Chapter 1

Introduction

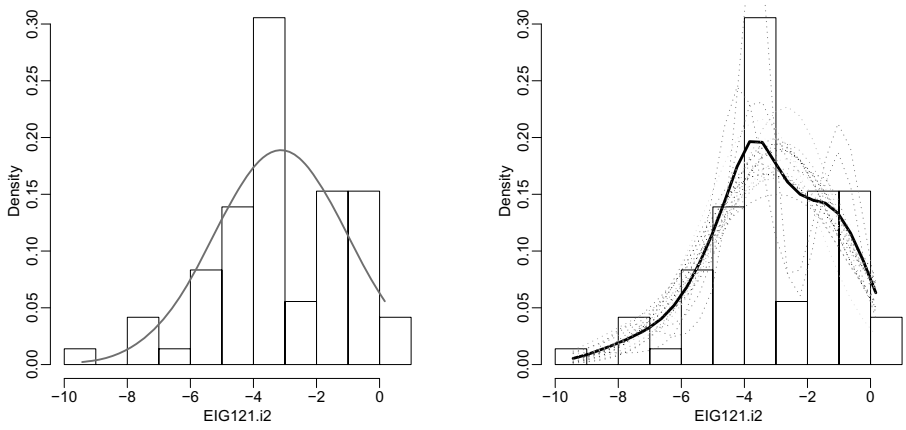
1.1. Definition

All models are wrong, but some are useful. Many statisticians know and appreciate G.E.P. Box's comment on statistical modeling (Box, 1979). Often the choice of the final inference model is a compromise of an accurate representation of the experimental conditions, a preference for parsimony and the need for a practicable implementation. However, these competing goals are not always honestly spelled out, and the resulting uncertainties are not fully described.

Over the last 20 years a powerful inference approach that allows us to mitigate some of these limitations has become increasingly popular. Bayesian nonparametric (BNP) inference allows us to acknowledge uncertainy about an assumed model while maintaining a practically feasible inference approach. We could take this feature as a pragmatic characterization of BNP as flexible prior probability models that generalize traditional models by allowing for positive prior probability for a very wide range of alternative models, while centering the prior around a parsimonious traditional model. A more formal definition of BNP is as probability models on infinite dimensional parameter spaces, such as functional spaces.

Example 1 (Density estimation) *Consider a simple random sample $y_i \sim F$ i.i.d., $i = 1, \ldots, n$, from some unknown distribution F. Bayesian inference requires that the model be completed with a prior for the unknown F in the sampling model. One could proceed by restricting F to a normal location family, $F = \mathsf{N}(\theta, 1)$. The model F is indexed by a finite dimensional parameter vector θ and the model is completed with a prior probability model for the finite dimensional θ. We are back to parametric Bayesian inference. Figure 1.1a shows the resulting inference conditional on an assumed random sample $\boldsymbol{y}$. Naturally, inference about the unknown F is restricted to the assumed normal location family and does not allow for multimodality or skewness. In contrast, a BNP model would proceed with a prior probability model $p(F)$ for the unknown distribution. Figure 1.1b contrasts the parametric inference with the flexible BNP inference under a Dirichlet process mixture prior.*

In Example 1 the infinite dimensional random quantity is an unknown distribution. Alternatively, the infinite dimensional quantity might be the unknown mean function $f(\cdot)$ in a regression problem, a response surface, a spectral density, or perhaps an autoregressive mean function in a nonparametric time series model. In the rest of these notes we will mostly focus on problems where the infinite dimensional quantity is an unknown probability measure $F(\cdot)$, as in example 1. The reason for this focus is simply tradition; most BNP models in the recent literature consider random probability measures.



(a) Parametric model (b) Nonparametric model

FIG 1.1. *Example 1. Inference on the unknown distribution F under a parametric model and nonparametric model. The histogram are the observed data $y_i \sim F$.*

1.2. BNP Models for Random Probability Measures

Figure 1.2 summarizes in a stylized diagram the relationships between some of the most popular BNP models for random probability measures (RPM). The diagram highlights the central role of the popular Dirichlet process (DP) model, which arises as a special case of several other BNP models. The diagram serves as a short outline of these notes. After a brief introductory definition of the models in the rest of this Introduction we will in the following chapters discuss some of the models in more detail.

1.2.1. Species Sampling Models

Species sampling models (SSMs) define an RPM $p(G)$ indirectly, by considering a predictive rule for $x_{n+1} \mid x_1, \ldots, x_n$ in a random sample $x_i \sim G$. For a discrete probability measure G, random sampling includes a positive probability for ties among the x_i. We use groups of tied sequence elements x_i to define clusters. These clusters will play a prominent role in the upcoming discussion. It is helpful to introduce some related notation. Let k_n denote the number of unique values ("species") among $(x_1, \ldots, x_n)$, let $x_j^\star$, $j = 1, \ldots, k_n$, denote the unique values, and let n_{nj} denote the number of x_i equal to the j-th unique value $x_j^\star$. Finally, $\boldsymbol{n} = \{n_{n1}, \ldots, n_{nk_n}\}$ characterizes the cluster sizes of the partition created by the ties. We drop the subindex $_n$ when the sample size n is understood from the context.

Definition 1 (Pitman, 1996) *An exchangeable sequence of r.v.'s $x_1, x_2, \ldots$ is a* species sampling sequence *(SSS) if $x_1 \sim G_0$ where G_0 is a non-atomic measure and*

$$x_{n+1} \mid x_1, \ldots, x_n \sim \sum_{j=1}^{k_n} p_j(\boldsymbol{n}_n)\delta_{x_j^\star} + p_{k+1}(\boldsymbol{n}_n)G_0, \tag{1.1}$$

where $p_j(\boldsymbol{n}_n) \geq 0$ and $\sum_{j=1}^{k_n+1} p_j(\boldsymbol{n}_n) = 1$.

