Dependent Dirichlet Processes

Steven N. MacEachern

May 8, 2000

Abstract

Dependent Dirichlet processes provide a means of modelling a collection of random distributions as related to, but not identical to, each other. A key point is that the *realizations* of these processes are dependent. This article develops the processes through definition and results. It also describes the basic tools for computations with the processes and suggests a number of applications in which they will be useful. The extension of the Dirichlet process to the dependent Dirichlet process can be paralleled for essentially any other nonparametric process which, for a single random distribution, relies on a countable collection of random variables. A parallel development for two such processes is indicated.

Key Words: Covariates; evolving distributions; Markov chain Monte Carlo; noise floor; nonparametric Bayesian analysis; random effects.

[*] Steven N. MacEachern is Associate Professor Department of Statistics, The Ohio State University, Columbus, OH 43210. e-mail: `snm@stat.ohio-state.edu`

[†] I would like to thank the participants of the Second International Workshop on Nonparametric Bayesian Methods for the valuable comments I received on this work at the conference. In particular, I would like to thank Stephen Walker for insightful comments about the theoretical underpinnings of dependent Dirichlet processes.

1 Introduction

The primary focus of nonparametric Bayesian methods, as with all Bayesian methods, is modelling the underlying phenomenon that has produced the data. To date, most Bayesian work in the area has relied on modelling concepts whose value has long since been established in a classical framework. The translation of the modelling concepts into Bayesian terms often provides a clarity of perspective that directly suggests a particular formalization of the concept and that leads to models that are both theoretically attractive and that have been shown to perform well in an empirical sense.

The familiar conceptual division of the effects in a model into fixed effects and random effects provides one example of the benefits that a Bayesian perspective brings to a modelling strategy. In the context of randomized block designs, Smith (1973) takes a Bayesian view, demonstrating how to recover the traditional classical estimates with Lindley and Smith's (1972) hierarchical model. One merely models the fixed effects with improper prior distributions, but must go through some rather contrived gynmastics to first model the random effects as independent draws from a normal distribution and to then obtain estimates conditional on a particular estimated value of the variance of the random effects distribution. More recently, Bush and MacEachern (1996) reexamine this modelling concept. They suggest that the natural form of a Bayesian model follows from the qualitative form of information that one is likely to have about the effects. For the fixed effects, where interest focuses on the individual effect, and where information is likely to be about each individual effect, create a prior distribution for these effects. This prior distribution is typically of low dimension, and will often assume a parametric form. For random effects, where interest focuses on the distribution of the effects, and where prior information is expressed about this distribution, create a prior distribution for this unknown distribution, say F. In a desire to have full support for the prior distribution, the natural modelling strategy is to place a nonparametric prior distribution on F. For problems that would be classically approached with a best linear unbiased predictor, the distribution of the random effects is naturally modelled with a nonparametric prior distribution, with inference still made about the individual effects. This Bayesian random effects modelling strategy has proven to be successful not only in the context of randomized block analyses, but also in such areas as repeated measures designs, random effects regression problems, and longitudinal and time-series analyses. Dey, Müller and Sinha (1998) contains several examples of such work.

The Bayesian modelling perspective and current computational methods allow us to divide large problems into more manageable components, to design a prior distribution for each component of the problem, and to link the components in a hierarchical fashion. This decomposition of a model into pieces that are linked together provides an easy approach to model building for multi-center clinical trials. In these trials, a therapy or therapies are examined at a number, say d, of locations. At each location, patients are recruited to the study, assigned to a therapy, and the results observed. A number of measurements are made on each patient. The model will contain two main components: One for the therapies, and one for the patient effects. The model for therapies expresses prior information about the therapies under study. The patient effects are naturally modelled as random effects, and the Bayesian random effects modelling strategy suggests that a nonparametric prior distribution be placed on the patient effects distribution. A standard semiparametric Bayesian model would pursue one of two courses. Either a single distribution would be presumed for all patient effects, and a nonparametric prior placed on this distribution, or, to account for differences in patient populations at different centers, nonparametric prior distributions would be placed on the d separate patient effect distributions. These two extremes of modelling would then take, with a foreshadowing of notation introduced in the next section to allow for covariates, $F_{x_1} = \ldots = F_{x_d}$ or it would take $F_{x_1}, \ldots, F_{x_d}$ to be mutually independent. A substantial limitation of current nonparametric Bayesian techniques is that they allow only modest improvements on these two extremes. In the first case, the distributions may be allowed to differ by a small number of parameters, perhaps locations and scales, but the distributions are identical in many ways; in the second case, the distributions may be linked together through hyperparameters that control the d nonparametric prior distributions. But, conditional on these hyperparameters, the realized distributions are independent. What is needed is a modelling strategy that allows the set of random effects distributions to be similar, but not identical, to each other. Tomlinson and Escobar (1998) provide an approach that they term analysis of densities. Müller, Quintana and Rosner (1999) discuss multi-center trials and provide one means of incorporating dependence among the random F_x. The focus of this paper is on the definition and development of a class of nonparametric processes that achieves this goal in a natural, extensible fashion.

Once described, the processes lend themselves to an enormous variety of uses. An immediate use, and the one illustrated in Section 4, is to create a more general version of the normal theory linear model. In this use, the covariates enter the model in the traditional fashion. The error distribution which can assume an arbitrary form, is allowed to evolve with the covariate, typically changing in a smooth fashion. By varying the parameters of the prior process, the error distribtuions range from a single normal distribution to a single nonparametric distribution through slowly varying nonparametric distributions to independent nonparametric distributions. Modest changes to this nonparametric Bayesian version of the basic linear model allow us to incorporate changes of scale and so to provide a competitor to weighted least squares. The generalized linear model is the next natural target. A means to this extension should be clear from the development of the processes and the work on the linear model.

Dependent nonparametric processes provide the solution to the modelling problem. The development of the processes will keep in mind slightly altered properties akin to the two desireable properties described by Ferguson (1973) as important for any nonparametric Bayesian procedure to have, and two additional properties that are generally desireable in this more complex setting.

1. The support of the prior distribution on $F_{x_1}, \ldots, F_{x_d}$, for any distinct $x_1, \ldots, x_d$, should be large.

2. The prior distribution, when used as a component in a Bayesian hierarchical model, should be amenable to updating in a reasonably painless fashion, either analytically or computationally.

3. The marginal distribution of F_x should follow a familiar distribution at any given level of the

covariate, x.

4. The realized distributions, F_x, should converge to the realized F_{x_0} as $x \to x_0$.

The development of these dependent nonparametric processes will be general enough to allow great flexibility in their use as a prior distribution, rendering them useful in a wide variety of Bayesian problems. Simplifications that enhance property 2 will be described, and some commentary is given on how to generalize the processes at the expense of properties 2, 3 and 4.

The remainder of this paper is organized as follows. Section 2 contains notation and a definition of the processes. Section 3 presents theoretical results, establishing that the processes have the desired properties. Section 4 presents an analysis of a set of "regression" data. Section 5 demonstrates the ease with which these new processes can be used to construct a prior that assigns probability one to the set of continuous distribution functions, providing an alternative to the Polya tree distribution. The concluding section describes a number of other examples where these prior distributions will find use. It also indicates how the strategy described herein can be used to extend other nonparametric processes that produce a random distribution function to processes that produce a collection of smoothly evolving random distribution functions. For additional motivation, see MacEachern (1999).

2 Notation and motivating ideas

The Dirichlet process enjoys a central role in nonparametric Bayesian analysis when a prior distribution is placed on the space of distribution functions. One of the biggest reasons for this central role is that there are many representations of the Dirichlet process, each of which provides insight into the structure of the prior distribution and several of which are useful when computing the posterior distribution of the process. For the purposes of this paper, the two most useful representations are Sethuraman's representation (1994) and the Polya urn scheme (Blackwell and MacQueen, 1973). A well-known fact is that a Dirichlet process with a sufficiently complex base measure (i.e., one supported at more than a finite number of points) assigns probability one to the set of countable, discrete distributions (Blackwell, 1973). See Ferguson (1973) for a definition of and fundamental results on the Dirichlet process.

With Sethuraman's representation of the Dirichlet process, a focus is placed on the description of a discrete distribution as a collection of locations and the mass associated with each location. The Dirichlet process has, as its parameter, a positive measure α that can be represented in terms of its mass, $M > 0$, and its shape, F_0. The parameter F_0 is a distribution function.

The end result of the Dirichlet process is to define a random distribution from which values of $\theta = (\theta_1, \ldots, \theta_p)$ are drawn. Hence, the focus is on a space appropriate for Θ. The Borel sets over $\mathcal{R}^p$ provide a workable space for all of the examples presented here, and so the processes will be defined in this setting. Generically, A will represent a measureable set, F will denote the random distribution function, and P will denote the probability measure corresponding to F. For the dependent Dirichlet process, a subscript x will refer to the value of a covariate in the space $\mathcal{X}$. $\mathcal{X}$ may be finite, countable, an interval of the real line, or a more general space. It will serve as the index space for a collection of stochastic processes.

Theorem 2.1 *The Dirichlet process (Sethuraman). Let $\theta_1, \theta_2, \ldots$ be a sequence of draws from F_0, and let $V_1, V_2, \ldots$ be a sequence of draws from the Beta(1,M) random variable. Assume that all random variables in the sequences $\theta_1, \theta_2, \ldots$ and $V_1, V_2, \ldots$ are mutually independent. Define $p_i = V_i \prod_{j<i}(1-V_j)$, for $i = 1, 2, \ldots$. For every measureable set A, take*

$$P(A) = \sum_{i|\theta_i \epsilon A} p_i.$$

Then F, corresponding to P, follows a Dirichlet process with base measure α determined by $M \cdot F_0$.

The θ_i are referred to as the locations and p_i is referred to as the mass assigned to location i. Note that the θ_i are merely possible values that the random variable θ may assume. They need not be location parameters for a further portion of a model. The relationship between α, P_0 and F_0 is $\alpha = M \cdot P_0$, with

F_0 the distribution function corresponding to P_0. When $\Theta = \mathcal{R}^p$, mention of P_0 is often avoided, and one relies on the relation $\alpha((-\infty, \theta]) = M \cdot F_0(\theta)$.

