SPATIAL NONPARAMETRIC BAYESIAN MODELS

Steven N. MacEachern, Athanasios Kottas, and Alan E. Gelfand
Department of Statistics, The Ohio State University, Columbus, OH 43210
Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708
Department of Statistics, University of Connecticut, Storrs, CT 06269

Key Words: Consistency, Dependent Dirichlet Process, Dirichlet Process, Logistic Regression, Overdisperstion, Point Referenced Spatial Data

1 Introduction and Motivation

The prior distribution is an essential ingredient of any Bayesian analysis, and it plays a major role in determining the final results. As such, Bayesians attempt to use prior distributions that have certain properties. Perhaps the main property is a desire to accurately reflect prior information, i.e., information external to the experiment at hand. We would supplement this vague property with a second equally vague property. The posterior distribution should exhibit behavior that is qualitatively acceptable.

The second property for prior distributions is vague, but carries with it several implications. An immediate implication is that we should dispense with parametric Bayesian models in all but the simplest of settings! This perhaps surprising implication follows from an examination of various cases. As a case in point, consider a survival analysis setting where there is a follow-up period of limited duration. With large samples, one could hope to learn the survival distribution over the follow-up period, but there is no hope of learning the exact distribution of survival times beyond the follow-up period. In Bayesian terms, with large samples, the posterior distribution within the follow-up period would be concentrated near the actual survival distribution while the posterior distribution beyond the follow-up period would not concentrate. In the limit, as the experiment tended to one of infinite size, the posterior distribution would ideally tend to a degenerate (and correct) distribution within the follow-up period but would not tend to a degenerate distribution beyond the follow-up period. In a similar fashion, if event times are recorded on a scale of limited precision (for example, in terms of months), one might hope to learn the survival distribution on the monthly scale, but would have no hope of learning the exact distribution on a finer scale. Parametric prior distributions will typically provide degenerate inference over the entire survival distribution and to infinite precision with only limited precision, limited follow-up data. Often, one only needs a discrete observation space consisting of $p+1$ possible values in order to obtain a degenerate posterior distribution in the limit for a p dimensional parametric model.

Survival analysis provides but one example of this general phenomenon. Consideration of such examples, where we desire a non-degenerate limiting posterior, leads us to the belief that, for a Bayesian prior distribution to accurately reflect prior opinion, the prior distribution must be nonparametric.

Conversely, instead of seeking to use prior distributions that do not have degenerate limiting behavior with data of restricted sort, we can seek to use prior distributions that do have partially degenerate limiting behavior. In the survival analysis setting, this would mean that the posterior predictive distribution on the monthly scale, and during the follow-up period, would tend to degeneracy. Furthermore, in this case, we would like the posterior predictive distribution to concentrate at the actual survival distribution. A minimal, though not sufficient, condition for ensuring that the posterior becomes degenerate at an appropriate point is that the prior distribution has full support. Imposing a parametric form on a distribution restricts its support. Again, nonparametric distributions provide a potential solution since they can have full support.

Nonparametric Bayesian methods provide a means of creating prior distributions that accurately reflect prior knowledge in the sense that they satisfy the basic desireable qualitative features of inference. In this work, we provide a framework that encompasses a wide range of nonparametric Bayesian models. The framework naturally suggests new classes of these models which are amenable to simulation based fits with the same technology used to fit finite mixture models and models based on Dirichlet process (DP) priors (reviewed in the next section).

This work was first presented at the ENAR meeting in the spring of 2001. Since this presentation, the authors have become aware of work that generalizes the DP by Hjort (2000) and by Ishwaran and James (2001). There is some overlap in the various generalizations. This talk and a companion talk, which described spatial applications of these models, were presented at the 2001 JSM.

2 The Dirichlet Process

Sethuraman's representation (Sethuraman and Tiwari, 1982; Sethuraman, 1994) of the DP provides a natural starting point for more general nonparametric Bayesian processes. It decomposes a distribution on a r.v. ξ into two parts: the locations and the masses associated with them.

The rule for generation of the locations is to draw a random sample from some distribution. Using θ_i to denote the i^{th} location, we have θ_i i.i.d. F_θ.