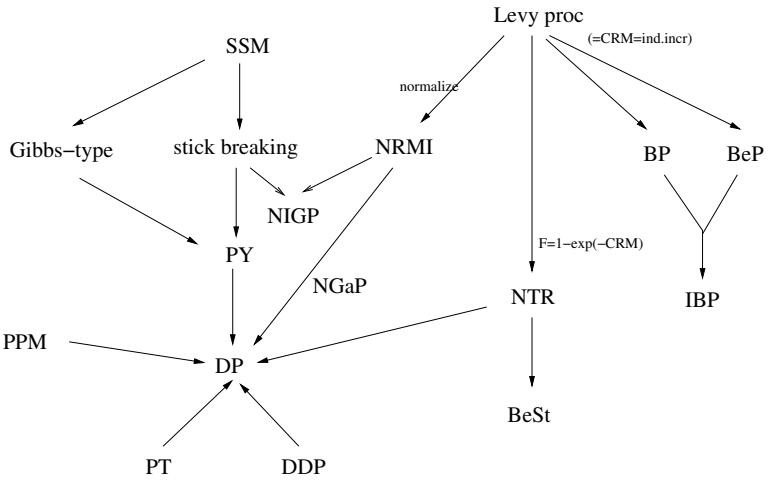


FIG 1.2. *Popular BNP models. An arrow from model A to B indicates that B is a special case (or variation) of A. Details of the models are discussed in the text. The graph includes species sampling models (SSM), the Pitman-Yor process (PY), the Dirichlet process (DP), the product partition model (PPM), Pólya trees (PT), dependent DP (DDP), normalized random measures with independent increments (NRMI), the normalized gamma process (NGaP), the normalized inverse Gaussian process (NIGP), neutral to the right processes (NTR), the Beta-Stacey process (BeSt), the beta process (BP), the Bernoulli process (BeP) and the Indian buffet process (IBP). The annotations are not exhaustive. For example, the NIGP is not the only NRMI that defines a SSM.*

The sequence of weights $\{p_j(\boldsymbol{n})\}$ is known as predictive probability function (PPF). Any SSS can be characterized by the PPF $\{p_j(\boldsymbol{n})\}$ and G_0. The opposite is not true. The critical property is the exchangeability of the sequence. Not every family of weights with $\sum_{j=1}^{k} p_j(\cdot) = 1$ and $p_j(\cdot) \geq 0$ characterizes an SSS because for an arbitrary choice of $\{p_j(\cdot)\}$ the implied sequence x_i might not be exchangeable.

At this moment the reader might wonder how the SSS defines a prior probability model for an unknown probability measure. The SSS defines a random probability measures as the de Finetti measure in the corresponding representation of the exchangeable sequence as a hierarchical model.

Theorem 1.2.1 (Pitman, 1996) *(x_i) is an SSS if and only if $x_i \sim G$, i.i.d., for*

$$G(\cdot) = \sum_{h=1}^{\infty} p_h \delta_{\widetilde{x}_h}(\cdot) + RG_0, \tag{1.2}$$

for some sequence of positive random variables (p_h) and R such that $1 - R = \sum_{i=h}^{\infty} p_h \leq 1$, $(\widetilde{x}_h)$ is a random sample from G_0, and $\{p_h\}$ and $\{\widetilde{x}_h\}$ are mutually independent.

In other words, a constructive definition of SSMs is possible as a discrete RPM with point masses at i.i.d. locations $\widetilde{x}_h$ and an arbitrary probability model for the weights p_h, subject only to $\sum p_h \leq 1$. Unless otherwise stated we will always assume $R = 0$. In contrast to the alternative definition through a PPF, any choice of G_0 and (p_h) will do. Exchangeability of the sequence x_i is already ensured by construction.

The Pitman-Yor (PY) process (Ishwaran and James, 2001; Pitman, 1995; Pitman and Yor, 1997) is a SSM with PPF

$$p_j(\boldsymbol{n}) \propto \begin{cases} (b+ka) & j = k+1 \\ n_{nj} - a & j = 1, \ldots, k, \end{cases} \tag{1.3}$$

for $0 \leq a < 1$ and $b > 0$. It is also known as the two-parameter Poisson Dirichlet process. Exploiting the construction of a SSM as a discrete RPM we can characterize the PY by a base measure G_0 and the law for the weights p_h in (1.2). The distribution of the weights p_h for the PY process can be described by a sequence of independent Beta random variables

$$p_h = v_h \prod_{l<h}(1-v_l), \qquad v_h \sim \mathsf{Beta}(1-a, b+ha), \tag{1.4}$$

for $h = 1, 2, \ldots$. We write $\mathsf{PY}(a, b, G_0)$ for a PY random probability measure with the joint distribution of the weights indexed by (a, b) and the locations of the point masses generated as i.i.d. sample from a base measure G_0.

Ishwaran and James (2001) refer to (1.4) as a stick breaking construction. The name arises from picturing (1.4) as repeated breaking of a stick of initial length 1. The first weight $p_1 = v_1$ is a beta random fraction of the stick, p_2 is a beta random fraction of the remaining stick of length $(1-p_1)$, etc. Sethurman (1994) introduced the construction for the special case of $a = 0$, which defines a Dirichlet process (DP), $\mathsf{DP}(b, G_0)$. Here we encounter for the first time the DP model. The characterization of the DP as a SSM is one of its many alternative defining properties. One of the reasons for the wide use of the DP prior is the simple form of the implied PPF (1.1). Assume $x_i \sim G$, i.i.d. and $G \sim \mathsf{DP}(M, G_0)$. Then

$$x_{n+1} \mid x_1, \ldots, x_n \sim \begin{cases} \delta_{x_j^\star} & \text{with prob } \propto n_{nj} \\ G_0 & \text{with prob} \propto M, \end{cases} \tag{1.5}$$

i.e., (1.3) with $a = 0$ and $b = M$.

The simple form of (1.5) greatly simplifies posterior simulations when the BNP model is used for statistical inference. Indeed, exchangeability of the x_i implies that the same rule applies for the complete conditional probability $p(x_i \mid x_\ell, \ell \neq i)$. We will come back to the DP several times in the following review before we discuss it in more detail in Chapter 3.

Gnedin and Pitman (2006) and Lijoi *et al.* (2007b) define another special case of SSMs. They consider Gibbs type priors that are characterized by a PPF

$$p_j(\boldsymbol{n}) \propto \begin{cases} V_{n+1,k} \frac{n_{nj}-\sigma}{n} & j = 1, \ldots, k \\ V_{n+1,k+1} & j = k+1, \end{cases}$$

with $\{V_{n,k},\ k \leq n\}$ a sequence of coefficients with $V_{1,1} = 1$ and subject to $V_{n,k} = V_{n+1,k+1} + (n - k\sigma)V_{n+1,k}$, and $0 \leq \sigma < 1$. The model defines a variation of PY priors. Conditional on $k_{n+1} = k_n$ the conditional PPF remains the same as under a PY prior. Only the probability of a new species, i.e., $p(k_{n+1} = k_n + 1 \mid \boldsymbol{n}_n)$, changes.