Dependent Dirichlet processes generalize the Dirichlet process to allow for a collection of nonparametric distributions, the realizations of which are dependent, with the level of the covariate, x, governing the degree of dependence. The idea that drives the dependence for one-dimensional θ is that, in the presence of a covariate, the location, θ, of the Dirichlet process can be replaced by the sample path of a stochastic process, denoted $\theta_{\mathcal{X}}$. This sample path provides the location at each value of the covariate. Similarly, the beta random variable, V, which, through Sethuraman's trick, produces the mass assigned to a value of θ, can be replaced by a process, denoted $V_{\mathcal{X}}$. This process produces the mass assigned to the value of θ_x for each value of the covariate. In higher dimensional settings, the same idea is extended to multivariate locations, either through multivariate stochastic processes or through collections of stochastic processes that yield $\theta_x = (\theta_{x,1}, \ldots, \theta_{x,p})$. This collection of stochastic processes is itself a multivariate stochastic process.

A stochastic process $Z_{\mathcal{X}} \overset{def}{=} (Z_x, x \epsilon \mathcal{X})$, is defined by its index set, $\mathcal{X}$, its state space, S, and the dependence relations among the Z_x. For the sequel, it is presumed that the state space for each Z_x is $\mathcal{R}^p$. In order to facilitate implementation of dependent Dirichlet processes, the definition of the processes introduces notation with some redundancy. The redundant notation makes a link between the stochastic process $Z_{\mathcal{X}}$ and a corresponding stochastic process $\theta_{\mathcal{X}}$ and a similar link between stochastic processes $U_{\mathcal{X}}$ and $V_{\mathcal{X}}$. The process $U_{\mathcal{X}}$ is real valued.

In the following definition, $F_{\mathcal{X}}$ represents the collection of p-dimensional random distribution functions. The mass parameter is $M_{\mathcal{X}}$. This parameter is finite and positive for all values of $x \epsilon \mathcal{X}$, but may vary with x. The set of base c.d.f.s is $F_{0,\mathcal{X}}$. Each of these c.d.f.s is a p-dimensional distribution function. Should need arise, $P_{0,\mathcal{X}}$ represents the collection of probability measures corresponding to $F_{0,\mathcal{X}}$, and the collection of base measures is $\alpha_{\mathcal{X}} \overset{def}{=} M_{\mathcal{X}} \cdot P_{0,\mathcal{X}}$. The stochastic processes $Z_{\mathcal{X}}$ and $U_{\mathcal{X}}$ provide draws that are turned into locations and probabilities, respectively. The transformations $T_{Z,\theta;\mathcal{X}}$

specify a mapping of Z_x into θ_x for each $x\epsilon\mathcal{X}$. The transformations $T_{U,V;\mathcal{X}}$ specify a mapping of U_x into V_x for each $x\epsilon\mathcal{X}$. The transformations must be such that if Z_x has some distribution, say G_x, $T_{Z,\theta;x}$ has the distribution $F_{0,x}$ for each $x\epsilon\mathcal{X}$. Similarly, if U_x has some distribution, say G_x, then $T_{U,V;x}$ must have the Beta$(1, M_x)$ distribution. All of these transformations are assumed to be measureable.

Definition 2.2 *Dependent Dirichlet processes are defined by the relation*

$$F_{\mathcal{X}} \quad \sim \quad DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}}),$$

where the parameters follow the restrictions described in the previous two paragraphs. Z_i *and* U_i*, for* $i = 1, 2, \ldots$*, are mutually independent realizations of the stochastic processes* $Z_{\mathcal{X}}$ *and* $U_{\mathcal{X}}$*. The processes* θ_i *and* V_i *are determined by the transformations of* Z_i *and* U_i*, respectively. At each* $x\epsilon\mathcal{X}$*, the distribution* F_x *is defined to be the discrete distribution characterized by* $P_x(A) = \sum_{i|\theta_{ix}\epsilon A} p_{ix}$*, or, equivalently, by* $F_x(\theta) = \sum_{i|\theta_{ix,j} \leq \theta_j} p_{ix}$.

The next section presents a set of sufficient conditions which ensure the existence of the process. The conditions also ensure that property 4 of the introduction is satisfied. Before turning to theoretical results, a simplification of notation is in order.

Definition 2.3 *The standard transformation from* $Z_{\mathcal{X}}$ *to* $\theta_{\mathcal{X}}$ *is defined by a pointwise, coordinatewise transformation based on the c.d.f.:*

$$\theta_{ix} = F_{0i,x}^{-1}(G_{ix}(Z_{ix}))$$

for i = 1, ..., p, and for each $x \in \mathcal{X}$*, where* G_{ix} *is the marginal c.d.f. of the* i^{th} *coordinate of* Z_x*, where* G_i *is assumed to be a continuous distribution, and where* $F_{0i,x}$ *is the* i^{th} *coordinate of the base c.d.f. for* θ_x*. Similarly, the standard transformation from* $U_{\mathcal{X}}$ *to* $V_{\mathcal{X}}$ *is*

$$V_{ix} = F_{Bx}^{-1}(G_x(U_x))$$

for each $x\epsilon\mathcal{X}$*, where* G_x*, assumed to be a continuous distribution, is the marginal c.d.f. of* U_x*, and where* F_{Bx} *is the Beta*$(1, M_x)$ *distribution.*

When describing dependent Dirichlet processes, if the standard transformation for either Z or U is used, it will be dropped from the definition of the process. Similarly, the $\mathcal{X}$ subscript will be dropped if the parameter does not vary as x varies. This convention is illustrated in Section 4.

3 Theoretical Results

The basic theoretical results for dependent Dirichlet processes ensure first that the processes exist, and then guarantee that the properties described in the introduction are satisfied. We first turn to the existence result.

Theorem 3.1 *Assume that $F_{\mathcal{X}} \sim DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}})$. Let $Z_{\mathcal{X}}$ be a separable stochastic process with state space $\mathcal{R}^p$, and let $U_{\mathcal{X}}$ be a separable real valued stochastic processes with $\mathcal{X}$ either a countable set or an open set in $\mathcal{R}^k$. Then, the dependent Dirichlet process exists and the distribution of $(F_{x_1}, \ldots, F_{x_d})$ is uniquely defined for each d-tuple $(x_1, \ldots, x_d) \epsilon \mathcal{X}_1 X \ldots X \mathcal{X}_d$.*

Proof. To ensure that the dependent Dirichlet process exists, we must show that it is well defined. First consider the case of a countable space $\mathcal{X}$. The countable collection of distributions, $F_{\mathcal{X}}$, is determined by a countable collection of random variables, U_{ix} and Z_{ix}, $i = 1, 2, \ldots$ and $x \epsilon \mathcal{X}$. This collection of variables has a well-defined distribution since the $U_{i\mathcal{X}}$ and $Z_{j\mathcal{X}}$ are mutually independent and since the stochastic processes $U_{\mathcal{X}}$ and $Z_{\mathcal{X}}$ are, by definition, well-defined. Measureability of the transformations, T, ensures that the distribution of $(\theta_{i\mathcal{X}}, V_{i,\mathcal{X}})$ is well-defined.

Next, consider the case where $\mathcal{X}$ is an interval of $\mathcal{R}^1$. A separable stochastic process has its entire sample path determined by its value at a dense set of points, and the rationals are dense in any open set in $\mathcal{R}^k$. From the preceeding paragraph, the distribution of $(U_{i\mathcal{X}}, Z_{i\mathcal{X}})$, $i = 1, 2, \ldots$ is well-defined when $\mathcal{X}$ is restricted to the rationals in $\mathcal{X}$. Since $Z_{\mathcal{X}}$ and $U_{\mathcal{X}}$ are assumed to be separable, the entire joint distribution of $U_{i\mathcal{X}}$ and $Z_{i\mathcal{X}}$ is well-defined.

The first property for the dependent Dirichlet process is that, for any distinct $x_1, \ldots, x_d \epsilon \mathcal{X}$, the distribution of $(F_{x_1}, \ldots, F_{x_d})$ should have large support. Ideally, the support of the distribution would be full with respect to every sensible metric on the space of distributions. However, the nonparametric nature of the problem precludes such statements. See Diaconis and Freedman (1986) and its discussion for viewpoints on the variety of relevant metrics in the case of a single nonparametric distribution. Instead, one must focus on a particular metric–here, a metric which induces the weak topology on the space of distribution functions for each component distribution. The following lemma describes the distribution of F_x at a single x. The lemma justifies the name dependent Dirichlet processes, and it leads the way to establishing the support result. It also establishes property 3 of the introduction.

Lemma 3.2 *Assume that $F_{\mathcal{X}} \sim DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}})$. Let $Z_{\mathcal{X}}$ be a stochastic process with state space $\mathcal{R}^p$, and let $U_{\mathcal{X}}$ be a real valued stochastic process with $\mathcal{X}$ either a countable set or an open set in $\mathcal{R}^k$. Then, for every $x \epsilon \mathcal{X}$, $F_x \sim Dir(M_x, F_{0,x})$.*

Proof. Fix some $x \epsilon \mathcal{X}$. The definition of dependent Dirichlet processes ensures that (i) the Z_{ix} are i.i.d. draws from some distribution, say G_x, and hence that the θ_{ix} are i.i.d. draws from $F_{0,x}$, that (ii) the U_{ix} are i.i.d. draws from some distribution, say G_x, and hence that the V_{ix} are i.i.d. draws from the Beta$(1, M_x)$ distribution, and that (iii) the entire collection of θ_{ix} and V_{ix} are mutually independent. These are exactly the features that Sethuraman uses to describe the Dirichlet process. Matching the parameter values yields the result.

The next theorem provides conditions that guarantee full support under metrics inducing the weak topology. The essential condition is that the process $Z_{\mathcal{X}}$ must be rich enough to ensure full support of the joint distribution at $(x_1, \ldots, x_d)$. The process $U_{\mathcal{X}}$ only needs a surprisingly limited amount of flexibility. The transformations from Z to θ and U to V must not be too bad. Since interest is in a collection of d distributions, the result relies on $d(F, G) = \sum_{i=1}^{d} d_L(F_i, G_i)$, with d_L representing the

Levy distance. It is easily verified that d is a distance.

