DEPENDENT NONPARAMETRIC PROCESSES

Steven N. MacEachern, The Ohio State University
Department of Statistics, 404 Cockins Hall, 1958 Neil Ave., Columbus, OH 43210
(snm@stat.ohio-state.edu)

Key Words: Dirichlet process, model adequacy, nonparametric Bayes, random effects, similarity.

1 Introduction and Motivation

The impetus behind the development of the models presented in this work has been of both practical and philosophical nature. From a practical viewpoint, the models are motivated by a common theme that I have noticed across a range of data sets: the standard assumptions that are used in the linear model are inappropriate.

A classic example is the modelling of weight as a function of some covariate, often height in adult populations or age for child populations. For data sets that I have seen collected on college student populations, the main features appear to be nonlinearity of the mean for weight given height, right skewness for the distribution of weight at a given height, heteroscedasticity, and a bumpiness in the distribution, particularly in the upper tail. These departures from the standard assumptions of a (perhaps linear) trend for the relationship between height and weight coupled with independent and identically distributed errors or errors that follow a scale family render any fit based on these assumptions suspect.

The problems with violations of the assumptions do not appear to be solved through conventional transformations of explanatory, response or both explanatory and response variables. The fundamental problem is that the conditional distribution of the response, given a particular level of the explanatory variable, neither follows one of the usual parametric families nor can be represented by a small, parametric expansion of the usual parametric families.

In studies of children, weight is often used as a proxy for the health of the child, both because of difficulties in measuring height (length) and because an unhealthy environment may well cause a reduction in both height and weight. The easily obtained measurement of weight is often tracked as a function of age, again an easily measured quantity. Figure 1 displays a kernel density estimate of the weight of $4,276$ girls, at birth, from a study of infants born in hospitals in Ohio. The data come from Amini et al. (1996) and are available as part of the Notz, Pearl and Stasny's Electronic Encyclopedia of Statistical Examples and Exercises (EESEE), available through W.H. Freeman. All of these girls were live, single births. Figure 2 displays a normal probability plot for these data. The figures clearly indicate the non-normality of the weight data. The kernel density estimate displays an unusual left hand tail which is not amenable to fits through introduction of an extra parameter or two to describe skewness and kurtosis. Other commonly used parametric families do not appear to fit the data. Consequently, these birthweight data cry out for a nonparametric fit.

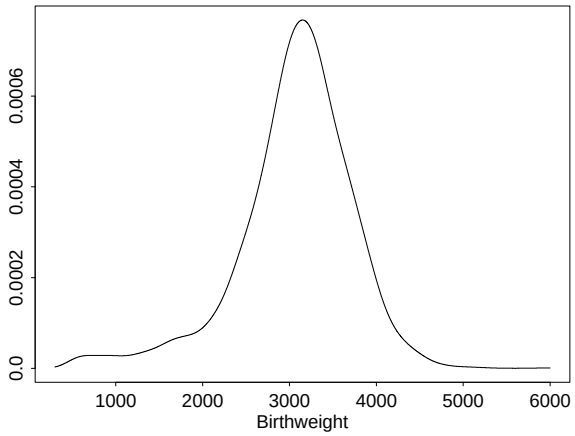

Figure 1: Kernel density estimate of birthweights.

Table 1 excerpts several values from a table of ages and weights of girls compiled by Bender and Remancus and available on the web at www.odc.com/anthro/deskref. The upper and lower standard deviations measure the size of typical departures from the "mean" in the direction of greater and lower weights. These values have been used to create percentiles for growth charts. An examination of the ratio of the upper to lower standard deviation shows a move from a left-skewed distribution to a right-skewed distribution.

Given the continuous nature of the aging and growth processes, these figures and tables suggest that (i) a nonparametric technique is needed to

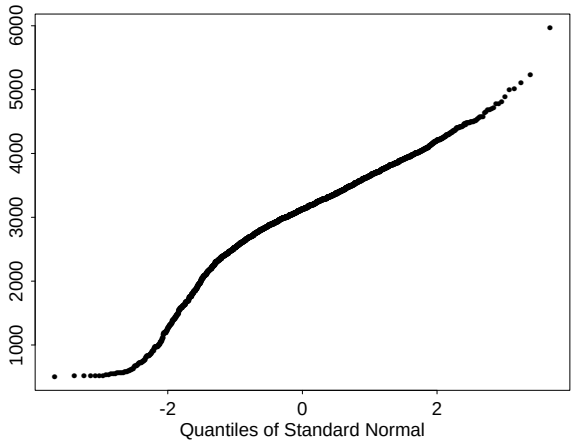

Figure 2: Normal probability plot for birthweights.