1.2.2. Stick Breaking Prior

One of the characteristics of the SSM construction is the unlimited flexibility in defining the joint distribution of the weights p_h. Ishwaran and James (2001) exploit

this flexibility to propose stick breaking priors for RPMs by generalizing the beta distribution of the fractions v_h in the construction of the PY process.

They propose two generalizations. First, they allow the number of non-zero weights to be finite,

$$G(\cdot) = \sum_{h=1}^{H} p_h \delta_{\widetilde{x}_h}(\cdot)$$

for $H \leq \infty$. Second, the beta prior for the fractions v_h is replaced by $v_h \sim \mathsf{Beta}(a_h, b_h)$, independently, $h = 1, \ldots, H-1$. For $H < \infty$ we add $v_H = 1.0$ to ensure $\sum p_h = 1.0$. The locations of the point masses remain unchanged as $\widetilde{x}_h \sim G_0$, i.i.d.

Naturally the DP remains a special case, with $a_h = 0$, $b_h = b$ and $H = \infty$. Ishwaran and James (2001) propose the model $a_h = 0, b_h = b$ and $H < \infty$ as a natural simplification of the DP prior. We refer to it as the finite DP, $\mathsf{DP}_H(b, G_0)$. An alternative version of the truncated DP prior is the ϵ-DP of Muliere and Tardella (1998), see §3.4.

1.2.3. Product Partition Models

While not strictly a prior for a random probability measure, we include the product partition model (PPM) in this review because of the close connection with popular BNP models for random probability measures. We have already seen how the random clustering that is defined by the ties in an i.i.d. sample from a discrete distribution G can be useful to characterize a discrete random probability measures $G \sim p(G)$. In many applications of BNP models the investigators are not primarily interested in the random probability measure G itself, but rather focus on the induced clustering. It is therefore useful to consider probability models for random cluster arrangements.

We need a minimum of notation. Let $S = \{1, 2, \ldots, n\}$ index a set of experimental units. A partition or cluster arrangement of S is a family of subsets $\rho_n = \{S_1, \ldots, S_k\}$ with $\bigcup S_j = S$ and $S_{j_1} \cap S_{j_2} = \emptyset$ for $j_1 \neq j_2$. When ρ_n is treated as a random quantity we have a random partition $p(\rho_n)$. For example, any discrete probability model G implies a random partition $p(\rho_n)$ by grouping random samples $x_i \sim G$, $i = 1, \ldots, n$, by unique values, as $S_j = \{i : \ x_i = x_j^\star\}$. Here $x_j^\star$ denotes the j-th unique value. The $x_j^\star$ are indexed by order of appearance. The random partition $p(\rho_n)$ is determined by the probability masses in G. The same remains true when G is an unknown discrete random probability measure with prior $p(G)$, but this is not the only interesting class of random partition models $p(\rho_n)$.

Hartigan (1990) introduces the product partition models (PPM). A random partition $p(\rho_n)$ is called a product partition model if it can be written as a product

$$p(\rho_n) = \prod_{j=1}^{k} c(S_j)$$

of factors that depend on S_j only, $j = 1, \ldots, k$. Let y_i denote an outcome for the i-th experimental unit, let $y_j^\star = \{y_i;\ i \in S_j\}$ denote the outcomes arranged by clusters and let $\boldsymbol{y} = (y_1, \ldots, y_n)$ denote the entire data. The PPM combines the prior $p(\rho_n)$ with a sampling model $p(\boldsymbol{y} \mid \rho_n)$ that factors similarly and assumes

exchangeability within each cluster

$$p(\boldsymbol{y} \mid \rho_n) = \prod_{j=1}^{k} p(y_j^\star)$$

for an exchangeable model $p(y_j^\star)$.

Again we run into the DP model as a special case. The random partition induced by the ties in a random sample $x_i \sim G$ with DP prior $G \sim \mathsf{DP}(M, G_0)$ forms a PPM with

$$p(\rho_n) \propto \prod_{j=1}^{k_n} M\,\Gamma(n_{nj}).$$

Recall that $n_{nj} = |S_j|$ is the size of the j-th cluster.

1.2.4. Pólya Trees

Essentially, the Pólya tree (PT) model defines a RPM G as a random histogram. The bins are created by nested partitions of the desired sample space B. The random probabilities for each bin are products of (independent) conditional probabilities of each layer of the nested partition sequence.

Figure 1.3 illustrates the construction. The bins B_ϵ are indexed by binary se-

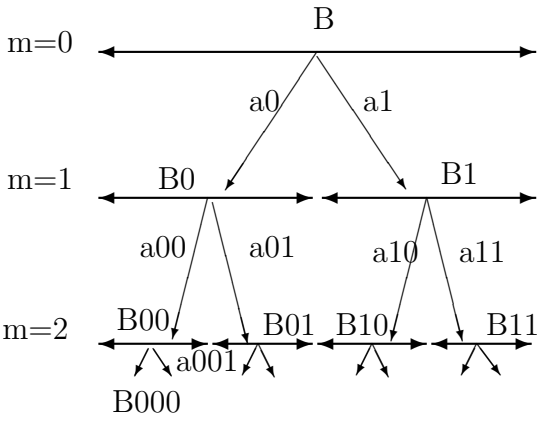


FIG 1.3. *PT prior for an RPM G. At level m of the nested partition sequence the sample space B is partitioned into $\{B_\epsilon\}$ indexed by binary sequences $\epsilon = \epsilon_1 \cdots \epsilon_m$ and defined by repeated splits into $B_\epsilon = B_{\epsilon 0} \cup B_{\epsilon 1}$.*

quences $\epsilon = e_1 \cdots e_m$ with $e_j \in \{0, 1\}$. The bins are created by nested partitions of the desired sample space B into $B = B_0 \cup B_1$, $B_0 = B_{00} \cup B_{01}$, etc. The random probabilities $G(B_\epsilon)$ are defined by the (independent) conditional probabilities $G(B_{e_1 \cdots e_j 0} \mid B_{e_1 \cdots e_j})$. Let $\epsilon = e_1 \cdots e_{j-1}$ and let $Y_{\epsilon 0} = G(B_{\epsilon 0 j} \mid B_\epsilon)$. The PT prior characterizes $p(G)$ as a prior probability model for all $Y_{\epsilon 0}$. It defines $p(G)$ by assuming

$$Y_{\epsilon 0} \sim \mathsf{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$$

independently across ϵ and $Y_{\epsilon 1} = 1 - Y_{\epsilon 0}$. In short, the definition of an RPM is reduced to independent beta priors for the conditional probabilities in the nested

partition sequence. The PT is indexed by the partition sequence $\mathcal{B} = \{B_\epsilon\}$ and the set of beta parameters, $\mathcal{A} = \{\alpha_\epsilon\}$. We write $G \sim \mathsf{PT}(\mathcal{A}, \mathcal{B})$.