Theorem 3.3 *Assume that $F_{\mathcal{X}} \sim DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}})$, with $\mathcal{X}$ either a countable set or an open set in $\mathcal{R}^k$. Take $x_1, \ldots, x_d$ to be a set of distinct points in $\mathcal{X}$. Define $Z_{x,d} = (Z_{x_1}^t, \ldots, Z_{x_d}^t)^t$ and $U_{x,d} = (U_{x_1}, \ldots, U_{x_d})^t$. Then, $(F_{x_1}, \ldots, F_{x_d}) \sim DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}})$. Assume that $Z_{x,d}$ induces a distribution on $\theta_{x,d}$ which is mutually absolutely continuous with $(F_{0,x_1}, \ldots, F_{0,x_d})$, and assume that $U_{x,d}$ induces a distribution on $V_{x,d}$ which, for some $\gamma > 0$ and every $0 < \delta_L < \delta_U < \gamma$ assigns positive probability to the event that $\delta_L < V_{x,i} < \delta_U$ for $i = 1, \ldots, d$. Consider any $(F_1, \ldots, F_d)$ which is absolutely continuous with respect to $(F_{0,x_1}, \ldots, F_{0,x_d})$. For every $\epsilon > 0$, under the metric d, the ϵ-ball about $(F_1, \ldots, F_d)$ is assigned positive probability by the dependent Dirichlet process.*

Proof. See Appendix A.

Corollary 3.4 *Under the assumptions of Theorem 3.3, consider any $(F_1, \ldots, F_d)$ in the closure (under the metric d) of the set of distributions that are absolutely continuous with respect to $(F_{0,x_1}, \ldots, F_{0,x_d})$. Then, for each $\epsilon > 0$, the ϵ-ball about $(F_1, \ldots, F_d)$ is assigned positive probability by the dependent Dirichlet process.*

Proof. Fix $(F_1, \ldots, F_d)$ and ϵ. There exists some distribution $(G_1, \ldots, G_d)$ which is absolutely continuous with respect to $(F_{0,x_1}, \ldots, F_{0,x_d})$ and which is within $\epsilon/2$ of $(F_1, \ldots, F_d)$. The $\epsilon/2$-ball about $(G_1, \ldots, G_d)$ is assigned positive probability, and it is contained in the ϵ-ball about $(F_1, \ldots, F_d)$.

Corollary 3.5 *Assume that $F_{\mathcal{X}} \sim DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}})$, with $\mathcal{X}$ either a countable set or an open set in $\mathcal{R}^k$. Take $x_1, \ldots, x_d$ to be a set of distinct points in $\mathcal{X}$. Define $Z_{x,d} = (Z_{x_1}^t, \ldots, Z_{x_d}^t)^t$ and $U_{x,d} = (U_{x_1}, \ldots, U_{x_d})^t$. Then, $(F_{x_1}, \ldots, F_{x_d}) \sim DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}})$. Assume that the distribution of $Z_{x,d}$ is mutually absolutely continuous with respect to the product of the distri-*

butions on $(Z^t_{x_1}, \ldots, Z^t_{x_d})^t$. Also assume that the distribution of $U_{x,d}$ is mutually absolutely continuous with respect to the product of d uniform $(0, 1)$ distributions. Consider any $(F_1, \ldots, F_d)$ which is absolutely continuous with respect to $(F_{0,x_1}, \ldots, F_{0,x_d})$. For every $\epsilon > 0$, under the metric d, the ϵ-ball about $(F_1, \ldots, F_d)$ is assigned positive probability by the dependent Dirichlet process. Furthermore, for any $(F_1, \ldots, F_d)$ in the closure (under the metric d) of the set of distributions that are absolutely continuous with respect to $(F_{0,x_1}, \ldots, F_{0,x_d})$, for each $\epsilon > 0$, the ϵ-ball about $(F_1, \ldots, F_d)$ is assigned positive probability by the dependent Dirichlet process.

Proof. Verify the conditions of Theorem 3.3 and Corollary 3.4.

Theorem 3.6 *Assume that $F_{\mathcal{X}} \sim DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}})$, with $\mathcal{X}$ either a countable set or an open set in $\mathcal{R}^k$. Let $Z_{\mathcal{X}}$ and $T_{Z,\theta;\mathcal{X}}$ be a stochastic process and collection of transformations such that the induced paths of $\theta_{\mathcal{X}}$ are continuous. Let $U_{\mathcal{X}}$ and $T_{U,V;\mathcal{X}}$ be a stochastic process and collection of transformations such that the induced paths of $V_{\mathcal{X}}$ are continuous. Then, for any $x_0 \epsilon \mathcal{X}$, $lim_{x \to x_0} P(d(F_x, F_{x_0}) > \epsilon) = 0$.*

Proof. Since the $V_{\mathcal{X}}$ have continuous paths, the induced $p_{\mathcal{X}}$ also have continuous paths for $i = 1, 2, \ldots$. Each of the following probabilities can be made arbitrarily large: (i) the probability that the distance from F_{x_0} to a finite approximation given by $\delta_{\theta_{N+1}}(1 - \sum_{i=1}^N p_i) + \sum_{i=1}^N \delta_{\theta_i} p_i$ is less than $\epsilon/3$, by choosing N sufficiently large, (ii) the probability that $|\theta_{i,x_0} - \theta_{i,x}| < \epsilon/3$, by considering only x near enough to x_0, and (iii) the probability that $\sum_{i \epsilon S} |p_{i,x_0} - p_{i,x}| < \epsilon/3$, for every subset of indices, S, of $1, \ldots, N+1$, by considering only x near enough to x_0. If each of these three events occurs, $d(F_x, F_{x_0}) < \epsilon$. Since the probability of any one of the events not occurring can be made arbitrarily small, the result follows.

Corollary 3.7 *Assume that $F_{\mathcal{X}} \sim DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}})$, with $\mathcal{X}$ either a countable* *set or an open set in $\mathcal{R}^k$. Further suppose that $Z_{\mathcal{X}}$ and $U_{\mathcal{X}}$ have continuous sample paths and that $T_{Z,\theta;\mathcal{X}}$ and $T_{U,V;\mathcal{X}}$ are continuous transformations. Then, for any $x_0 \epsilon \mathcal{X}$, $lim_{x \to x_0} P(d(F_x, F_{x_0}) > \epsilon) = 0$.*

Proof. Continuity of the two stochastic processes and the related transformations ensure that the conditions of Theorem 3.6 are met.

The posterior distribution of $F_{\mathcal{X}}$, given $\theta_1, \ldots, \theta_{n_i}$, for $i = 1, \ldots, d$, observed at covariate values $x_1, \ldots, x_d$, is not typically a dependent Dirichlet process. However, for some particular distributions with modest sample sizes, the posterior distribution is amenable to numerical integration; for larger sample sizes, the posterior distribution may be described by means of a simulation method. The essential analytical result for many modern simulation methods for high-dimensional problems, such as the Gibbs sampler, the Metropolis-Hastings algorithm, or the sequential importance sampler, is the joint distribution of a collection of parameters which implies a set of conditional distributions. In the nonparametric context, one must take care to describe the distribution over the uncountable set of parameters by means of finite dimensional distributions. The next result describes the essentials of relevant joint distributions. For clarity, the result is expressed in terms of a density rather than a distribution.

In the result itself, f_{θ_x} represents the density of the vector $(\theta_{x_1}, \ldots, \theta_{x_d})$ induced by $Z_{\mathcal{X}}$ and $T_{Z,\theta;\mathcal{X}}$, and f_{V_x} represents the density of the vector $(V_{x_1}, \ldots, V_{x_d})$ induced by $U_{\mathcal{X}}$ and $T_{U,V;\mathcal{X}}$. The parameter m is the number of components of the mixture that are explicitly described, $Y_{xl} \epsilon (1, \ldots, m)$ is the component to which the l^{th} observation at covariate level x belongs, and W_{xl} is the value of the l^{th} observation at covariate level x. Hence $W_{xl} = \theta_{Y_{xl}x}$, provided $Y_{xl} \leq m$. The joint distribution of $\{W_{xl} | Y_{xl} > m\}$ is denoted by $g_s(\cdot)$. It is usually a density with respect to an unusual dominating measure, though this is of no concern for the straightforward uses of the result. This form facilitates calculation of conditional densities and/or ratios of densities.

The joint density of $(\theta_{ix}, v_{ix}, Y; i = 1, \ldots, m)$ is

$$\begin{aligned} f(\theta_{ix}, v_{ix}, Y_{xl}; i = 1, \ldots, m) &= [\prod_{i=1}^{m} f_{\theta_x}(\theta_{ix})][\prod_{i=1}^{m} f_{V_x}(V_{ix})][\prod_{x \epsilon (x_1, \ldots, x_d)} (\prod_{i=1}^{m} V_{ix} \prod_{j<i} (1 - V_{jx}))^{I(Y_{xl}=i)}] \\ &\cdot [\prod_{x \epsilon (x_1, \ldots, x_d} (\prod_{j=1}^{m} (1 - V_{jx}))^{\sum I(Y_{xl}>m)} \sum_{s} g(Y_{xl}) P(s|Y_{xl} > m)] \qquad (1) \end{aligned}$$

where this last expression depends only on the components of Y which are attached to components of the mixture beyond the first m. The W_{xl} associated with components among the first m are defined to equal the corresponding θ_{xl}.

The result allows us to consider two cases when writing expressions for the conditional distribution of Y_{xl}. The first case is for small values of Y_{xl}, where an observation is joining one of the first m components of the mixture. The expression for the conditional density (up to a constant of proportionality) is found for these components by considering the ratio of such expressions. The density for terms beyond the first m components cancel from these ratios. The second case corresponds to Y_{xl} joining a component beyond the first m. In this case, as the result will be used, there are no observations attached to components beyond the first m. Consequently, the conditional distribution of $(V_{ix}, \theta_{ix}, i > m$, will not depend on i, and the conditioning event will contain no information about which component (beyond the first m) will be joined. This leads to a simple algorithm for implementing conditional simulation techniques such as the Gibbs sampler.

3.1 A computational strategy

The distribution just presented is at the heart of a basic computational strategy. This strategy is a minor variant on Gelfand and Smith's (1990) Gibbs sampler. First, the algorithm, presented without data or hyperparameters.