The rule for generation of the masses is given in two steps. First, a random sample of beta variates, V_i i.i.d. Beta(1,M), is drawn independently of the θ_i. Here, M is the mass parameter of the DP. Next, these variates are turned into a set of probabilities through the relationship $p_1 = V_1$ and $p_i = V_i \prod_{j<i}(1-V_j)$, for $i = 2, 3, \ldots$.

The final distribution on ξ is (almost surely) discrete. For each measureable set A,

$$P(\xi \in A | \{\theta_i, V_i\}_{i=1}^{\infty}) \quad = \quad \sum_{i:\theta_i \in A} p_i. \qquad (1)$$

Importantly, this construction leads to a more general class of distributions than does Ferguson's (1973) DP definition. With Ferguson's development, the θ_i may be scalars or vectors; with Sethuraman's development, the θ_i may be stochastic processes. As MacEachern (2000b) has shown, letting x represent the index of the real-valued or vector-valued stochastic process, and identifying x with a covariate or covariates, the more general version of the DP provides a collection of distributions, indexed by the covariate. These collections of distributions have many attractive properties and are useful in a wide variety of modeling situations, especially when used as a component in a hierarchical model.

Following the framework that leads to (1), we can describe much more general classes of models. For the purposes of this paper, we will restrict discussion to countable (including finite) mixture models, although there is both motivation and scope for extending the results to models that do not naturally have such a representation. We plan to examine some of these other models in further work.

3 Building blocks

Sethuraman's construction of the DP consists of two building blocks: The distribution for a θ_i, and the distribution for a V_i. These building blocks are coupled with an assumption of independence of all the θ_i and V_i. The remarkably clean construction suggests immediate ways to generalize the DP. A first direction for generalization is to allow the θ_i to be more general than a vector. A second direction is to allow the V_i to come from some distribution other than a beta. This can occur in two ways: the V_i may themselves be stochastic processes, indexed by the same x that indexes θ_i or the V_i or $V_i(x)$ can have other distributions on the unit interval. This is the heart of the dependent Dirichlet process (DDP) (MacEachern, 2000a, b). Through appropriate choice of distributions for θ_i and V_i, one can obtain a wide variety of useful models. In this section, we describe building blocks for a more general process.

3.1 The locations

Suppose that we wish to form a distribution for a random variable taking values in $\mathcal{R}^p$. For models not involving a covariate, our basic requirement would be that F_θ is a distribution on $\mathcal{R}^p$. In the event that we wished to have a family of distributions indexed by a covariate, we would take θ_i to be a stochastic process, indexed by the covariate x. We would require that the range of the covariate be contained in the index set of the stochastic processes. The stochastic process would be vector (of dimension p) valued.

A capsule description of these two models for the locations follows.

- DP. The θ_i are i.i.d. from some distribution F_θ.
- DDP/DP. The θ_i are a random sample of stochastic processes. $\theta_i(x)$ assumes values in $\mathcal{R}^p$ for each x.

3.2 The masses

There is more flexibility in the building blocks for the masses. A primary division is whether the masses are or are not indexed by a covariate. The DDP allows the masses to be indexed by a covariate whereas the DP does not. The V_i must, of course, assume values in the set $[0, 1]$. More properly, for the constructions below, they must have the property that $lim_{n\to\infty} \prod_{i=1}^{n}(1-V_i) = 0$, almost surely. When indexed by x, we require that the above condition either hold for almost all x or for all x.

This division of building blocks for the masses leads to two different sorts of models in the context where there is a covariate x. When V_i does not vary with x (or equivalently, when $V_i(x)$ does not vary with x), the model produces a single mixture distribution. This distribution may be real or vector valued, as has traditionally been the case with the DP. Or, as can happen when the locations are stochastic processes, the distribution may yield an uncountable collection of distributions, with a distribution for each level of the covariate, x. When $V_i(x)$ does vary with x, however, there is no unique representation of the collection of distributions as a single distribution. The model merely specifies a (perhaps uncountable) collection of conditional distributions.

A brief description of building blocks for the masses follows.

- DP (also single-p DDP). The V_i are i.i.d. Beta$(1, M)$.