Again the DP arises as a special case when $\alpha_\epsilon = \alpha_{\epsilon 0} + \alpha_{\epsilon 1}$. For example, $\alpha_{\epsilon_1 \cdots \epsilon_m} = c2^{-m}$ implies a $\mathsf{DP}(c, G_0)$ with G_0 defined as the distribution with dyadic quantiles given by B_ϵ. Thus $G_0(B_0) = 0.5$, $G_0(B_{01} = 0.25)$, etc.

1.2.5. DDP

Many applications call for more than one random probability measure G. For example, the generic regression problem of predicting an outcome y conditional on a covariate x could be described as inference for the conditional distributions $G_x(\cdot) = p(y_i \mid x_i = x)$ for $x \in X$. When $p(y_i \mid x_i = x)$ is indexed by finitely many parameters we are back to parametric, possibly non-linear regression. However, when the investigator is unwilling or unable to restrict $p(y_i \mid x_i)$ to a parametric family, then the problem becomes one of inference for a family of random probability measures $\mathcal{G} = \{G_x,\ x \in X\}$, indexed by the covariates x. We thus need a BNP prior $p(G_x;\ x \in X)$ for the entire family. In the application to nonparametric regression as well as many other applications it is natural to require that G_x be dependent across x. Surely we would not expect G_x to change substantially for minor changes of x.

MacEachern (1999) introduced the class of dependent DP (DDP) priors to construct such prior models $p(\mathcal{G})$. In particular, the marginal distribution $p(G_x)$ remains a DP prior, $G_x \sim \mathsf{DP}(c, G_{0x})$. But the model allows for the desired dependence. Recall that the DP prior is a special case of a SSM. A DP random measure can therefore be written as an infinite discrete probabilty measure with independent locations for the point masses

$$G_x(\cdot) = \sum_h p_h \delta_{\widetilde{x}_{hx}}(\cdot). \tag{1.6}$$

We will discuss the DDP prior in more detail in Chapter 5. One important feature is the independence of the point mass locations $\widetilde{x}_{hx}$ across h. The DDP construction leaves the independence across h untouched, but adds dependence for $\widetilde{x}_{hx}$ across x to induce the desired dependence of G_x across x. The notation in (1.6) implies the use of common weights p_h across all G_x. This variation of DDP models is known as the common weight DDP. In general, the weights could have an additional $_x$ index, defining $G_x(\cdot) = \sum p_{hx} \delta_{\widetilde{x}_{hx}}(\cdot)$.

1.2.6. Completely Random Measures

A rich variety of BNP models are based on completely random measures (CRM) (Kingman, 1993). A random measure μ is a CRM when $\mu(B_1)$ and $\mu(B_2)$ are independent for any two non-overlapping measurable sets B_1, B_2 of some space X. The independence property implies in particular that μ can not be a probability measure, lest the restriction to total mass 1.0 induces dependence between $\mu(B_1)$ and $\mu(B_2)$.

As a consequence of the desired independence a CRM must always be discrete, i.e., it can be written as a sum of point masses. An alternative construction of CRMs that will turn out to be useful in the upcoming discussion is based on a

Poisson process. Let $N(\cdot)$ denote a Poisson process on $X \times R^+$ with intensity $\nu(\cdot)$. Then

$$\mu(A) \equiv \int_A \int_{R^+} sN(dx, ds)$$

for measurable $A \subset X$. In words, each point (x, s) of the Poisson process in $X \times R^+$ defines a locations x and weight s for a point mass of μ. The intensity $\nu(\cdot)$ is known as Levy intensity, which also features in the Levy-Khintchine representation for μ. For $X = R$, $\mu_x \equiv \mu((-\infty, x])$ is also known as increasing additive process or independent increments process.

CRMs are useful tools to define BNP priors for random probability measures. The simplest construction is to normalize a CMR to define

$$G \equiv \mu/\mu(X).$$

Regazzini *et al.* (2003) introduce such random probability measures as normalized random measures with independent increments (NRMI).

Here we run again into the DP prior. The original discussion of the DP in Ferguson (1973) discusses as an alternative defining property of the DP the construction as an NRMI, using a normalized version of a gamma process. The definition of the DP as a normalized gamma process immediately implies another useful characterization. Let $\boldsymbol{w} \sim \mathsf{Dir}(a_1, \ldots, a_k)$ denote a Dirichlet distribution for a random vector of weights $\boldsymbol{w}$. Recall that a Dirichlet random vector can be generated by normalized gamma random variables, as $w_i = x_i/(\sum_j x_j)$ for $x_i \sim \mathsf{Gamma}(\alpha_i, \theta)$, i.i.d. Let $\{A_1, \ldots, A_k\}$ denote a partition of the sample space. The nature of the DP as a normalized gamma process implies $(G(A_1), \ldots, G(A_k)) \sim \mathsf{Dir}(a_1, \ldots, a_k)$ with $a_j = \alpha G_0(A_j)$.

There is at least one other NRMI model that allows a similarly simple characterization. Lijoi *et al.* (2005) introduce the normalized inverse Gaussian process (NIGP) as an NRMI. Alternatively, the NIGP can be defined by the following property. For any partition $(A_1, \ldots, A_k)$ of the sample space,

$$(G(A_1), \ldots, G(A_k)) \sim \mathsf{NIG}(MG_0(A_1), \ldots, MG_0(A_k)),$$

where $\mathsf{NIG}(a_1, \ldots, a_k)$ denotes a normalized inverse Gaussian distribution. The NIG distribution is a parametric probability model for a (finite) vector of weights that add up to 1. The definition starts with the inverse Gaussian distribution, $\mathsf{IG}(\alpha, \gamma)$ with p.d.f.

$$p(x) \propto x^{-3/2} e^{-\frac{1}{2}\left(\frac{\alpha^2}{x} + \gamma^2 x\right) + \gamma\alpha},$$

where $x \geq 0$ and $\alpha > 0$. Now, let $x_j \sim \mathsf{IG}(a_j, 1)$, $j = 1, \ldots, k$, denote k independent inverse Gaussian random variables. The NIG is the distribution of the normalized values $w_j = x_j/\sum x_\ell$. We say $\boldsymbol{w} \equiv (w_1, \ldots, w_k) \sim \mathsf{NIG}(a_1, \ldots, a_k)$. Despite the name, the inverse Gaussian has no obvious relation with the normal distribution.