Algorithm 3.1. At any given time, a collection of values of $\theta_{ix}, V_{ix}, W_{xl}$ and Y_{xl} are stored for $l = 1, \ldots, n_i$, for $x = x_1, \ldots, x_d$, and for some $i = 1, \ldots, m$. The value of m fluctuates throughout a simulation. In the sequel, *rest* refers to all parameters not appearing to the left of the conditioning bar.

0. Initialize the $\theta_{ix}, V_{ix}, W_{xl}$, and Y_{xl} in some state that is consistent with the data.

1. Generate $(\theta_{ix}, W_{xl})|rest$ for $x = x_1, \ldots, x_d$ and $l = 1, \ldots, n_i$. Update each coordinate immediately upon its generation.

2. Generate $(W_{xl}, Y_{xl})|rest$ for $x = x_1, \ldots, x_d$ and $l = 1, \ldots, n_i$. Update each pair immediately upon its generation. If $Y_{xl} > m$ for the current m, generate new components up to m.

3. Generate $V_{ix}|rest$ for $l = 1, \ldots, m$.

Repeat steps 1 through 3 a large number of times. Discard the early parameter vectors as burn-in and retain the rest for estimation.

Inclusion of data and hyperparameters follows in the usual fashion for Gibbs sampling methods for models based on the Dirichlet process. Details of the conditional densities from which observations are generated are provided in Appendix B.

There are many other useful sets of conditional distributions that can be written down, and that prove useful for computations, particularly when the model exhibits a great degree of conditional conjugacy. These simplifications of the model allow one to appeal to computational strategies designed to improve simulation for models based on a single Dirichlet process. In particular, the following "single p" model allows one both to use single Dirichlet process tricks for computation, and, as importantly, it allows one to often use code written for the models based on a single Dirichlet process.

3.2 The single p model

An important class of models is the "single p" dependent Dirichlet process. This class of models retains the great flexibility of the dependent Dirichlet process, yet leads to computational simplifications that render it attractive. The idea behind the model is that the mass, p_{ix}, associated with the location θ_{ix} does not vary with x. The implication of this restriction is that the overall model may be viewed as a countable mixture of stochastic processes, with the mixing weights matching those from a single Dirichlet process model. When $\mathcal{X}$ is finite, the distribution is, in fact, a single Dirichlet process. This match with a known process leads to an eventual simplification of computational strategies. But first, a definition and result for the single p model.

Definition 3.8 *The single p dependent Dirichlet process is defined by $F_{\mathcal{X}} \sim DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}})$, where $U_{i\mathcal{X}} = U_{ix_0}$, and where $T_{U,V;\mathcal{X}} = T_{U,V;x_0}$ for some x_0. These restrictions force $M_{\mathcal{X}} = M_{x_0}$.*

The formal specification of the single p model ensures that the masses do not vary with x. Since the $U_{\mathcal{X}}$ process does not vary with x and the transformations are all the same, the V_{ix} will not vary with x, and hence neither will the p_{ix}. A natural concern with this restriction is whether the resulting model has full support or whether the restriction has in some way reduced the support of the distribution. The next theorem provides conditions that ensure full support for the prior distribution.

Theorem 3.9 *Assume that $F_{\mathcal{X}} \sim DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}})$, with $\mathcal{X}$ either a countable set or an open set in $\mathcal{R}^k$. Also assume that the process $U_{\mathcal{X}}$ and the transformation $T_{U,V;\mathcal{X}}$ are such that $V_x = V_{x_1}$ for all $x\epsilon\mathcal{X}$. Take $x_1, \ldots, x_d$ to be a set of distinct points in $\mathcal{X}$. Define $Z_{x,d} = (Z_{x_1}^t, \ldots, Z_{x_d}^t)^t$ and $V_{x,d} = (V_{x_1}, \ldots, V_{x_1})^t$. Then, $(F_{x_1}, \ldots, F_{x_d}) \sim DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}})$. Assume that $Z_{x,d}$ induces a distribution on $\theta_{x,d}$ which is mutually absolutely continuous with $(F_{0,x_1}, \ldots, F_{0,x_d})$. Consider any $(F_1, \ldots, F_d)$ which is absolutely continuous with respect to $(F_{0,x_1}, \ldots, F_{0,x_d})$. For every $\epsilon > 0$, under the metric d, the ϵ-ball about $(F_1, \ldots, F_d)$ is assigned positive probability by the dependent Dirichlet process. Furthermore, for any $(F_1, \ldots, F_d)$ in the closure (under the metric d) of the set of distributions that are absolutely continuous with respect to $(F_{0,x_1}, \ldots, F_{0,x_d})$, for each $\epsilon > 0$, the ϵ-ball about $(F_1, \ldots, F_d)$ is assigned positive probability by the single p dependent Dirichlet process.*

Proof. Follow the construction for Theorem 3.3 given in Appendix A. Since the V_{ix} do not vary with x, the parameter M_x cannot vary with x. The condition of Theorem 3.3 on the support of the V_{ix} is, by definition, satisfied since the V_{ix} are i.i.d. Beta$(1, M)$ variates.

The single p models do provide a restriction on the dependent Dirichlet process, preventing one from obtaining independent F_i at distinct levels of the covariate. To see this, merely note that the p_i are the same for all levels of the covariate. Hence if F_i contains a large lump of mass at a point, all of the F_j contain the same large lump of mass.

Computations for the single p model follow from those for the Dirichlet process. With these strategies, one need only track which observations are tied to which component process, dispensing with labels for the component processes and avoiding explicit generation of the p_i. This results in algorithms with superior performance, and it also simplifies the task of programming the algorithms. The computations for the single p models are based on the following result.

Theorem 3.10 *Assume that $F_{\mathcal{X}}$ follows a single p dependent Dirichlet process with parameters $(M, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U, T_{Z,\theta;\mathcal{X}}, T_{U,V})$. Restrict consideration to a finite set of distinct values for x, say $(x_1, \ldots, x_d)$. Define $\tilde{\theta} = (\theta_{x_1}, \ldots, \theta_{x_d})$, and let F_0 represent the distribution of $\tilde{\theta}$, induced by $F_{0,\mathcal{X}}, Z_{\mathcal{X}}$, and $T_{Z,\theta;\mathcal{X}}$. Then the joint distribution of $(F_1, \ldots, F_d)$ follows a Dirichlet process with mass M and base distribution F_0.*

The theorem enables one to perform simulations based on a single Dirichlet process model. The computational strategies for these models are now well developed. See Escobar (1994), Escobar and West (1995), MacEachern (1994), Bush and MacEachern (1996), MacEachern and Müller (1998), MacEachern (1998), Neal (1998), and Green and Richardson (1997) for computational techniques and tips on how to tailor the computational strategy to the problem at hand.

The beauty of the single p models is that they preserve the essence of the problem, providing a full nonparametric prior distribution at each level of the covariate while ensuring local dependence of the realized distributions. Yet they retain great simplicity for prior specification and computation. They thus provide a general framework that covers a vast territory.

3.3 Connections

Dependent Dirichlet processes form a very general class of models. They include a number of useful subclasses, such as a single Dirichlet process or a collection of independent Dirichlet processes, either explicitly or as limiting cases. This section describes a number of connections to other models. It also indicates how the development of other dependent nonparametric processes can be accomplished.

The dependent Dirichlet process model can be used to obtain a single random distribution that follows a Dirichlet process. To do so, one need only eliminate the dependence of the stochastic processes on x. This is accomplished by taking $F_{\mathcal{X}} = F_{x_0}$ and by choosing $M_{\mathcal{X}} = M_{x_0}$, $F_{0,\mathcal{X}} = F_{0,x_0}$, $Z_{\mathcal{X}} = Z_{x_0}$, $U_{\mathcal{X}} = U_{x_0}$, $T_{Z,\theta;\mathcal{X}} = T_{Z,\theta;x_0}$ and $T_{U,V;\mathcal{X}} = T_{U,V;x_0}$ for some $x_0 \epsilon \mathcal{X}$.

Pursuing the notion of a single nonparametric distribution to drive an analysis, one sees that this deterministic link lets us describe collections of distributions such as an arbitrary scale family. To do this, the random F_{x_0} serves as the base distribution for the family. This distribution follows a Dirichlet process, and the $\theta^*_{ix_0}$ are i.i.d. draws from F_{0,x_0}. The θ^*_{ix} with scale $s(x)$ are, deterministically, $\theta^*_{ix} = \theta^*_{ix_0} s(x)/s(x_0)$. They are i.i.d. from $F_{0,x}$. The masses are calculated from the draws V_{ix_0}, with $V_{ix} = V_{ix_0}$. Together, these features imply that the family $F_{\mathcal{X}}$ is a scale family, with scale parameter $s(x)$. Such distributions are of use in problems where a classical statistician might use weighted least squares instead of least squares. Alternatively, one can model the single nonparametric distribution as a Dirichlet process and also model the scale parameter explicitly, yielding the identical model for the data.

Extending this line of thought to problems where one would consider fitting a generalized linear model is not as straightforward without the dependent Dirichlet process. In these settings, the single nonparametric distribution becomes a collection of nonparametric distributions indexed by the covariate. The baseline distributions change with the covariate, and so the realized distributions must change as well. The change is not typically anything as simple as a scale change. For example, the distributions may be Poisson distributions which change as their mean changes. To fit this type of model into the single distribution, dependent Dirichlet process framework, the deterministically linked set of Dirichlet processes is obtained by allowing $F_{0,\mathcal{X}}$ and $T_{Z,\theta;\mathcal{X}}$ to vary with x while maintaining $Z_{\mathcal{X}} = Z_{x_0}$, $U_{\mathcal{X}} = U_{x_0}$, $T_{Z,\theta;\mathcal{X}} = T_{Z,\theta;x_0}$ and $T_{U,V;\mathcal{X}} = T_{U,V;x_0}$. This approach allows one to pursue models similar to those in Newton, Czado and Chappell (1996), Mukhopadhyay and Gelfand (1997), Kleinman and Ibrahim (1998), Carota and Parmigiani (1997) and Dominici and Parmigiani (1998) but which allow some dependence among the distributions at various levels of the covariate.

The parametric model corresponding to $F_{0,\mathcal{X}}$ is most easily obtained by taking the limit of the deterministically linked dependent Dirichlet processes as the mass parameter, M, tends to ∞. Since this model is a limit of dependent Dirichlet processes, one typically obtains inferences that tend to those of the parametric model as the limit of posteriors is taken.