- Countable mixture. The V_i are i.i.d. from some distribution F_V. F_V assigns all of its mass to the interval $[0, 1)$, and some of its mass to the interval $(0, 1)$.

- Finite mixture. The V_i are i.i.d. from some distribution F_V. F_V assigns all of its mass to the interval $[0, 1]$, and positive mass to the singleton $\{1\}$. The mass at 1 ensures that the mixture will be finite; a consequence of this model is that the number of components in the mixture follows a geometric distribution.

- DDP. The V_i form a random sample of stochastic processes. For each x, $V_i(x) \sim$ Beta$(1, M(x))$. The mass parameters $M(x)$ may vary with x. When the possible values for x have more than countable cardinality, additional conditions are needed to ensure that all of the mass is used for each x.

- Countable mixture analog of the DDP. The V_i form a random sample of stochastic processes. For each x, there is some $F_{V(x)}$ that assigns all of its mass to the interval $[0, 1)$, and some of its mass to the interval $(0, 1)$. $F_{V(x)}$ is not necessarily a beta distribution. Again, additional conditions are required in order to ensure that all of the mass is used for each x.

- Finite mixture analog of the DDP. The V_i form a random sample of stochastic processes. For each x, there is some $F_{V(x)}$ that assigns all of its mass to the interval $[0, 1]$ and positive mass to the singleton $\{1\}$. Again, additional conditions are required in order to ensure that all of the mass is used for each x. Assumptions about the distribution of the process $V(x)$ determine features of the finite mixture. For example, if the path $V_i(x)$ is either entirely in the interval $(0, 1)$ or is identically equal to 1, all of the conditional distributions indexed by x will have the same number of mixture components. For stochastic processes whose paths behave differently, the number of components in the mixture may vary with x. With appropriate conditions on the paths $V_i(x)$, one can guarantee that the number of components in all of the mixture distributions is uniformly bounded.

The assumption that the V_i are i.i.d. from some distribution can obviously be relaxed. Relaxing the assumption of identical distributions provides more flexible tail behavior. For example, the V_i might be independent $Beta(1, M_i)$ variates. Retaining an independence structure for the V_i, at least for large i, is computationally advantageous.

4 More general processes

To create a novel model, we need only ensure that we select a building block for the locations and one for the masses. Existence of a joint probability space is guaranteed by standard results in probability theory. See, for example, Ash's (1972) description of product measure. His description includes a countable or finite number of components. It also applies to components that are stochastic processes as well as real or vector valued random variables.

Once a joint probability space has been defined, one merely needs to check that the purported distributions (conditional on x) are indeed distributions. This leads to a condition on the locations that, for each x, across i, the locations represent the same type of quantity. A set A must be measureable across i in order for one to accumulate the corresponding p_i as in equation (1) above. In typical applications with a covariate, θ_i will be a vector valued stochastic process. A random sample of these locations will automatically satisfy this condition.

The condition on the masses is more difficult to check. Since the $V_i(x)$ are, in all cases, confined to the interval $[0, 1]$, the construction of the $p_i(x)$ ensures that $\sum_{i=1}^{\infty} p_i(x) \leq 1$ for all x. Thus, we have defined a collection of sub-distributions. In order to obtain a collection of distributions, we need to ensure that all of the probability is used for each x. If the covariate space is finite or countable, all of the building blocks described above lead to a valid collection of distributions. That is, since at each level of the covariate the model almost surely defines a distribution, the finite or countable collection of distributions is also almost surely defined. For a covariate space that has larger cardinality, various conditions on the paths of the $V_i(x)$ guarantee the almost sure existence of the entire collection of distributions.

Instead of placing conditions on the paths of stochastic processes, an alternative approach that turns any sub-distribution into a distribution is to add one more component to the mixture. We call this component the null component. Formally, we define one more location, θ_0, that is independent of and that has the same distribution as the other locations. We assign mass $p_0(x) = 1 - \sum_{i=1}^{\infty} p_i(x)$ to this component at covariate level x. Thus, the null component sweeps up all of the left-over mass at whatever values of x are short of mass. This technique is useful for finite mixture models where one might focus on the first several components of the mixture and attempt to sweep the remaining details of a distribution under the rug.