1.2.7. NTR Priors

Normalization is not the only mechanism to construct nonparametric Bayes models from CRMs. Another popular class of models based on CRMs are neutral to the right (NTR) priors. NTR priors are nonparametric priors for random distributions on the real line. Typical applications are to modeling event time distributions in survival analysis.

The defining property of NTR models are independent normalized increments. An RPM G is NTR if the normalized increments

$$G(\,(t_{i-1}, t_i]\,)/G(\,(t_{i-1}, \infty)\,),$$

$i = 1, \ldots, M$, are independent for any $t_0 < t_1 \ldots < t_M$. Docksum (1974, Theorem 3.1) shows that G is NTR if and only if its distribution function can be written as $1-\exp(-\mu(\infty, t])$ for some CRM μ on the real line with $\lim_{t\to\infty} \mu(0, t] = \infty$ almost surely. The DP is again a special case; Doksum (1974) shows that the DP is NTR and gives the specific CRM μ that defines the DP as NTR prior.

1.2.8. Indian Buffet Process

The Indian buffet process (IBP) defines a random binary matrix $\boldsymbol{Z}$, not (naturally) a random probability measure. The name is explained by the analogy to a large buffet in an Indian restaurant. Customers arrive at the buffet to select dishes. Let $Z_{ik} \in \{0, 1\}$ denote an indicator for the i-th customer selecting the k-th dish. The first customer selects a number k_1 of dishes. We index the dishes in the sequence of first selection. Thus the first customer, $i = 1$, selects dishes $k = 1, \ldots, k_1$. This defines $Z_{ik} = 1$, $i = 1$ and $k = 1, \ldots, k_1$. Let $K_i = \sum_{j\le i} k_j$ denote the number of distinct dishes selected by the first i customers. The next, $(i + 1)$-st customer selects or does not select some of the previously selected dishes, defining $Z_{i+1,k} \in \{0, 1\}$, $k = 1, \ldots, K_i$. In addition to dishes selected by previous customers the next customer selects a number k_{i+1} of new dishes $k = K_i + 1, \ldots, K_i + k_{i+1}$, defining $Z_{i+1,k} = 1$. We set $Z_{j,k} = 0$ for earlier customers, $j \le i$.

Let $\boldsymbol{Z}_i = [Z_{jk};\ j = 1, \ldots, i,\ k = 1, \ldots, K_i]$ denote the selections of the first i customers. The random matrix is defined by specifying the probability $p(Z_{i+1,k} = 1 \mid \boldsymbol{Z}_i)$ of the next customer choosing already earlier selected dishes and the distribution of the number of new dishes, $p(k_{i+1})$. Let $m_{-(i+1),k}$ denote the number of customers before $(i + 1)$ selecting dish k. We assume

$$p(Z_{i+1,k} = 1 \mid \boldsymbol{Z}_i) = \frac{m_{-(i+1),k}}{i+1}$$

and $k_{i+1} \sim \mathsf{Poi}\{\alpha/(i + 1)\}$. For n customers the process defines the random binary $(n \times K_n)$ matrix $\boldsymbol{Z}_n$, with a random number of columns K_n, $K_n \sim \mathsf{Poi}\,(\alpha \sum_{i=1}^{n} 1/i)$. In an alternative construction Thibaux and Jordan (2007) show that the IBP can be constructed by a Beta process and a Bernoulli process.

The IBP is useful as a prior model for (possibly overlapping) random subsets. Interpret Z_{ik} as an indicator for experimental unit i being in the $k-$subset. Then $S_k = \{i :\ Z_{ik} = 1\}$ denotes the k-th subset. We see the parallel to the random partition that is defined by, for example, the PPM. While the PPM defines a prior for a random partition $p(S_k;\ k = 1, \ldots, K)$ with $S_{k_1} \cap S_{k_2} = \emptyset$ and $\cup_k S_k = \{1, \ldots, n\}$, the IBP defines a prior $p(S_k;\ k = 1, \ldots, K)$ for a family of subsets with possibly overlapping subsets, and with $\cup_k S_k \subseteq \{1, \ldots, n\}$. In an application the experimental units and subsets could be, for example, proteins and different molecular pathways. A protein i could be part of multiple pathways, i.e., $i \in S_{k_1} \cap S_{k_2}$, and some proteins might not be in any pathway of interest, i.e., $\cup_k S_k \neq \{1, \ldots, n\}$.

1.3. BNP Models for Random Functions

The models that were introduced in §1.2 are priors $p(G)$ for RPMs. Recall the earlier definition of BNP as probability models on infinite dimensonsal spaces. Besides random distributions, another large number of BNP models defines priors $p(f)$ for random functions f. The main motivating applications is to priors for non-linear regression mean functions $f(\cdot)$.

1.3.1. Gaussian Process

We first discuss Gaussian processes as nonparametric priors $p(f)$ for a function $f(\cdot)$ on $\mathbb{X}$. Consider any finite collections of $n \geq 1$ points $x_1, \ldots, x_n \in \mathcal{X}$, and let $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))$. A stochastic process $\{f(x) : x \in \mathbb{X} \subset \mathbb{R}^d\}$ is said to follow a Gaussian process with mean function $g(x)$ and symmetric covariance function $\gamma(x, x')$, denoted $f \sim \mathsf{GP}\{g(x), \gamma(x, x')\}$, if

$$\begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_n) \end{pmatrix}, \begin{bmatrix} \gamma(x_1, x_1) & \gamma(x_1, x_2) & \cdots & \gamma(x_1, x_n) \\ \gamma(x_2, x_1) & \gamma(x_2, x_2) & \cdots & \gamma(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(x_n, x_1) & \gamma(x_n, x_2) & \cdots & \gamma(x_n, x_n) \end{bmatrix} \right) \tag{1.7}$$

or, more succinctly, $\boldsymbol{f} \sim \mathsf{N}(\boldsymbol{g}, \boldsymbol{\Gamma})$. In our notation we distinguish the random function $f(\cdot)$ versus the finite-dimensional vector $\boldsymbol{f}$. The collection of finite-dimensional distributions described above defines a proper stochastic process since it satisfies Kolmogorov's consistency conditions. Indeed, for any collection of measurable sets $A_1, \ldots, A_n$ the joint distribution $\nu_{(x_1,\ldots,x_n)}$ for $(f(x_1), \ldots, f(x_n))$ satisfies

$$\nu_{(x_1,\ldots,x_n)}(A_1, \ldots, A_n) = \nu_{(x_{\pi_1},\ldots,x_{\pi_n})}(A_{\pi_1}, \ldots, A_{\pi_n})$$

for any permutation $\pi_1, \ldots, \pi_n$ of the integers $\{1, \ldots, n\}$ and