The far extreme from complete dependence is independence. To obtain a collection of independent Dirichlet processes, simply take $(Z_{ix_1}, \ldots, Z_{ix_k}, U_{ix_1}, \ldots, U_{ix_k})$ mutually independent for every distinct k-tuple $(x_1, \ldots, x_k)$. The mutual independence of the variates used to construct $F_1, \ldots, F_k$ ensures that the realized distributions are independent.

The great strength of the dependent Dirichlet process lies between the extremes of complete dependence and independence. The strength of the dependence in the realized F_i will be governed by parameters of the model. If the driving stochastic processes exhibit strong dependence over the range of covariate levels in use, then the realized F_i will be strongly related to each other, and the realized distributions will slowly evolve as the covariate changes. If the dependence of the stochastic processes is weak, then the realized F_i will be only weakly related to each other, and the realized distributions will show substantial, quick changes as the covariate changes. The parameters that govern the strength of this dependence are either specified as part of the prior distribution, or one may place a distribution on the parameters.

More generally, the technique used to construct the dependent Dirichlet process can be used to construct other dependent nonparametric distributions. The general approach to construction of a prior distribution with full support over the space of distribution functions is to take a countable collection of mutually independent random variables, and to use these variates to describe a distribution function. For conciseness, the distribution function is presumed to be on the unit interval, although the Dirichlet process uses the θ_i and V_i to construct a discrete distribution (Sethuraman, 1994). Finite mixture models with no upper bound on the number of support points are typically constructed via a variate that determines the number of support points, and, conditional on that variate, a collection of θ_i and p_i that produce a discrete distribution. The ϵ-Dirichlet process (Tardella and Muliere, 1998) provides a means of approximating a Dirichlet process with a finite mixture model. The Polya tree first fixes a partition of the unit interval, and then uses a collection of mutually independent beta random variates to describe the random distribution (Ferguson, 1973; Lavine, 1992, 1994). The generalized Polya tree fixes a partition of the unit interval and then uses a collection of mutually independent variates (that may or may not be beta random variates) to describe the random distribution (Walker and Muliere, 1997). Series expansions provide a representation of a density as the normalized sum of random coefficients times basis functions, perhaps exponentiated. The prior distribution can usually be described by a variate that determines the number of non-zero coefficients in the series and a variate that determines the coefficient for each term in the series (for example, Petrone, 1999a, 1999b). These cases represent but only a few of the approaches to placing a distribution on the space of distribution functions. Other approaches that are computationally accessible include hybrid models. For example, a hybrid model for a discrete distribution might rely on explicit description of the first few components, as per a finite mixture model, followed by a tail of the remaining components determined by the Dirichlet process. In each of these cases, and in many other settings, a collection of dependent nonparametric distributions can be created by merely replacing each random variate by a stochastic process indexed by a covariate, as has been done here for the dependent Dirichlet process. Computations for many of these models, such as the finite mixture models or the hybrid models follow the general strategies provided herein.

4 Regression Modelling

The dependent Dirichlet process (DDP) model provides an alternative to the usual normal theory linear model, having the advantages of allowing non-normal error distributions and also of allowing the error distribution to evolve as the level of the covariate changes. This section analyzes a set of data from Weisberg (1985) which was provided to him by S.J. Gould. The data analyzed here consist of measurements of perimeter and area of 24 Romanesque churches (Birkin has been removed from the original data set). Of interest is the relationship between area and perimeter, as well as the predictive distribution of area for Romanesque churches not included in the analysis. Following Weisberg's discussion, we examine the data after both perimeter and area have been subjected to log transformations.

Consider a least squares analysis and the accompanying residual plots. The residuals from such an analysis indicate a lack of fit of the normal theory linear model: plots of log area against log perimeter and of residuals against fitted values show (to my eye) some lack of fit, particularly for the small churches; a normal probability plot of the residuals shows a distinctly non-linear pattern; boxplots of the residuals, splitting the churches evenly into two groups based on perimeter, show a change in the skewness of the residual distribution. Figure 1 provides some of these plots. All three of these features suggest that one use a DDP model when analyzing the data.

The structure of the DDP model follows, with x representing the centered log perimeter of a church and y representing the log area of a church.

$$\begin{aligned}
F_{\mathcal{X}} &\sim DDir(M, N(0,\sigma_\theta^2), G(c_0, c_e, a_e), \cdot, \cdot, \cdot) \\
\beta &\sim N(\mu_\beta, \Sigma_\beta) \\
\theta_i &\sim F_{x_i} \\
\epsilon_i &\sim N(0, \sigma_e^2) \\
Y_i &= \beta_0 + \beta_1 x_i + \theta_i + \epsilon_i.
\end{aligned}$$

This specification of the model makes use of the convention that only non-standard components are specified and that, when a subscript is omitted, the parameter does not vary with the subscript. Thus, the DDP has a constant value of M, the stochastic process for $U_{\mathcal{X}}$ is constant in $\mathcal{X}$ and is just a Beta$(1, M)$ variate, and the transformation from U to V is the identity transformation. The stochastic process for $Z_{\mathcal{X}}$ is a Gaussian process with an exponential variogram having parameters c_0, c_e and a_e, as described in Cressie (1993). The identity transformation takes Z to θ. The prior distribution for β is normal with the specified mean and variance.

The prior for the DDP model was constructed in several stages. The prior for the regression coefficients, β_0 and β_1, was chosen to have a mean vector of $(2.307, 2)$. The second component of the prior mean corresponds to the notion that the churches are similar in shape, with larger churches scaled up versions of smaller churches. In the absence of any theory, the first component was chosen to match the value of $\hat{\beta}_0$ in the least squares analysis. The variance matrix for β was selected to provide little information about the values of the parameters. The least squares analysis provides an information matrix for the vector β. To ensure that the prior on β would play a minimal role in the analysis, the prior was chosen to have a ratio of prior precision to information of $1:100$. This led to a prior variance matrix for β of diag$(0.0757, 0.1439)$.

Relying on the least squares analysis, the mean square error was used to provide an estimate of σ^2. This value, 0.01817, was then split into two components–σ_e^2 and σ_θ^2. In keeping with the arguments used for kernel density estimation, σ_e was set to $1.06 \cdot 0.1348 \cdot 24^{-1/5} = 0.75675$. The variance of an observation, given β, is $\sigma_e^2 + \sigma_\theta^2$. Setting this to 0.01817 yields $\sigma_\theta = 0.11155$.

Several choices were tried for the DDP prior distribution. All relied on a mass parameter of 10 and the Gaussian process with an exponential covariance function described above. The parameters of the covariance function were chosen to enforce continuity of the sample paths of the stochastic processes for θ. Values of the correlation parameter were selected to produce a correlation between the θ_{ix} values for θ_{ix} located at the quartiles of the data set of $0.5, 0.8, 0.9$ and 0.99. The analyses based on these prior distributions are contrasted below.

A Gibbs following the computational strategy of Bush and MacEachern (1996) sampler was used to fit the single p DDP model. A burn-in period of $1,000$ scans was used, followed by an estimation period of $20,000$ scans. The output of the estimation period was systematically subsampled at a rate of 1 in 10, to reduce the need for storage space. Table 1 provides a summary of least squares fits and DDP fits of the model. Since the parameters β_0 and β_1 are not identifiable in the DDP model, the reported values of β_0 and β_1 have been computed by ordinary least squares so as to provide a linear inference for the model. See discussion of the noise floor in the next subsection.

Figure 2 presents pairs of posterior predictive distributions for log areas of churches not included in the data set. The predictive distributions are Rao-Blackwellized estimates obtained from the Gibbs sampler run. The change in skewness of the predictive distributions, from a positive skewness for small churches to a negative skewness for large churches is typical of predictive densities for other levels of the covariate. This feature of the data is easily missed with a standard regression analysis. The predictive distributions exhibit less variability as the correlation parameter grows. This is to be expected since the greater the correlation, the more similar are values of θ_{ix_j} as x_j varies. A greater similarity of these values effectively implies a greater pooling of information about the unknown distribution, and will generally result in tighter predictive distributions. Similarly, we might expect the reduction in the change in skewness that we observe, as the greater correlation parameter expresses greater similarity between the distributions F_x for different values of x. In all cases, the distinctive, evolving nonparametric character of the predictive distributions indicates the need for the DDP analysis.

4.1 The noise floor

The church data was examined, in part, to see whether a theory of shape developed for living organisms would extend to other classes of objects. The theory, alluded to in Gould (1973), suggests that β_1 should be smaller than 2. As with any theory of this sort which ignores "small order" terms, we would not expect the churches to obey a straight line law exactly, and would even be unlikely to believe that the churches would obey a straight line plus mean zero error model. Instead, my actual belief would encompass a richer class of models, where the simplification to a line plus mean zero error formulation is used as a summary of the richer distribution. Walker and Gutierrez-Pena (1999) suggest a strategy for problems that are fundamentally nonparametric in nature. They recommend that one model the data in the larger, nonparametric space, but that inference be made in the restricted space. Applying this approach to the regression setting, we fit the DDP regression model to the data. After fitting the model to the data, we seek the best linear description of the fit of the model.

The core elements of the regression problem are the mean response given the covariate, the variability of the response given the covariate, and also the marginal distribution of the covariate. This marginal distribution is typically ignored in a regression analysis, for, when the linear model is presumed to hold (i.e., when one works in the smaller space), the marginal distribution is irrelevant for the *definition* of β. However, when the model is not linear in x, the marginal distribution on x affects the very definition of β. With straightforward notation, a population version of the least squares fit minimizes the expression $\int(\mu_{y|x} - \beta_0 - \beta_1 x)^2 m(x)dx$ over β_0 and β_1. Call the minimizing vector β_m. Unless $\mu_{y|x}$ is linear in x, β_m depends on m. Such dependence implies that, even asymptotically, there is no single best choice of β_m: there is a noise floor below which one cannot pass.

To obtain a linear fit to the DDP model, a marginal distribution was chosen for x to match the observed distribution of the covariate. At this point, a line was fit to the posterior predictive means,

$$\hat{\mu}_{x_i} \quad = \quad E[\beta_0 + \beta_1 x_i + E[\theta|x_i]|(Y_1, \ldots, Y_n)]$$

with a weighted least squares technique. The weights were chosen to equal the posterior precision of the predictive distribution for churches not appearing in the data set. The weighted fit is provided by the first line of Table 2. Alternatively, ordinary or generalized least squares could have been used to produce a linear fit to the DDP model.