A particularly appealing class of models is one we call *head and tail models.* These models depart from the structure outlined above. We define a finite mixture model, that may depend on a covariate, for the first few components. Hence we have a distribution on the first k locations and masses, θ_i, V_i, $i = 1,...,k$. These parameters need not be independent nor identically distributed. Call this the head of the model. We define a countable (or finite) mixture model for the remaining components by choosing building blocks for the locations and masses described above. Call this the tail of the model.

Head and tail models are attractive because they allow us to address the first property for Bayesian prior distributions: accurate reflection of prior information. A traditional shortcoming in the use of nonparametric Bayesian procedures has been a limitation on the form of information injected into the prior distribution.

As a case in point, we consider problems falling under the heading of classification or cluster analysis. As described above, many nonparametric Bayesian models can be described as mixture distributions, and mixtures lie at the heart of the classification problem. With the traditional formulation, there are a finite number of groups, each with a probability distribution. In many instances, the labels on the classes are meaningful, and there may be substantial information about the parameter values associated with the different classes. The head of the model allows one to use a prior distribution that identifies some components of the mixture and that places informed prior distributions on their parameters. The distributions for these classes may be dependent, for example, they may capture an ordering apparent in previous data sets, or they may arise through combination of estimates based on previous data sets. The head of the model also allows one to express informed opinion as to the relative prevalence of the different classes. When several classes are approximately equal in prevalence, one would write a distribution reflecting this by placing an appropriate joint distribution on V_1,..., V_k which cannot easily be obtained from independent V_i.

The tail of the model ensures, or at least makes it possible, that we satisfy the second property for prior distributions. The tail can guarantee full support for the prior distribution which in turn provides the dual benefits of not forcing degenerate limiting behavior where we don't want it and of making consistent estimation possible.

Head and tail models are computationally attractive. Since the head of the model follows a finite mixture distribution, Markov chain Monte Carlo simulation strategies developed for such distributions enable us to perform all of the requisite conditional generations. The tail matches a nonparametric Bayesian model, and so we make use of computational strategies developed for these models.

5 Consistency in a generalized logistic regression setting

Consider the following extension of the logistic regression model where x_i is a possibly vector-valued covariate,

$$\begin{aligned} \xi_i | x_i &\sim F_{x_i} \\ Y_{ij} | \xi_i &\sim \text{Binomial}(n, \xi_i), \end{aligned}$$

where the usual conventions of conditional independence apply. We supplement the model with an assumption that data are collected at a fixed set of m design points, $x_1, \ldots, x_m$. Assume that the design is balanced and that the replication at each design point tends to ∞.

Suppose that the true sampling distribution of Y_{ij}, conditional on x_i, is given by

$$m_{x_i}(y) = \int \binom{n}{y} \xi^y (1-\xi)^{n-y} dF_{x_i}(\xi),$$

where F_{x_i} is an arbitrary distribution with support on the unit interval. Under this assumption, and an additional assumption that the implied prior distribution on the sampling distribution for Y_{ij} has full support, we conclude that the posterior predictive distribution, $\hat{m}_{x_i}$ at each $x_i \in \{x_1, \ldots, x_m\}$ tends to the sampling distribution at that point, m_{x_i}.

To establish this result, we repeat, with minor modification, an argument that appears, among other places, in MacEachern, Clyde and Liu (1999). The argument turns the problem into a finite dimensional inference problem and then relies on the standard asymptotic argument to show convergence of the posterior (here the posterior predictive) distribution to the sampling distribution.

- There is a one-to-one match between the first n moments of F_{x_i} and probabilities for the $n+1$ possible outcomes for $Y|x_i$. Consequently, the set of m distributions $m_{x_1}, \ldots, m_{x_m}$ is determined by the mn dimensional parameter which consists of the first n moments of Y at each of the m levels of the covariate. Consider the vector $(Y_{1j}, \ldots, Y_{mj})$ to consist of m conditionally independent components.