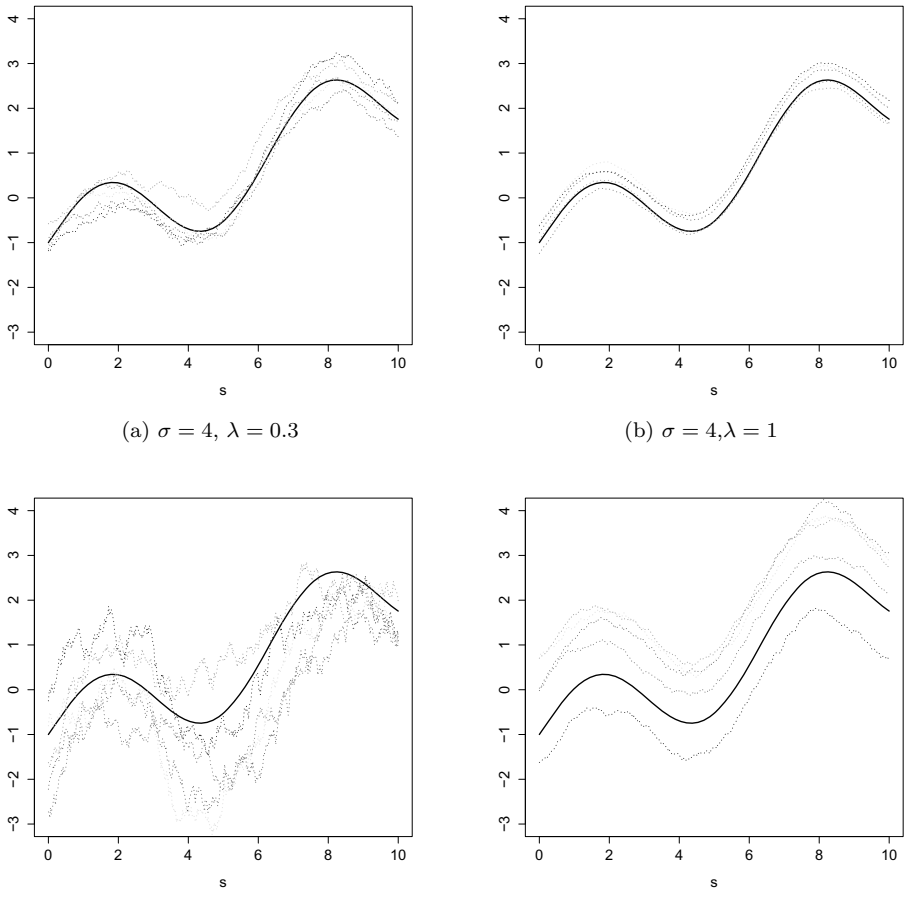
$$\nu_{(x_1,\ldots,x_{n-1},x_n)}(A_1, \ldots, A_{n-1}, \mathbb{R}) = \nu_{(x_1,\ldots,x_{n-1})}(A_1, \ldots, A_{n-1}).$$

Example 2 (Realizations from a Gaussian process) *Consider a Gaussian process on $\mathcal{X} = [0, 10]$ with mean function $g(x) = \sin(x) - \cos(x/4) + 0.15x$ and covariance function $\gamma(x, x') = \sigma^2 \exp\{-|x - x'|/\lambda\}$. For any collection of points, $x_1, \ldots, x_n \in \mathcal{X}$, we can obtain a realization from the stochastic process on these locations by sampling from the multivariate normal distribution (1.7). Figure 1.4 shows realizations from the process on a fine regular grid on $\mathcal{X}$, for different values of σ and λ. The simulations illustrate the effect of these two parameters on the realizations of the process; the range parameter λ controls the local variability, while σ controls the global variability in the realizations.*

In addition to controlling how close the realizations from the process are to the mean function, the covariance function also controls other important properties such as smoothness. For example, the exponential covariance function that we used in example 2 implies that realizations are almost surely not differentiable anywhere, hence the jagged look of the curves. A more detailed discussion of these issues can be found, for example, in Banerjee and Gelfand (2002).

One of the appealing features of the Gaussian process is its tractability. Given the values $\boldsymbol{f}_o = (f(x_1), f(x_2), \ldots, f(x_n))'$, predictions for the value of the function at new levels of the covariates

$$\boldsymbol{f}_p = (f(x_{n+1}), \ldots, f(x_{n+m}))'$$



(a) $\sigma = 4$, $\lambda = 0.3$ (b) $\sigma = 4$,$\lambda = 1$

(c) $\sigma = 100$, $\lambda = 0.3$ (d) $\sigma = 100$, $\lambda = 1$

FIG 1.4. *Realizations from a Gaussian process with mean function* $g(x) = \sin(x) - \cos(x/4) + 0.15x$ *and exponential covariance function* $\gamma(x, x') = \sigma^2 \exp\{-|x - x'|/\lambda\}$ *over a regular grid of 500 points on the interval* $[0, 10]$. *Each panel corresponds to a different combination of the parameters* σ *and* λ. *Within each panel, the solid line shows the mean function* g *and the dotted lines show five different realizations generated under the corresponding values of* λ *and* σ.