A great benefit of recasting a traditional parametric analysis into a nonparametric modelling framework followed by a parametric inference is that one can assess limits on the precision of the parametric fit. To illustrate this point, we consider a collection of linear inferences made on the basis of the DDP regression model with a correlation of 0.99. A class of marginal distributions was created by assigning weight 1/12 to each of a dozen successive cases in the ordered (by x) data. For each marginal distribution, a line was fit with weighted least squares. Table 2 summarizes these fits. The largest fitted slope was 1.69 while the smallest fitted slope was 1.46. Even if more data were available, we would be unlikely to have the ability to distinguish between slopes in a range of roughly 0.23. This range gives us an idea of the noise floor for the inference we intend to make.

It should be noted that further Bayesian formalization of this problem, not pursued here, can also place a distribution on $m(x)$. Adding a loss function and applying Bayesian reasoning creates a fully Bayesian approach to making inference about the linear summary of a potentially non-linear model. Loosely, for models that are nonlinear, where the distribution on $m(x)$ does not change as more data is collected, the posterior distribution on the optimal β_m remains nondegenerate, even as the sample size tends to ∞.

5 A prior for a continuous cdf

Bayesians often rely on qualitative judgements to simplify prior elicitation, demanding, for example, that a distribution have tails that decay at a particular rate or that the posterior mean of a parameter be a monotone function of some statistic. These qualitative judgements, though rarely a complete representation of one's prior beliefs, must be reflected in the prior distribution. In nonparametric Bayesian analysis, a common qualitative judgement is that an unknown distribution is continuous. The mathematical question then becomes how to construct a prior distribution that assigns probability one to the set of continuous distributions. The practical question is how to solve the mathematical problem within computational constraints and without introducing artificial or unappealing features to the prior.

Polya trees and their generalizations (Lavine, 1992, 1994; Walker and Muliere, 1997) provides one means of specifying a prior on the space of continuous distributions. Cleverly creating a family of Polya trees can lead to computational improvements, as in Berger and Guglielmi (1999). The down-side of the Polya tree solution is that one is forced to specify a tree structure as part of the prior distribution. The tree structure introduces an element of artificiality into the problem.

Dependent Dirichlet processes provide a second solution to the problem. An unobserved covariate is introduced into the model. The distribution F_x follows some dependent Dirichlet process, and a continuous distribution, G, is placed on X. The end result is that, while F_x may be discrete for each x, the marginal $F_Y(y) \overset{def}{=} \int F_x(y)dG(x)$ is a continuous distribution. In situations where the unobserved distribution of X represents heterogeneity of experimental units, this model provides a plausible mechanism for producing the distribution as well as a mathematical solution to the problem. This approach easily extends to situations where covariates are measured with error, allowing a novel Bayesian approach to errors in variables models wherein the conditional distributions of the response variables are nonparametric and continuous for each *measured* value of the explanatory variables. The following theorem provides a set of conditions that guarantee that the distribution of F, marginalized over X, is continuous. The conditions are sufficient, but not necessary.

Theorem 5.1 *Assume that $F_{\mathcal{X}} \sim DDir(M_{\mathcal{X}}, F_{0,\mathcal{X}}, Z_{\mathcal{X}}, U_{\mathcal{X}}, T_{Z,\theta;\mathcal{X}}, T_{U,V;\mathcal{X}})$, where $\mathcal{X}$ is an interval of $\mathcal{R}^1$, and where $F_{0,x}$ is a continuous distribution for each $x \epsilon \mathcal{X}$. Further assume that the stochastic process $Z_{\mathcal{X}}$ and the transformation $T_{Z,\theta;\mathcal{X}}$ are such that the distribution of $\theta_{ix'}|(\theta_{ix_1}, \ldots, \theta_{ix_n})$ is continuous, provided $x' \ni (x_1, \ldots, x_n)$, for every $x', x_1, \ldots, x_n \epsilon \mathcal{X}$. Assume that $X_1, X_2, \ldots$ form a random sample from a continuous distribution, say G, on $\mathcal{X}$. Then, the distribution defined by $F_Y(y) \overset{def}{=} \int F_x(y)dG(x)$ is continuous with probability* 1.

Proof. To show that the distribution F_Y is continuous, it suffices to show that $P(Y_{n+1} \epsilon (Y_1, \ldots, Y_n)) = 0$.

$$P(Y_{n+1} \epsilon (Y_1, \ldots, Y_n)) \leq \sum_{i=1}^{n} P(Y_{n+1} = Y_i)$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} P(Y_{n+1} = Y_i | X_{n+1} = X_1) P(X_{n+1} \epsilon (X_1, \ldots, X_n)) \\
&\quad + P(Y_{n+1} = Y_i | X_{n+1} \neq X_1) P(X_{n+1} \ni (X_1, \ldots, X_n)) \\
&= 0,
\end{aligned}
$$

because $P(X_{n+1} \epsilon (X_1, \ldots, X_n)) = 0$ follows from the assumption that the X_i form a random sample from a continuous distribution and since, with $X_{n+1} \ni (X_1, \ldots, X_n)$, the distribution of $\theta_{j,X_{n+1}} | (\theta_{j,X_1}, \ldots, \theta_{j,X_n}$ is continuous for each j.

6 Conclusions

Ferguson's (1973) early paper on nonparametric Bayesian methods lays out the Dirichlet process, thus describing a means of placing a distribution on the space of distribution functions. The value of being able to work in a large space for distributions, rather than a tightly constrained parametric family is immediately obvious: it enables one to model departures from, say, normality that are both routine (such as a bit of skewness) and that are non-routine (such as multimodality). Ferguson's paper indicates a wide variety of uses for nonparametric Bayesian techniques and this portion of his paper is so convincing that it has led to a large body of work both on theoretical properties of the models and on their practical uses. The subsequent development of computational techniques for fitting complex models that involve a nonparametric Bayesian component has extended their use to pieces of essentially any Bayesian model, and the success of these methods is well documented across a range of applications (see, for example, West, Müller and Escobar, 1994, the papers in Dey, Müller and Sinha, 1998, and Walker, Damien, Laud and Smith, 1999). Typically, the additional flexibility of a nonparametric form allows one to better model the data, and this superior modelling capability translates directly into better out of sample predictions.

Dependent nonparametric processes provide the next logical step in the development of Bayesian models. They allow us to move beyond a single unknown distribution, to enter the realm of dependent nonparametric distributions. The approach pursued here, with the development of the dependent Dirichlet process, allows us to work with a continuous covariate space, where the covariate indexes the random distribution. This covariate space can be of quite general form, with the results here proven for open sets of $\mathcal{R}^k$. As indicated briefly in Section 3.3, once the processes are defined, there is an abundance of natural modelling strategies. To focus on the normal theory linear model, we have several immediate generalizations: The first is to replace the normal distribution for errors with a nonparametric error distribution, as can be done with a single mixture of Dirichlet processes model. Alternatively, the normal error distribution can be replaced with an evolving error distribution through use of a mixture of dependent Dirichlet processes, as illustrated in Section 4. The index through which dependence among nearby error distributions is created in Section 4 is the covariate itself. In higher dimensional problems, either the entire covariate space can index the dependence or a lower dimensional summary of the covariate space (such as the mean for the response) can index the dependence. While it is natural to write down models which allow some non-linearity in the relationship, one can also write down mixture of dependent Dirichlet process models which retain linearity of the response's mean in the covariates. All of these generalizations allow for incorporation of a scale parameter in the error distribution. Incorporation of such a parameter enables one to model heteroscedasticity of the errors. Similar extensions allow one to generalize the generalized linear model.

The phenomenon of evolving error distributions is one which appears, to me, to be quite prevalent. In order to create accurate predictive densities, it is essential to work with models that allow the error distribution to evolve. The dependent Dirichlet process provides a simple, natural approach to modelling evolving distributions. In many situations where the error distributions change shape as the covariate (or as the mean of the response) changes, there also appear to be mild non-linearities in the regression. These non-linearities may be mild enough, as they seem to be for the church data, that one still wishes to work with a conceptual, linear summary of the regression. By modelling the data in a larger space that allows for some non-linearity, but by making an inference in the smaller (linear) space, one is led to the notion of a noise floor and the need to acknowledge that one may never be able to distinguish between some collections of parameter values. The notion of a noise floor applies in many situations, and is just as applicable in classical as Bayesian settings. Further development of the concept of a noise floor and its application should help to clarify the importance (or lack thereof) of retaining a linear regression when fitting extensions of the linear model.

The applications of dependent nonparametric processes mentioned to this point are heavily used in applied statistics, and these alone would justify study of the techniques. However, the impact of the processes is not limited to these settings. Dependent Dirichlet processes (or other dependent nonparametric processes) can be used as pieces of more complex Bayesian models in a relatively straightforward way. The hierarchical modelling technique which drives the modern Bayesian models splits a large model up into more manageable components and then links these components together in relatively simple fashions. Dependent nonparametric processes provide potential replacements for components of the hierarchical model, allowing the components to interact in a more complex manner while still retaining the relatively simple links between components. The techniques will find wide use in the future.

7 References

BERGER, J.O. AND GUGLIELMI, A. (1999). Bayesian testing of a parametric model versus nonparametric alternatives. *Technical Report*, **99-04**, Institute of Statistics and Decision Sciences, Duke University.

BLACKWELL, D. (1973). Discreteness of Ferguson selections. *The Annals of Statistics*, **1**, 356–358.

BLACKWELL, D. AND MACQUEEN, J.B. (1973). Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, **1**, 353–355.

BUSH, C.A. AND MACEACHERN, S.N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, **83**, 275–285.

CAROTA, C. AND PARMIGIANI, G. (1997). Semiparametric regression for count data. *Technical Report*, **99-17**, Institute of Statistics and Decision Sciences, Duke University.

CRESSIE, N.A.C. (1993). *Statistics for Spatial Data, revised edition.* New York: Wiley.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates (with discussion). *The Annals of Statistics*, **14**, 1–67.

Dey, D., Müller, P., and Sinha, D. eds. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics.* New York: Springer-Verlag.