- The prior distribution on $(F_{x_1}, \ldots, F_{x_m})$ induces a prior distribution on the first n moments of $\xi_{x_1}, \ldots, \xi_{x_m}$. In turn, this induces a distribution on $(Y_{1j}, \ldots, Y_{mj})$.

- A data value, described as $(Y_{1j}, \ldots, Y_{mj})$, can also be viewed as a multinomial observation.

- Asymptotically, the likelihood of the mn dimensional parameter concentrates on the closest parameter values in the support of the prior distribution. Proximity is measured by the Kullback-Liebler divergence.

- Consistency follows from full support for the mn dimesional moment parameter. This full support is a consequence of full support of the distributions $F_{x_1}, \ldots, F_{x_m}$.

To tidy up the details of this argument, some care is needed with the metrics that underlie a description of full support. MacEachern (2000b) contains details of the unusual metric on a collection of distribution functions.

The model that we describe here extends previous Bayesian modelling efforts on logistic regression. Consider the following collection of models. The first model is a straight logistic regression model. With $g(x) = logit(x'\beta)$ for some vector β, we have $Y_{ij}|x_i \sim$ Binomial$(n, g(x'\beta))$. This model is very restrictive. The two main implications are that the covariate must enter the argument for $g(\cdot)$ in a linear fashion and that all of the Bernoulli trials at covariate level x_i are judged to be i.i.d., even across the binomial samples. Such a model has a restrictive mean structure and does not allow the Y_{ij} to exhibit overdispersion. Consequently, it will be inappropriate for many binomial-logistic regression data sets.

The second model remedies the problem of a restrictive mean structure. This can be accomplished by allowing a more general, possibly nonparametric form for $g(\cdot)$, thus including probit and other link functions. Typically, we would retain an assumption of monotoncity of the link function $g(\cdot)$. The restrictions on the mean structure can also be relieved by replacing the linear form for $x'\beta$ with a more flexible form. The two main approaches are through model building–adding more covariates, perhaps through creation of new covariates which are functions of those in the model–and through nonparametric regression. However, these models still imply that $Y_{ij}|x_i$ follows some binomial distribution.

The third model remedies the problem of a fixed dispersion or of a conditional binomial distribution for $Y_{ij}|x_i$. This model adds what is often described as a random effect associated with each sample. Thus, expanding on the first model, $x_i'\beta$ would be replaced by $x_i'\beta + \epsilon_{ij}$, yielding a distribution for $Y_{ij}|x_i$ that is a mixture of binomials. The usual assumption on the ϵ_{ij} is that they are independent and normally distributed with mean 0 and some fixed variance, σ^2. Since $g(\cdot)$ is a nonlinear function, $E[Y_{ij}|x_i, \beta, \sigma^2]$ is a function of σ^2. Thus, the standard way to add overdispersion to the model also perturbs the mean structure.

An alternative formulation of the third model that avoids this problem is to add a stage to the model where $\mu_{ij} \sim$ Beta(α_1, α_2) with the restriction that $\alpha_1/(\alpha_1 + \alpha_2) = g(x'\beta)$. This formulation of the model allows one to add overdispersion directly while not affecting the mean structure of the model. It also has the advantage of matching the beta-binomial model that many would use if many "binomial" samples were collected at a single level of the covariate.

The second and third models are often used separately, though they can be used in conjunction with one another. Mukhopadhyay and Gelfand (1997) have created such a family of nonparametric Bayesian models. They provide a model that incorporates nonparametric components in two places, yielding an arbitrary, monotonic increasing link and also an arbitrary distribution for the random effects. However, their model relies on additive random effects resulting in a restriction on its support.

The models that we describe in this paper represent a natural completion of the modelling strategies described above, allowing us to obtain full support for our prior distribution. To obtain full support, we need to choose appropriate building blocks for the nonparametric prior distribution. The key features in this context are to allow an arbitrarily large number of components in the model–whether this is accomplished by assigning probability 1 to countable mixture distributions or by assigning positive probability to finite mixture distributions with an arbitrarily large number of components–and to allow the θ_i associated with the various components to differ for the different levels of the covariate.