can be obtained by noting that the joint distribution for $\boldsymbol{f} = (\boldsymbol{f}_o', \boldsymbol{f}_p')'$

$$\begin{pmatrix} \boldsymbol{f}_o \\ \boldsymbol{f}_p \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} \boldsymbol{g}_o \\ \boldsymbol{g}_p \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Gamma}_{oo} & \boldsymbol{\Gamma}_{op} \\ \boldsymbol{\Gamma}_{po} & \boldsymbol{\Gamma}_{pp} \end{bmatrix} \right),$$

where $\boldsymbol{g}_o$ and $\boldsymbol{g}_p$ denote the marginal means of $\boldsymbol{f}_o$ and $\boldsymbol{f}_p$, $\boldsymbol{\Gamma}_{oo}$, $\boldsymbol{\Gamma}_{pp}$ denote their marginal variance matrices, and $\boldsymbol{\Gamma}_{op} = \boldsymbol{\Gamma}_{po}'$ denotes the cross-covariance matrix. From this, standard results for the normal distribution yield

$$\boldsymbol{f}_p \mid \boldsymbol{f}_o \sim \mathsf{N}\left\{ \boldsymbol{g}_p + \boldsymbol{\Gamma}_{po}\boldsymbol{\Gamma}_{oo}^{-1}(\boldsymbol{f}_o - \boldsymbol{g}_o), \boldsymbol{\Gamma}_{pp} - \boldsymbol{\Gamma}_{po}\boldsymbol{\Gamma}_{oo}^{-1}\boldsymbol{\Gamma}_{op} \right\}.$$

In this way, the predictive distribution $p(\boldsymbol{f}_p \mid \boldsymbol{f}_0)$ is obtained by marginalizing with respect to the random function $f(\cdot)$. In other words, inference can proceed without

the need to record the infinite dimensional $f(\cdot)$. This is critical for practical implementation, computation and extrapolation. The predictive distribution implies that the optimal predictor for $\boldsymbol{f}_p$ under squared-error loss is simply

$$\widehat{\boldsymbol{f}}_p = \boldsymbol{g}_p + \boldsymbol{\Gamma}_{po}\boldsymbol{\Gamma}_{oo}^{-1}(\boldsymbol{f}_o - \boldsymbol{g}_o) \tag{1.8}$$

If $\boldsymbol{f}_o$ is observed, then equation (1.8) can be used to estimate $\boldsymbol{f}_p$ for any value of the covariate x. In that case, $\widehat{f}(x_i) = f(x_i)$ for the covariate values $x_1, \ldots, x_n$ where the function was observed and (1.8) acts as an interpolator at $x_{n+1}, \ldots, x_{n+m}$. However, in practice we often observe $\boldsymbol{f}_o$ only indirectly through some noisy observations $\mathbf{y}_o = (y_1, \ldots, y_n)$ with $E(y_i) = f(x_i)$. If we assume normal residuals we get a hierarchical model

$$\mathbf{y}_o \mid \boldsymbol{f}_o \sim \mathsf{N}(\boldsymbol{f}_o, \tau^2\mathbf{I}), \qquad \boldsymbol{f} \sim \mathsf{N}(\boldsymbol{g}_o, \boldsymbol{\Gamma}_{oo}), \tag{1.9}$$

which implies $\mathbf{y}_o \sim \mathsf{N}(\boldsymbol{g}_o, \boldsymbol{\Gamma}_{oo} + \tau^2\mathbf{I})$. Then, a posteriori,

$$\boldsymbol{f}_o \mid \mathbf{y}_o \sim \mathsf{N}\left\{\left(\boldsymbol{\Gamma}_{oo}^{-1} + \frac{1}{\tau^2}\mathbf{I}\right)^{-1}\left(\boldsymbol{\Gamma}_{oo}^{-1}\boldsymbol{g}_o + \frac{1}{\tau^2}\mathbf{y}_o\right), \left(\boldsymbol{\Gamma}_{oo}^{-1} + \frac{1}{\tau^2}\mathbf{I}\right)^{-1}\right\},$$

and the optimal predictor under squared error loss for $\boldsymbol{f}_p$ is given by

$$\widehat{\boldsymbol{f}}_p = \boldsymbol{g}_p + \boldsymbol{\Gamma}_{po}\boldsymbol{\Gamma}_{oo}^{-1}\left\{\left(\boldsymbol{\Gamma}_{oo}^{-1} + \frac{1}{\tau^2}\mathbf{I}\right)^{-1}\left(\boldsymbol{\Gamma}_{oo}^{-1}\boldsymbol{g}_o + \frac{1}{\tau^2}\mathbf{y}_o\right) - \boldsymbol{g}_o\right\}. \tag{1.10}$$

In this case, the predictor acts as a smoother rather than an interpolator, with the value of τ^2 controlling the level of smoothing. In particular, note that if $\tau^2 \to 0$ then (1.10) converges to (1.8).

An excellent reference on Gaussian process models for regression is Rasmussen and Williams (2006). For applications of Gaussian processes in spatial statistics, see Cressie (1993) and Banerjee *et al.* (2004).

1.3.2. Models Based on Basis Representations

An alternative strategy to create rich models for random functions is to consider representations of the unknown function f in terms of a basis system. This approach reduces the problem of modeling a random function to that of modeling the coefficients associated with the bases.

In the sequel, let $f \in \mathcal{F}$, where $\mathcal{F}$ represents an appropriate function space, and let $(\phi_l(x))$ be a system of basis functions spanning $\mathcal{F}$, implying

$$f(x) = \sum_{\ell=1}^{\infty} \beta_\ell \phi_\ell(x). \tag{1.11}$$

Nonparametric Bayesian inference on f can now be carried out by introducing a prior distribution for coefficients $\boldsymbol{\beta}$. For any $x \in \mathcal{X}$, the optimal estimator under squared error loss is

$$\widehat{f}(x) = \sum_{\ell=1}^{\infty} \mathsf{E}\left(\beta_i \mid \mathbf{y}\right) \phi_\ell(x).$$

A popular example of this approach is wavelet regression. See, for example, Müller and Vidakovic (1999) and Vidakovic (1998).