Table 1: Girls, weight for age. Age is in months, LSD, USD and RSD are the "lower standard deviation", "upper standard deviation", and ratio of the USD to LSD. All weights are in kilograms.

Age	LSD	Mean	USD	RSD
0	0.5	3.2	0.4	0.80
1	0.6	4.0	0.5	0.83
20	1.2	11.2	1.2	1.00
40	1.6	14.8	2.1	1.31
60	1.9	17.7	2.8	1.47
100	3.8	26.0	5.8	1.53

model the conditional density of weight given age and (ii) the error distribution in the model must evolve as age changes. The evolution should take the form of a smoothly changing distribution, where the conditional distribution of weight at nearby ages is nearly the same. This evolution of a nonparametric residual distribution is naturally approached from a nonparametric Bayesian viewpoint.

The philosophical motivation for this problem lies in the familiar statement often attributed to Box that while all models are wrong, some models are useful. The linear model and its many generalizations have shown great success across a wide variety of problems. Much of this success is due to the robustness of the fitting procedure to modest departures from the presumed residual structure. Though the model may be inadequate due to omission of relevant factors, model misspecification and an incorrect residual structure, the fit may nevertheless be in the right ballpark. Variations on the central limit theorem ensure that fitting techniques largely based on the mean response at a given covariate level will asymptotically yield good fits for the main trend in response given covariate. However, when such a model is used for predictive purposes, attention to modelling the residual structure is essential–there is no central limit theorem at work for a prediction about a single future observation.

An additional motivation for these models is found in a Bayesian treatment of random effects. As argued in Bush and MacEachern (1996), an adequate Bayesian model for random effects must place a prior distribution on the distribution from which effects are drawn. This distribution must have large support, suggesting a nonparametric prior distribution. In contrast, the prior distribution for fixed effects will typically focus on the effects themselves, resulting in a parametric form. For models with both sorts of effects (the so-called mixed model), prior distributions will naturally incorporate both parametric and nonparametric components. In many instances, one wishes to work with several random effects distributions, as in the setting of a multi-center clinical trial. In this setting, the distribution of patient specific parameters at each center is modelled as a collection of random effects. Current models restrict one to a collection of distributions that are either independent realizations of nonparametric distributions (perhaps given a few hyperparameters) or a single realization of the nonparametric distribution (perhaps adjusted by a parameter or two to account for location and scale differences). The trick is to write a model for the general random effects distribution that allows one to capture the notion of a collection of similar but not identical nonparametric distributions. The key modelling concept is that the realizations of the random distributions should be dependent. It should be noted that Muller et al.(1999) provide a recent alternative approach to this problem.

This work lays out the basic Bayesian approach to evolving, nonparametric distributions and, as a byproduct, provides a class of models for collections of nonparametric random effects distributions. These evolving distributions represent a generalization of Bayesian methods which place a prior distribution over the space of distribution functions. They enable us to write sensible models that capture the phenomenon described above: conditional distributions that can only be described in a nonparametric fashion, that are not amenable to description as a nonparametric location scale family and that change smoothly as the covariate changes. The class of models developed here contains a number of parameters that allow one to describe a great variety of behaviors. The new models are also designed to fit in

as a component of a hierarchical Bayesian model, and so provide a replacement for a parametric residual structure in essentially any parametric Bayesian model. The models play off of standard nonparametric Bayesian models, and so the Markov chain Monte Carlo techniques developed for fitting the standard models extend to the new models. These methods prove useful in a wide variety of modelling contexts, some of which are described in the concluding section.