6 DDP modeling for spatial data analysis

Consider point referenced spatial data assumed to form a sample from a realization of a random field $\{Y(s) : s \in D\}$, $D \subseteq R^d$. Denote by $s_1, ..., s_n$ the locations in D where the data $\mathbf{Y}' = (Y(s_1), ..., Y(s_n))$ are collected. Typically, a Gaussian random field is assumed resulting to a multivariate normal specification for $\mathbf{Y}$. Within hierarchical modeling, observations are assumed conditionally independent given model parameters at the first stage, with spatial dependence introduced at the second stage in the prior distribution of the parameters. However, the underlying parametric distributional assumptions result in models that do not have full support and might fail to reveal important features of the data.

A semiparametric model can be developed employing a single-p DDP prior on the random field F_D. Here, the V_i are i.i.d. Beta$(1, M)$ and the locations, $\theta_{i,D} = \{\theta_i(s) : s \in D\}$, are random realizations from some base random field F_{0D} over D. For instance, F_{0D} might be a mean zero stationary Gaussian random field. Attractively, although the prior for F_D is *centered* around a stationary process, it can be shown that random realizations have nonconstant variance and are nonstationary.

To overcome the almost sure discreteness of the prior on F_D we can mix it against a white noise process (nugget process with zero mean and variance τ^2) to create random processes G which have continuous support. More explicitly, if θ_D is a realization from F_D and $\mathbf{Y}_D - \theta_D$ is a realization from the white noise process then marginally $\mathbf{Y}_D$ arises from the process G which can be defined as the convolution

$$G\left(\mathbf{Y}_D \mid F_D, \tau^2\right) = \int \mathcal{K}\left(\mathbf{Y}_D - \theta_D \mid \tau^2\right) F_D\left(d\theta_D\right).$$

Differentiating to densities,

$$g\left(\mathbf{Y}_D \mid F_D, \tau^2\right) = \int k\left(\mathbf{Y}_D - \theta_D \mid \tau^2\right) F_D\left(d\theta_D\right). \tag{2}$$

Here $\mathcal{K}$ is the distribution function and k is the density function of the white noise process.

For the finite set of locations $s_1, ..., s_n$, (2) implies that the joint density of $\mathbf{Y}$ given F_D and τ^2 is almost surely of the form $\sum_{i=1}^{\infty} p_i f_{N_n}(\mathbf{Y} \mid \theta_i, \tau^2 I_n)$, where $\theta_i = (\theta_i(s_1), ..., \theta_i(s_n))$, i.e., a countable location mixture of normals. A constant mean term μ can also be added to the kernel of this mixture. The full Bayesian model is completed with priors on μ, τ^2, the parameters of the covariance matrix of F_{0D} and possibly M. Simulation based model fitting is routine for this single-p DDP model building on existing techniques for DP based models.

An important extension is to spatial-temporal modeling which requires two single-p DDP priors. Even more interesting, but also more challenging at least computationally, is the extension of the models to incorporate general DDP priors so that the V_i also depend on spatial location. The full development of the methodology outlined in this section will be reported elsewhere.

7 References

Ash, R.A. (1972). Real Analysis and Probability. New York: Academic Press.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. Ann.Statist., **1**, 209–230.

Hjort, N.L. (2000). Bayesian analysis for a generalised Dirichlet process prior. Technical Report, University of Oslo.

Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick stick-breaking priors. J. Amer. Statist. Assoc., **96**, 161–173.

MacEachern, S.N. (2000a). Decision theoretic aspects of dependent nonparametric processes. To appear in the Proceedings of ISBA 2000.

MacEachern, S.N. (2000b). Dependent Dirichlet processes. Unpublished Manuscript, Department of Statistics, The Ohio State University.

MacEachern, S.N., Clyde, M.A. and Liu, J. (1999). Sequential importance sampling for nonparametric Bayes models: the next generation. Can. J. Statist., **27**, 251–267.

Mukhopadhyay, S. and Gelfand, A.E. (1997). Dirichlet process mixed generalized linear models. J. Amer. Statist. Assoc., **92**, 633–639.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statist. Sinica, **4**, 639–650.

Sethuraman, J. and Tiwari, R.C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Statist. Decis. Theory Related Topics III*, vol. 2, 305–315. New York: Academic.