Dominici, F. and Parmigiani, G. (1998). Bayesian semi-parametric analysis of toxicology data. *Technical Report*, **98-09**, Institute of Statistics and Decision Sciences, Duke University.

Escobar, M. (1988). Estimating the means of several normal populations by estimating the distribution of the means. Ph.D. thesis, Yale University.

Escobar, M. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, **89**, 268–277.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–588.

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

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 378–409.

Gould, S.J. (1973). The shape of things to come. Systematic Zoology, **22**, 401–404.

Green, P. and Richardson, S. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 731–792.

Kleinman, K.P. and Ibrahim, J.G. (1998). A semiparametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine*, **17**, 2579–2596.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, **20**, 1203–1221.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, **22**, 1161–1176.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 1–42.

MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics B: Simulation and Computation*, **23**, 727–741.

MacEachern, S.N. (1998). Computations for MDP models. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey and P. Müller and D. Sinha eds.), 23–43, New York: Springer-Verlag.

MacEachern, S.N. (1999). "Dependent nonparametric processes," in *ASA Proceedings of the Section on Bayesian Statistical Science.*

MacEachern, S.N. and Muller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**, 223–238.

Mukhopadhyay, S. and Gelfand, A.E. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association*, **92**, 633–639.

Müller, P., Qintana, F. and Rosner, G. (1999). Hierarchical meta-analysis over related nonparametric Bayesian models. *Technical Report*, **99-22**, Institute of Statistics and Decision Sciences, Duke University.

Neal, R.M. (1998). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, to appear.

Newton, M.A., Czado, C. and Chappell, R. (1996). Semiparametric Bayesian inference for binary regression. *Journal of the American Statistical Association*, **91**, 142–153.

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics*, **27**, 105–126.

Petrone, S. (1999b). Random Bernstein polynomials. *Scandinavian Journal of Statistics*, **26**, 373–393.

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

Smith, A.F.M. (1973). Bayes estimates in one-way and two-way models. *Biometrika*, **60**, 319–329.

Tardella, L. and Muliere, P. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics*, **26**, 283–297.

Tomlinson, G. and Escobar, M. (1998). Analysis of densities. *Technical Report*, University of Toronto.

Walker, S.G., Damien, P., Laud, P.W. and Smith, A.F.M. (1999). Bayesian nonparametric inference for distributions and related functions (Discussion: 510–527). *Journal of the Royal Statistical Society, Series B*, **61**, 485–509.

Walker, S.G., Gutierrez-Pena, E. (1999). Robustifying Bayesian Procedures (with discussion). In *Bayesian Statistics 6* (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.), 685–710. New York: Springer-Verlag.

Walker, S. and Muliere, P. (1997). Beta-Stacy processes and a generalization of the Polya urn scheme. *The Annals of Statistics*, **25**, 1762–1780.

Weisberg, S. (1985). *Applied Linear Regression, second edition.* New York: Wiley.

West, M., Müller, P., and Escobar, M. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty: A tribute to D.V. Lindley* (A.F.M. Smith and P. Freeman eds.), 363–386, New York: Wiley.

8 Appendix A

8.1 Proof of Theorem 3.3

The distance measure used in the proof is $d(F,G) = \sum_{i=1}^{d} d_L(F_i, G_i)$, the sum of the d Levy distances across the distributions. It is easily verified that this is still a distance measure. Since $d(F,G) \leq d \cdot max_{1 \leq i \leq d}(d_L(F_i, G_i))$, proximity under this distance measure implies proximity under Levy distance for each of the distributions.

We first focus on the case of $d = 1$, with F a p-dimensional distribution. The proof has two parts. The first part shows that there exists a suitable discrete approximation, say G, of any F in the set described in the statement of the theorem. Such a G lies within $\epsilon/2$ of F. The second part shows that the prior assigns positive probability to the $\epsilon/2$-ball about G. Finally, the triangle inequality is used to show that the prior assigns positive probability to every ϵ-ball about F.

Part one. Suitable approximation of F.

Find some compact set $C = [-c, c]^p$ for which $P_F(x \epsilon C) > 1 - \epsilon/4$, and where c is a multiple of $\epsilon/4$. Partition this set into disjoint hypercubes of side $\epsilon/4$, with, say, the convention that the north and east surfaces of a hypercube are included in it and the south and west surfaces are not. Define G to be the finite, discrete distribution that assigns the mass in each hypercube from F to the midpoint of the hypercube. Assign the mass from F that lies outside the set C to the point $(-c-\epsilon/8, \ldots, -c-\epsilon/8)$. Verification of the inequalities defining the Levy metric ensure that $d(F,G) < \epsilon/2$.

Part two. Prior probability near G.

The region C is comprised of $N = (8c/\epsilon)^p$ hypercubes. Consider a distribution H which assigns to each hypercube a mass within $\epsilon/(4N)$ of what G assigns to that hypercube. Verification of the Levy inequalities implies that $d(G,H) < \epsilon/2$. Consider the probability assigned to the set of such H, say $\mathcal{H}$. The next paragraph shows that this probability is positive.

By assumption, the distribution of F is absolutely continuous with respect to $F_{0,d}$. The locations are a sequence of i.i.d. draws from $F_{0,d}$, and so the distribution of each θ_i^* assigns positive probability to each hypercube where F assigns mass, and hence where G assigns mass. Thus, each finite sequence assigning $\theta_1^*, \ldots, \theta_n^*$ to the hypercubes where G has mass has positive probability. This part of the argument is completed by focusing on the u_i. The u_i are transformed into v_i and thence into p_i. Define p_{jH} to be the mass assigned by some H to hypercube j. Let p_{0H} represent the mass of H that lies outside of the set C. Define p_{jG} in a similar fashion. It is sufficient to show that $|p_{jH} - p_{jG}| < \epsilon/(4N)$ for $j = 0, \ldots, N$. This is accomplished by ordering the hypercubes and then creating a finite set of small lumps of mass that nearly sum to the desired p_{iG} in each hypercube. These lumps of mass account for nearly all of the mass of H. Formally, choose a collection of large enough n_j, set $n = \sum_{j=0}^{N} n_j$, and choose some $\delta_L < \delta_U$ sufficiently close together and sufficiently small that, provided $\delta_L < v_i \leq \delta_U$ for $i = 1, \ldots, n$, the induced $|p_{jH} - p_{jG}| < \epsilon/(4N)$ for $j = 0, \ldots, N$. The sequence of locations that correspond to these first n lumps of mass in the distribution are independent of the u_i, and are i.i.d. draws from $F_{0,d}$, and so have positive probability of the first n_0 lying outside the set C, the next n_1 lying in hypercube 1, and so on. Since the events of the p_{iH} and θ_i^* lying in the desired intervals and region of R^p are independent and each has positive probability, we have that $P(d(H,G) < \epsilon/2) > 0$. Application of the triangle inequality yields $d(F,G) + d(G,H) \leq d(F,H)$, implying the result.

The case of $d > 1$ is made difficult by the fact that the u_i, hence the v_i and so eventually the p_i may vary from one distribution to another. This prevents us from stringing together the d distributions to form a single distribution in pd dimensions. Instead, the proof parallels that of the case $d = 1$. For the first part of the proof, a suitable approximation to F_i is found such that $d_L(F_i, G_i) < \epsilon/(2d)$. To do this, a cube C_i is set up for each of the d distributions, and the previous part one argument is repeated. For the second part of the proof, a repetition of the argument relies on the d locations $(\theta^*_{i1}, \ldots, \theta^*_{id})$ simultaneously falling into appropriate hypercubes. The conditions of the theorem ensure that this can happen with positive probability. The argument also relies on the existence of a large, finite collection of v_{ij} that enable the creation of masses for the hypercubes that lie very close to the hypercube masses of the approximating discrete G_i. Again, the conditions of the theorem ensure that we can find δ_L and δ_U small and sufficiently close together that the p_{jH_i} and the p_{jG_i} are uniformly close. Independence of the Z_X and the U_X ensure that each of the neigborhoods $\mathcal{H}_i$ receive positive probability. To complete the proof, $d(F, H) = \sum_{i=1}^{d} d_L(F_i, H_i) < \epsilon$ for all $H \epsilon \mathcal{H}$, ensuring that the ϵ-ball about F is assigned positive probability.

9 Appendix B

Generation of $(\theta_{ix}, \{W_{xl} | Y_{xl} = i\}) | rest$ follows from the joint density in equation (1). Use the density with a value of $m > Y_{xl}$ to obtain the density from which to generate. Since conditioning on $Y_{xl} = i$ ensures that $W_{xl} = \theta_{ix}$, the density can be expressed in a form that does not explicitly involve W_{xl}. For the basic model this means either a single θ_{ix} is generated from

$$f_{\theta_{ix}} = f_{\theta_{ix}}(\cdot | \theta_{ix'}), \text{ conditioning on all } x' \neq x \tag{2}$$

or the vector of θ_{ix}, with i fixed, are generated from

$$f_{\theta_{ix}} = f_{\theta_{ix}}(\theta_{ix}).$$

In the first case, the W_{xl} for which $Y_{xl} = i$ are set to the new θ_{ix}. In the latter case, the W_{xl} for which $Y_{xl} = i$ are set to θ_{ix}, where x ranges over $x_1, \ldots, x_d$.

Generation of $(Y_{xi}, W_{xi})|rest$ follows from the joint density in equation (1). This is accomplished in two stages, first generating Y_{xi} and then generating W_{xi}. In the first stage, generate $s \sim U(0,1)$. Set

$$Y_{xi} \quad = \quad min\{l| \sum_{j=1}^{l} p_{jx} \geq S\},$$

where $p_{jx} = V_{jx} \prod_{k<j}(1 - V_{kx})$. If $\sum_{j=1}^{m} p_{jx} < S$, then generate enough additional components as $V_{m+k,x} \overset{i.i.d.}{\sim} Beta(1, M_x)$ variates until Y_{xi} is defined, and change the value of m to this larger value. Retain the newly generated V_{ix}. Supplement these by generating values for $V_{ix'}$, $x' \neq x$, and values for θ_{ix}, $x \epsilon x_1, \ldots, x_d$. The $V_{ix'}$ are generated from the density for $V_{ix'}|V_{ix}$, and the θ_{ix} are generated from $f_{\theta_{ix}}$. In the second stage, set $W_{xi} = \theta_{Y_{xi}x}$.