2 Dependent Dirichlet Processes

The brief technical development to follow outlines a particular type of dependent nonparametric process, namely dependent Dirichlet process, or for short, the DDP. I have chosen to focus the development on this process because of the central role played by the Dirichlet process in nonparametric Bayesian modelling, because of the connections between models based on the Dirichlet process and finite mixture models–incidentally, many of the same goals that can be accomplished with the DDP can also be accomplished by working with finite mixture models–because the computational tractability of the Dirichlet process extends directly to simple DDP models and with slight additional work to more complex DDP models, and also because the DDP models comprise a rich enough class of models to allow us to capture the desired behavior.

The starting point for describing the DDP is a single Dirichlet process, described through Sethuraman's representation. The Dirichlet process places a distribution on the space of distribution functions by creating a distribution on the set of countable mixture distributions. In a single dimension, the random distribution function, F, can be described by the expression $F(y) = \sum_{i=1}^{\infty} p_i I(\theta_i^* \leq y)$, where the p_i sum to 1. The Dirichlet process provides the distribution for F by placing a distribution on the θ_i^* and the p_i. These distributions are governed by the parameter of the Dirichlet process, a non-null, finite measure, called α. For distributions on the real line, α is typically taken to be a measure that is absolutely continuous with respect to Lebesgue measure. The measure α is often described by its mass, M, and its shape, a probability measure or the corresponding distribution function, F_0. The vector θ^* and the vector p are mutually independent. The distribution on the θ_i^* is that of an independent and identically distributed sample from F_0. The distribution on the p_i follows from a distribution on an implicit set of parameters, the v_i. The v_i are an independent and identically distributed set of Beta$(1, M)$ random variates. The prescription for moving from the v_i to the p_i is to define $p_i = v_i \prod_{j=1}^{i-1}(1 - v_j)$.

The transition from a Dirichlet process to the "single-p DDP model," a special case of the DDP model, is most easily seen for a set of distributions that evolve over a single dimension. Referring to the dimension over which the distributions evolve as the covariate, we write $\mathcal{X}$ for the covariate space, and consider the set of distributions $F_x, x \epsilon \mathcal{X}$. These F_x are the random distributions. For the single-p models, the collection of random distributions will be specified by $F_x(y) = \sum_{i=1}^{\infty} p_i I(\theta_{ix}^* \leq y)$, for each $x \epsilon \mathcal{X}$, where the p_i sum to 1. In this model, one merely replaces θ_i^* with $\theta_{ix}^*, x \epsilon \mathcal{X}$.

The DDP model places a distribution on $F_{\mathcal{X}}$ in the following fashion. The p_i are once again random variables of a single dimension while the one-dimensional θ_i^* have been replaced by stochastic processes $\theta_i^*(x)$, $x \epsilon \mathcal{X}$, called $\theta_{i\mathcal{X}}^*$ to simplify notation. To complete the description of the distribution on $F_{\mathcal{X}}$, one parallels the development of the Dirichlet process. The vector v and the countable collection of stochastic processes, θ^*, are mutually independent. The $\theta_{i\mathcal{X}}^*$ form a random sample of stochastic processes (i.e., the $\theta_{i\mathcal{X}}^*$ are independent of one another, and they all follow some given distribution). In the case that $F_{0x} = F_0$, not depending on x, $\theta_{i\mathcal{X}}^*$ will often be a stationary stochastic process with continuous paths and index set $\mathcal{X}$. The distribution on p is exactly the same as the distribution of p for the Dirichlet process. The underlying random sample of v_i gives rise to the p_i through the same expression used for the Dirichlet process.

The simplest case of the DDP model occurs when the shapes of the conditional base measures, F_{0x}, do not depend on x. The parameters that govern this single-p DDP model, then, are a mass parameter, M, used to generate the v_i and from these the p_i, a shape for the base measure (more properly the distribution function correspoding to that shape), F_0, and a stationary stochastic process, $\theta_{\mathcal{X}}$ which has F_0 as the marginal distribution for each θ_x. When F_0 is normal, a Gaussian process can be used for $\theta_{\mathcal{X}}$.

