to draw on both the theory of algorithms and mathematical statistics. Once a wider range of results is available, they may in turn serve to guide the development of new models, if it is possible to establish how different methods of model construction affect the statistical properties of the constructed model.

### Cross Reference

Dirichlet Processes, Gaussian Processes, Bayesian Methods, Prior Probabilities.

### Recommended Reading

In addition to the references embedded in the text above, we recommend the books [11, 12] and the review articles [13, 14] on Bayesian nonparametrics. References for DPs and DP mixture models can be found in the DP entry [?], for GPs in the GP entry [?], while for most of the other examples can be found in the chapter [3] of the book [11].

- [1] L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006.
- [2] D. B. Dunson. Nonparametric Bayes applications to biostatistics. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press, 2010.
- [3] Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press, 2010.
- [4] B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Accademia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Matematiche e Naturale*, 4, 1931.

---

<sup>5</sup><http://mathstat.helsinki.fi/openbugs>

<sup>6</sup><http://research.microsoft.com/en-us/um/cambridge/projects/infernet>

- [5] P. Diaconis and D. Freedman. On the consistency of Bayes estimates (with discussion). *Annals of Statistics*, 14(1):1–67, 1986.
- [6] S. Ghosal. The Dirichlet process, related priors and posterior asymptotics. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press, 2010.
- [7] B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, 34:837–877, 2006.
- [8] P. Orbanz. Construction of nonparametric Bayesian models from parametric Bayes equations. In *Advances in Neural Information Processing Systems*, 2010.
- [9] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.
- [10] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.
- [11] N. Hjort, C. Holmes, P. Müller, and S. Walker, editors. *Bayesian Nonparametrics*. Number 28 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.
- [12] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer, 2002.
- [13] S. G. Walker, P. Damien, P. W. Laud, and A. F. M. Smith. Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society*, 61(3):485–527, 1999.
- [14] P. Müller and F. A. Quintana. Nonparametric Bayesian data analysis. *Statistical Science*, 19(1):95–110, 2004.