Let $\mathcal{F} = L^2(\mathbb{R})$ denote the space of square-integrable functions on $\mathbb{R}$. A basis for $\mathcal{F}$ is obtained by translations and dyadic dilations of a mother wavelet $\psi(x) \in L^2(\mathbb{R})$, so that

$$\psi_{j\ell}(x) = 2^{j/2}\psi(2^j x - \ell), \qquad \ell \in \mathbb{Z}, j = 0, \ldots, 2^j - 1,$$

and $(\psi_{j\ell})$ forms an orthonormal basis for L^2. The fact that shifted and scaled versions of $\psi(\cdot)$ form an orthonormal basis is a defining characteristic of a wavelet function. Transformation and dilation of an arbitrary function would not necessarily define an orthonormal basis. Hence, any function $f \in L^2$ can be written as

$$f(x) = \sum_j \sum_\ell \beta_{j\ell} 2^{j/2} \psi(2^j x - \ell). \tag{1.12}$$

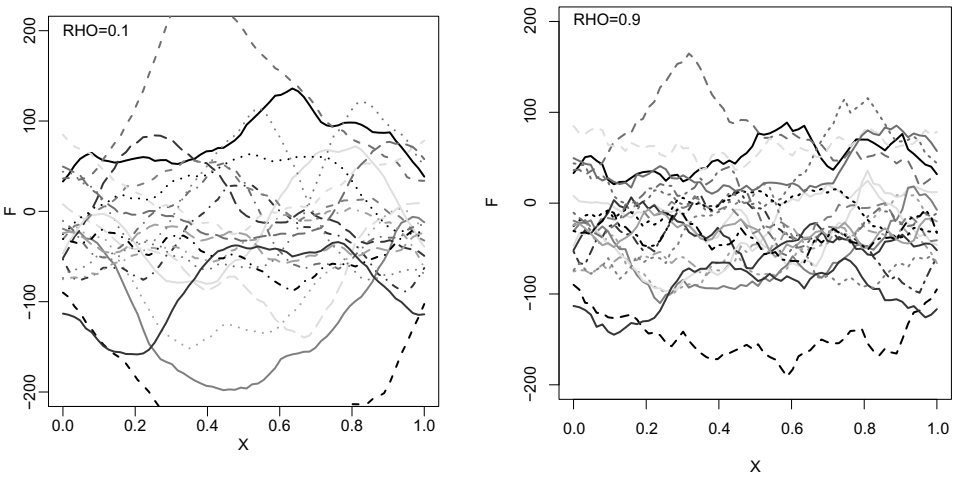
The representation of a function with respect to a wavelet basis can be thought of as a localized version of a Fourier transform. The localization is provided by the index ℓ, while the index j is the level of detail explained by the basis function, with larger values of j corresponding to basis functions explaining higher frequency properties of the function.

Since in practice just a finite number of the coefficients $(\beta_{j\ell})$ can be estimated from a finite sample of size n, the estimation problem is often regularized by assuming that coefficients at the higher levels of details are zero, say for $j \geq \lfloor \log_2 n \rfloor$. In addition, variable selection priors (such as zero inflated Gaussians or double exponential priors) can be used to further reduce the number of coefficients to be estimated.

An important practical feature of the wavelet bases is the existence of a superfast algorithm to carry out the transformation from $\boldsymbol{f}$ to $\boldsymbol{\beta}$ and the reconstruction from $\boldsymbol{\beta}$ to $\boldsymbol{f}$ when $\boldsymbol{f}$ is evaluated over a regular grid. The algorithm is known as the pyramid scheme and allows easy implementation of nonparametric regression if the data are observed on a regular grid. In the absence of a regular grid computation becomes challenging and the practical advantage of wavelet bases fades. An excellent introduction to wavelet appears in Vidakovic (1999).

Example 3 (Wavelet prior for a periodic function.) *Figure 1.5 shows prior simulations for a random function $f \sim p(f)$ with $p(f)$ defined as a prior on wavelet coefficients. The function is defined on $[0, 1]$ and is constrained to $f(0) = f(1)$. The prior is a multivariate normal dependent prior on the wavelet coefficients $\beta_{j\ell}$ in (1.12). We start with a regular grid $\{\ell/n;\ \ell = 0, \ldots, n = 2^J\}$ on $[0, 1]$, and define a multivariate normal prior on $d_\ell = f(\ell/n) - f[(\ell-1)/n]$, $\ell = 1, \ldots, n$. Working with the differences makes it easy to impose the constraint $f(0) = f(1)$. A multivariate normal on $\boldsymbol{d} = (d_1, \ldots, d_n)$ is defined with mean 0 and $\mathsf{Cor}(d_k, d_\ell) = \exp(-\rho|k-\ell|)$. By the pyramid scheme this implies a multivariate normal prior on all $\beta_{j\ell}$. See Berger* et al. *(2012) for details. Modeling in the wavelet domain allows us to later add prior information about the spikiness of the function, formalized by selecting wavelet coefficients $\beta_{j\ell}$ with prior probability $p(\beta_{j\ell} = 0) = 1 - \alpha^j$ (see §2.2.2).*

Models based on basis representations are particularly attractive because model fitting can often be carried out using tools for linear regression. This requires that the basis system is fixed in advance, and that the number of basis functions that are used in the representation is finite. As we discussed above for the case of wavelet bases, this last requirement is often satisfied by truncating the basis system and introducing regularization priors on the remaining coefficients.



(a) $\rho = 0.1$ (b) $\rho = 0.9$

FIG 1.5. *Prior simulation of random curves $f \sim p(f)$ using a dependent prior on the wavelet coefficients in (1.12), subject to $f(0) = f(1)$, with high (panel a) and low (panel b) prior correlation.*

1.3.3. Basis Representation and Gaussian Process Priors

There is a close connection between models based on basis representations and those based on Gaussian process priors. The Kaurhunen-Loève representation theorem (Karhunen, 1947; Lòeve, 1978) states that if f follows a Gaussian process prior with mean $g(x) = 0$ and covariance function $\gamma(x, x')$, then it admits a representation of the form (1.11), where each β_ℓ is independently distributed as a normal random variable and the functions $\{\phi_\ell(x)\}$ are the eigenfunction of the covariance function $\gamma(x, x')$, i.e., they satisfy the integral equations

$$\lambda_k \phi_k^*(x) = \int \gamma(x, x')\phi_k^*(x')dx', \qquad \int \phi_k^*(x)\phi_\ell^*(x)dx = \begin{cases} 1 & k = \ell \\ 0 & k \neq \ell. \end{cases} \tag{1.13}$$

Similar results can be obtained for Gaussian processes with more general mean function $g(x)$ by expanding $g(x)$ in terms of the orthonormal basis functions (ϕ_k). Hence, when we fit a Gaussian process model to data we are implicitly estimating a model that uses an infinite-dimensional basis representation where the basis functions satisfy the constrains in (1.13) and where each β_ℓ is given an independent Gaussian prior.