The existence of the single-p DDP models is easily demonstrated when the stochastic process $\theta_{\mathcal{X}}$ has continuous paths with probability 1. In this case, the entire path of the stochastic process is determined by its value at a dense set of points. Since the rational numbers form a dense set in $\mathcal{R}^1$, pinning down each path at each of the rationals pins down the entire collection of paths. The additional variables, either the v_i or the p_i, are a countable collection of real valued random variables. Together,

the θ_{ix} at the rationals and the v_i determine the entire collection of $F_x, x\epsilon\mathcal{X}$. The question of existence of the distributions thus hinges only on the the existence of a joint distribution for a countable number of real-valued random variables. That this entire countable collection of real valued random variables has a well-defined distribution follows directly from the independence of the v_i and $\theta_{i\mathcal{X}}$ and the fact that the stochastic process, $\theta_{\mathcal{X}}$ is itself well-defined. This proof extends to more complex covariate spaces and to DDP's based on many stochastic processes which do not have continuous sample paths.

The DDP models can be developed in much greater generality than the stationary single-p DDP model. Key extensions and how they can be incorporated are the following: First, the stochastic process that drives the evolution of the F_x need not be stationary. This can be accomplished either by letting the correlation structure vary as x varies, and/or by allowing differing shapes for the base measures, so that F_{0x} varies with x. Such a modification will be useful when using the DDP models for the residual structure in the generalized linear model.

Second, the p_i can be allowed to vary with x. The model is referred to as a DDP model when the p_i do vary with x, but is no longer a single-p DDP model. To formally accomplish this, the individual variates v_i are replaced by stochastic processes, $v_{i\mathcal{X}}$, with the result being that the point masses at x are determined by the expression $p_{ix} = v_{ix}\prod_{j=1}^{i-1}(1-v_{jx})$. It is worth noting that when the $v_{i\mathcal{X}}$ are stochastic processes, the mass of the base measure is allowed to vary with x, yielding a parameter $M_{\mathcal{X}}$. This enables one to model the degree of proximity to a parametric form as a function of the covariates.

Third, the covariate space is in no way limited to a single dimension. There are many situations where the models will be used with a finite or countable discrete covariate space. In these settings, the stochastic processes simplify to random vectors of finite or countable length. An alternative view of the single-p model with a finite covariate space is that the distribution $F_{\mathcal{X}}$ follows a Dirichlet process, with each of the dimensions producing its own univariate distribution. The novelty of this model is its use in a hierarchical setting where an observation in the model would have a distribution depending only on one coordinate (or a set of coordinates) of the multivariate F, a use distinctly different than previous uses of the Dirichlet process. This model provides a means of generating a finite set of dependent distributions, and it would be appropriate in such settings as the multi-center clinical trial described earlier. When the covariate space is not of finite dimension, even the single-p DDP model is an extension of the Dirichlet process.

The DDP models allow great flexibility in their implementation. Although at first glance, one seems faced with the task of specifying stochastic processes that have F_{0x} as the perhaps unusual marginal distribution for each x, there are simple recipes for creating such stochastic processes. A particularly easy recipe for obtaining the target F_{0x} is to begin with a stochastic process that has continuous marginal distributions, say Z_x with marginal distribution G_x. Then define $\theta^*_{ix} = F^{-1}_{0x}(G_x(z_x))$, where the inverse cumulative distribution function is defined in the usual way. Since $G_x(z_x)$ follows the uniform distribution, θ^*_{ix} will follow the marginal distribution F_{0x}.

The DDP models have been developed with an eye toward their use as components in hierarchical Bayesian models. In particular, the models have been motivated as appropriate for use in the residual portion of a linear model or one of its many generalizations. In settings where the response variable is discrete, this approach is promising, although in settings where the response variable is thought of as continuous, use of these models in conjunction with a lower stage of the model to produce a continuous distribution is a sounder approach.