Generation of $V_{ix}|rest$ follows from the joint density in equation (1), with m taken to exceed i. The density is proportional to

$$f_{v_x}(V_{ix}|V_{ix'}) \prod_{l=1}^{n_x} V_{jx}^{I(Y_{xl}=j)} \prod_{l=1}^{n_x} (1 - V_{jx})^{I(Y_{xl}>j)}.$$

Dependent Dirichlet processes find use in more complex models. In these settings, the W_{xl} often represent parameter vectors indexing parametric distributions from which data values are drawn. This portion of the model impacts the distribution of θ_{xi}. Multiply (x.x) by the likelihood of all observations for which $Y_{xl} = i$. It also impacts the distribution of Y_{xi}. Letting D_{xl} represent the data attached to W_{xl}, and letting $g(D_{xl})$ represent its density, the expressions become

$$f_{\theta_{ix}} \quad = [f_{\theta_{ix}}(\theta|\theta_{ix'})][\prod_{l=1}^{n_x} g(D_{xl}|\theta_{ix})^{I(Y_{xl}=i)}]$$

and

$$\begin{aligned} Y_{xj} \quad & \propto \quad [V_{jx} \prod_{k<j}(1 - V_{kx})][g(D_{xi}|\theta_{jx})] \\ & \qquad \text{for } j = 1, \ldots, m, \text{ and} \\ & \propto \quad [[\prod_{k \leq m}(1 - V_{kx})][\int g(D_{xi}|\theta_{jx}) f(\theta_{jx}) d\theta_{jx} \\ & \qquad \text{for the event that } j > m, \end{aligned}$$

with the constraint that $\sum_{j=1}^{\infty} p_{jx} = 1$. As before, if $j > m$, generate enough additional components as $V_{m+k,x} \overset{i.i.d.}{\sim} Beta(1, M_x)$ variates until Y_{xi} is defined, update m, and generate values for the new θ_{ix}. The expression for the conditional density of V_{ix} remains unchanged.

A further portion of the model often involves a distribution over the hyperparameters that control the dependent Dirichlet process. In an MCMC simulation, these parameters would also be updated, conditional on m and $\theta_{ix}, i = 1, \ldots, m$. These updates are standard, with teh model passing a "likelihood" of $\prod_{i=1}^{m} f_{\theta_{ix}}(\theta_{ix})$ to the distribution of the hyperparameters.

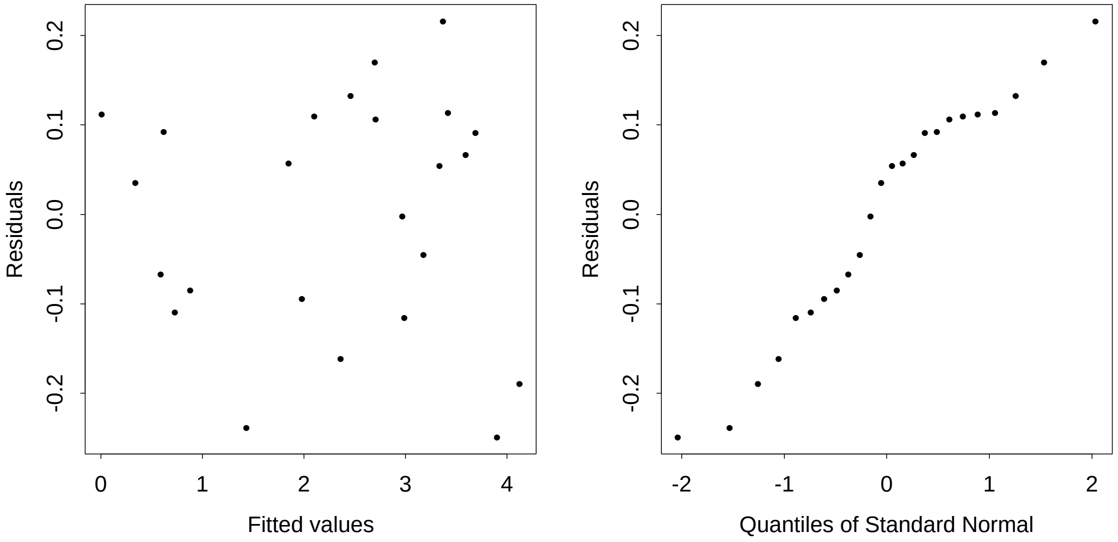

Figure 1: Residual and normal probability plots from a least squares analysis of the church data.

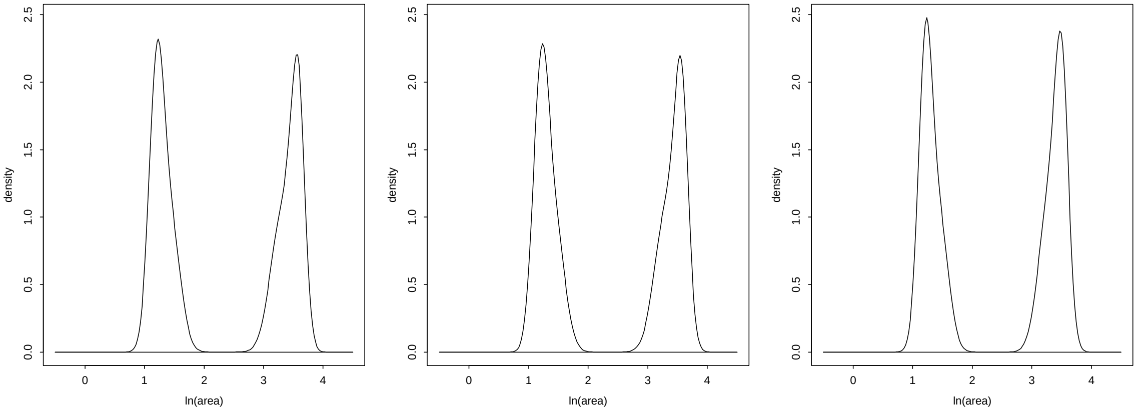

Figure 2: Predictive density plots for two churches, DDP model. From left to right, correlation parameters are 0.8, 0.9 and 0.99.

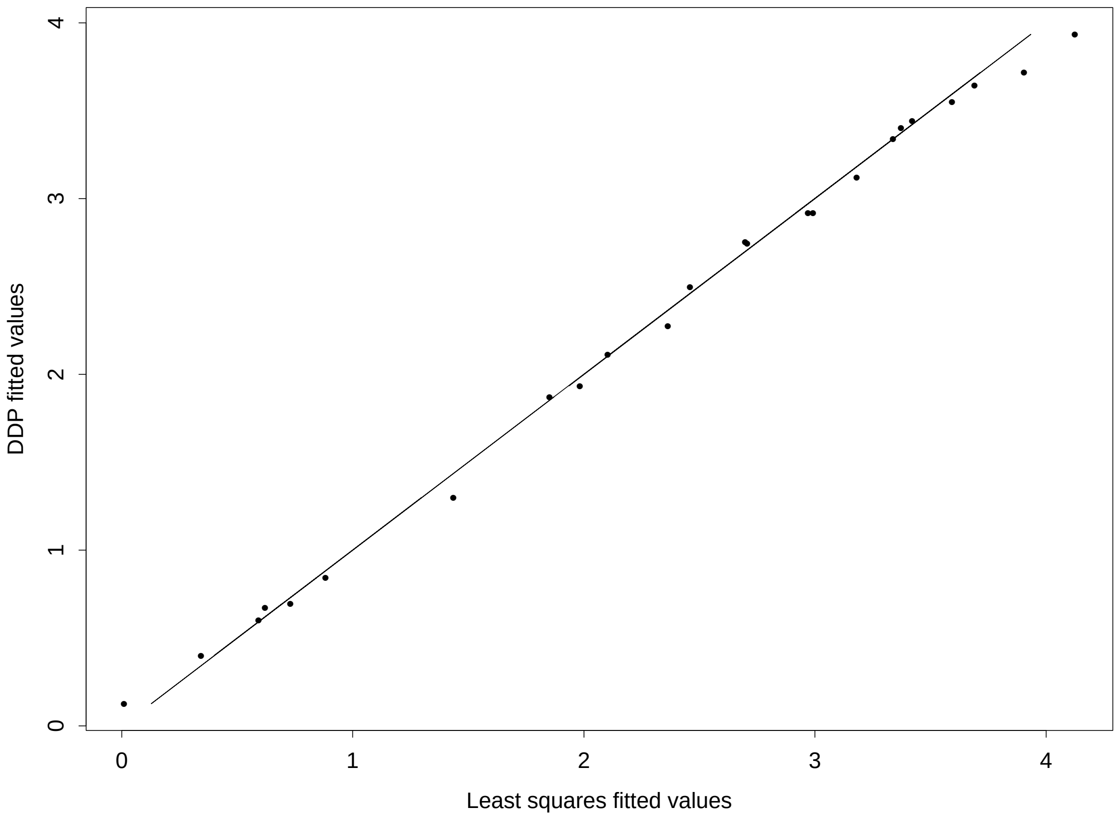

Figure 3: Plot of fits from DDP model against fits from least squares analysis. The line appearing in the figure corresponds to equal fits for the two models.

Table 1: Estimated coefficients under several models. For the DDP models, the values of slope and intercept are estimated by ordinary least squares with a marginal distribution on x equal to a discrete uniform distribution on the 24 observed x values.

coefficient	least squares	DDP 0.5	DDP 0.8	DDP 0.9	DDP 0.99
intercept	2.299	2.283	2.282	2.282	2.281
slope	1.658	1.626	1.620	1.618	1.617

Table 2: Estimated coefficients under several marginal distributions for x. The marginal distribution is uniform on the set of observed x values indicated in the first column. All fits are based on the DDP model with a correlation of 0.99.

data set	intercept	slope
all data	2.279	1.613
1:12	2.268	1.611
2:13	2.288	1.653
3:14	2.300	1.684
4:15	2.291	1.676
5:16	2.281	1.690
6:17	2.277	1.678
7:18	2.278	1.684
8:19	2.293	1.651
9:20	2.284	1.685
10:21	2.294	1.656
11:22	2.287	1.664
12:23	2.366	1.498
13:24	2.387	1.455