The DDP models may be fit directly or they may be used as a component in a hierarchical Bayesian model with use of Markov chain Monte Carlo simulation techniques. These techniques are now well developed for the Dirichlet process, and these techniques carry over in a straightforward fashion to the DDP models. In the simplest case, where the covariate space is finite, as mentioned above, the single-p DDP model may be viewed as a Dirichlet process used in a novel fashion. The computational methods described in MacEachern (1998) may be used, with no changes whatsoever, for this model. In more complex settings, where the DDP may be used to model the residual structure, computations will often be done conditional on the observed levels of the covariate. Once again, with the single-p DDP model, the computational strategies developed for the Dirichlet process carry over without change. As is standard in predictive problems, forecasts at unobserved levels of the covariate can be produced through simulations that require only the output of the simulation used to fit the model. The distribution of F_x at unobserved x can be handled in the same fashion. When one moves from the single-p DDP model to more general DDP models, the current computational strategies need to be extended to allow the p_i to vary with x. This can be accomplished through

use of Metropolis-Hastings steps.

3 Properties of Dependent Dirichlet Processes

The DDP model has a number of desireable properties which will be more rigorously established elsewhere. These properties extend Ferguson's (1973) early motivation of the Dirichlet process to settings that involve covariates. Four desireable properties of models for a collection of random distribution functions are (i) that the support of the distribution on $F_{x_1}, \ldots, F_{x_d}$ should be large, (ii) that the distribution, when used as a component in a Bayesian hierarchical model should be amenable to updating, (iii) that the marginal distribution of F_x should follow a familiar distribution at any given level of the covariate, and (iv) that the realized distributions, F_x should converge to the realized F_{x_0} as $x \to x_0$. A brief synopsis of some of these properties and loose versions of conditions that yield them follows:

1. The prior distribution on $F_{x_1}, \ldots, F_{x_d}$ has full support, provided the stochastic process $\theta_{\mathcal{X}}$ is rich enough.

2. As mentioned in the preceeding section, the DDP models are amenable to simulation based fits.

3. The marginal distribution, F_x follows a well-known distribution. In fact, $F_x \sim \mathrm{Dir}(M_x, F_{0x})$ for each $x \epsilon \mathcal{X}$, where the right hand side of the above expression indicates a Dirichlet process.

4. The distributions F_x are continuous in x. This feature is what produces distributions that evolve as the covariate changes. It can be obtained by working with stochastic processes which produce continuous paths for θ_x and for p_x.

   F_{x_1} and F_{x_2} tend toward independent distributions as x_1 and x_2 become more distant. To accomplish this, we need θ_{x_1} to tend toward independence from θ_{x_2} and also v_{x_1} to tend toward independence from v_{x_2}. This can be accomplished by writing stochastic processes which yield the decay toward independence.

5. In addition, a spectrum of inference can be captured, ranging from a nearly parametric inference (take M_x nearly ∞ for all values of x) to inference that relies on a single nonparametric distribution (take the stochastic process $\theta_{\mathcal{X}}$ for which $\theta_{ix} = \theta_i$ for all x) to inference that shows a strong dependence between distributions with nearby x (take slowly varying stochastic processes for θ_x and/or v_x) to inference that encourages quick changes in the distributions as x changes (take θ_x that change quickly in x).

One of the beauties of the DDP model is that we can exploit the wealth of knowledge about nonparametric Bayesian models and stochastic processes to select parameter values that produce the behavior we want. We also create a set of models where we have a clear understanding of the behavior of the conditional distribution, an essential feature when we wish to use the models as a component in a large, hierarchical Bayesian model.

4 Other Dependent Nonparametric Processes

The strategy used to develop the DDP can be used to develop many other dependent nonparametric processes. The general method by which a finite dimensional nonparametric distribution is created is to lay out a structure whereby a countable set of (usually independent) variates produces a distribution. Distributions produced in this fashion include finite mixture distributions (when the number of components in the mixture is unbounded, the distributions qualify as nonparametric), Polya trees and series expansions.

In order to create a dependent nonparametric process that generalizes one of these methods of constructing a random distribution, one need only replace each of the countable set of variates with a stochastic process. The marginal distribution of the dependent nonparametric process at any given level of the covariate will match the basic nonparametric distribution. Full support of the models generally follows from use of stochastic processes with rich enough support. Local dependence of the distributions follows from the dependence within realizations of the stochastic process.

For a finite mixture distribution, the variates deterimining the random distribution are one which indicates the number of components of the mixture, say k, and k more which describe the components of the mixture. These latter variates are often of dimension greater than 1.

For a Polya tree, the variates determining the random distribution are a collection of beta variates with parameters specified at each split of the tree. Note that dependent Dirichlet processes constructed by introducing the dependence through the Polya

tree that matches the Dirichlet process has a different feel than the DDP model described in this paper.

For a series expansion, the distribution (or density) is represented as a finite or infinite series. The variates determine the number of terms in the series and the coefficients of the terms.

5 Uses of Dependent Dirichlet Processes

DDP models are naturally used to model the residual structure in the linear model. The traditional linear model takes $Y_i = x_i\beta + \epsilon_i$, with the ϵ_i a random sample of normal variates with mean 0 and variance σ^2. The DDP models allow us to replace the assumption of normality for the residuals with the assumption that the error distributions evolve with the covariate. The random sample of normal variates is replaced by a sample of independent variates where $\epsilon_i \sim F_{x_i}$ for $i = 1, \ldots, n$. To adjust for the approximate continuity of the response variable, an additional stage is added to the hierarchy to smooth the response distribution, just as is done with models based on the Dirichlet process.

A second expansion on the linear model pursues a line of thought motivated by the generalized linear model. A value, $x_i\beta$, is computed for each observation and linked to the response through the identity transformation. The residual distribution is indexed by the value $x_i\beta$ rather than by x_i itself, so that the residuals correspond to a sample of independent variates where $\epsilon_i \sim F_{x_i\beta}$. The advantages of this approach are parsimony in the modelling process and some computational simplification due to the reduction in dimension of the space that indexes the random distributions.

Expansion of parametric models along the lines outlined above can be done in very general settings. The generalized linear model is the most natural for generalization, as it already focuses on relaxing the assumption of normal errors. The generalized linear model replaces the normal distribution with a more general parametric form and replaces the identity link function with a more general form, though still "parametric" in basic developments of the model. The most natural use of the DDP in this setting is to retain the link function used in the generalized linear model and to take the base measure F_{0x} to be the form that would be used in the parametric version of the model. Models for a discrete response such as Poisson or binomial would rely the DDP directly, thus providing a generalization of the work by Carota and Parmigiani (1997) to incorporate the notion of an evolving response distribution. Further work would allow flexibility in the form of the link function, thus allowing one to incorporate both a nonparametric trend in the model as well as a nonparametric residual structure.

The models have a considerable variety of additional uses. One such use is in modelling a continuous, nonparametric distribution. Currently, the main nonparametric Bayesian approach to directly modelling a continuous distribution is to use a Polya tree, with particular choices for the splits that determine the tree and the number of parameters (number of balls) in each of the urns. With appropriate choices of these parameters, one guarantees that the random distribution is continuous. The drawback of this method is that the tree structure implicit in this model induces discontinuities in the density of the distribution. The DDP models suggest an alternative approach. Write a DDP model for a collection of distributions, $F_{\mathcal{X}}$, and supplement this with a distribution, G, on the unobserved covariate, x. Define $F(y) = \int F_x(y)dG(x)$. Although each of the conditional distributions, F_x, is discrete, marginalization over a covariate for which G is continuous can result in a continuous F.

Work that has already been done makes it clear that these models have much to contribute in a wide variety of applications. I, along with a number of coauthors, intend to work on the uses for DDP models described above as well as a number of variations on these themes in the years to come.

6 References

Amini, S. B., Catalano, P. M., and Mann, L. I. (1996), Births to Unmarried Mothers: Trends and Obstetrical Outcomes, unpublished manuscript.

Bush, C. A. and MacEachern, S. N. (1996). A Semi-parametric Bayesian Model for Randomized Block Designs. Bka., **83**, 275-286.

Carota, C. and Parmigiani, G. (1997). Semiparametric Regression for Count Data. Tech report 97-17, ISDS, Duke University.

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

MacEachern, S. N. (1998). Computational Methods for Mixture of Dirichlet Process Models, in "Practical Nonparametric and Semiparametric Bayesian Statistics", D. Dey, P. Muller, D. Sinha (eds.), 23-44.

Müller, P., Quintana, F. and Rosner, G. (1999). Hierarchical Meta-Analysis over Related Non-parametric Bayesian Models. Tech Report 99-22, ISDS, Duke University.