Dirichlet Process Mixtures of Generalized Linear Models

Lauren Hannah LHANNAH@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08544, USA

David Blei BLEI@CS.PRINCETON.EDU
Department of Computer Science
Princeton University
Princeton, NJ 08544, USA

Warren Powell POWELL@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08544, USA

Editor: ????

Abstract

We propose Dirichlet Process-Generalized Linear Models (DP-GLM), a new method of nonparametric regression that accommodates continuous and categorical inputs, and any response that can be modeled by a generalized linear model. We prove conditions for the asymptotic unbiasedness of the DP-GLM regression mean function estimate and give a practical example for when those conditions hold. Additionally, we provide Bayesian bounds on the distance of the estimate from the true mean function based on the number of observations and posterior samples. We evaluate DP-GLM on several data sets, comparing it to modern methods of nonparametric regression like CART and Gaussian processes. We show that the DP-GLM is competitive with the other methods, while accommodating various inputs and outputs and being robust when confronted with heteroscedasticity.

1. Introduction

In this paper, we examine a Bayesian nonparametric solution to the general regression problem. The general regression problem models a response variable Y as dependent on a set of d-dimensional covariates X,

$$Y \mid X \sim f(m(X)). \tag{1}$$

In this equation, $m(\cdot)$ is a deterministic mean function, which specifies the conditional mean of the response, and f is a distribution, which characterizes the deviation of the response from the conditional mean. In a regression problem, we estimate the mean function and deviation parameters from a data set of covariate-response pairs $\{(x_i, y_i)\}_{i=1}^N$. Given a new set of covariates x_{new}, we predict the response via its conditional expectation, $\mathbb{E}[Y \mid x_{\text{new}}]$. In Bayesian regression, we compute the posterior expectation of these computations, conditioned on the data.

Regression models have been a central focus of statistics and machine learning. The most common regression model is linear regression. Linear regression posits that the conditional mean is a linear combination of the covariates x and a set of coefficients β, and that the response distribution is Gaussian with fixed variance. Generalized linear models (GLMs) generalize linear regression to a diverse set of response types and distributions. In a GLM, the linear combination of coefficients and covariates is passed through a possibly non-linear *link function*, and the underlying distribution of the response is in any exponential family. GLMs are specified by the functional form of the link function and the distribution of the response, and many prediction models, such as linear regression, logistic regression, multinomial regression, and Poisson regression, can be expressed as a GLM. Algorithms for fitting and predicting with these models are instances of more general-purpose algorithms for GLMs.

GLMs are a flexible family of *parametric* regression models: The parameters, i.e., the coefficients, are situated in a finite dimensional space. A GLM strictly characterizes the relationship between the covariates and the conditional mean of the response, and implicitly assumes that this relationship holds across all possible covariates (McCullagh and Nelder, 1989). In contrast, *nonparametric* regression models find a mean function in an infinite dimensional space, requiring less of a commitment to a particular functional form. Nonparametric regression is more complicated than parametric regression, but can accommodate a wider variety of response function shapes and allow for a response function that adapts to the covariates. That said, nonparametric regression has its pitfalls: A method that is too flexible will over-fit the data; most current methods are tailored for specific response types and covariate types; and many nonparametric regression models are ineffective for high dimensional regression. See Hastie et al. (2009) for an overview on nonparametric regression and Section 1.1 below for a discussion of the current state of the art.

Here, we study Dirichlet process mixtures of generalized linear models (DP-GLMs), a Bayesian nonparametric regression model that combines the advantages of generalized linear models with the flexibility of the nonparametric regression. With a DP-GLM, we model the joint distribution of the covariate and response with a DP mixture: the covariates are drawn from a parametric distribution and the response is from a GLM, conditioned on the covariates. The clustering effect of the DP mixture leads to an "infinite mixture" of GLMs, a model which effectively identifies local regions of covariate space where the covariates exhibit a consistent relationship to the response. In combination, these local GLMs can represent arbitrarily complex global response functions. Note that the DP-GLM is flexible in that the number of segments, i.e., the number of mixture components, is determined by the observed data.

In a DP-GLM, the response function consists of piece-wise GLM models when conditioned on the component (i.e., covariate region).

A DP-GLM represents a collection of GLMs, one for each component, or region of the covariate space. In prediction, we are given a new covariate vector x, which induces a conditional distribution over those components, i.e., over the coefficients that govern the conditional mean of the response. Integrating over this distribution, the predicted response mean is a weighted average of GLM predictions, each prediction weighted by the probability that covariates x are assigned to its coefficients. Thus, the expected response function $\mathbb{E}[Y \mid x]$ is not simply a piece-wise GLM. The sharp jumps between segments are smoothed out by the uncertainty of the cluster assignment for the covariates that lie near the boundaries of the regions. (Furthermore, all of these computations are estimated in a Bayesian way. We estimate and marginalize with respect to a posterior distribution over components and parameters conditioned on a data set of covariate-response pairs.) This model is a generalization of the Dirichlet process-multinomial logistic model of Shahbaba and Neal (2009).

As an example, consider a simple linear regression setting with a single covariate. In this case, a GLM is a linear regression model and the DP mixture of linear regression models finds multiple components, each of which is associated with a covariate mean and a set of regression parameters. Conditioned on the region, the linear regression DP-GLM gives a piece-wise linear response function. But during prediction, the uncertainty surrounding the region leads to a response function that smoothly changes at the boundaries of the region. This is illustrated in Figure 1, where we depict the maximum a posteriori (MAP) estimate of the piece-wise linear response and the smooth MAP response function $\mathbb{E}[Y \mid x_{\text{new}}]$.

The data in Figure 1 are heteroscedastic, meaning that the variance is not constant across the covariate space. Most regression methods, including parametric methods, Gaussian process and SVM regression, assume that the variance is constant, or varies in a known fashion. While these models can be modified to accommodate heteroscedasticity, DP-GLM accommodates it natively by incorporating variance as a variable within each cluster.

In the remainder of this paper, we will show that DP-GLMs are an effective all-purpose regression technique that can balance the advantages of parametric modeling with the flexibility of nonparametric estimation. They readily handle high dimensional data, model heteroscedasticity in the errors, and accommodate diverse covariate and response types. Moreover, we will show that DP-GLMs exhibit desirable theoretical properties, such as the asymptotic unbiasedness of the conditional expectation function $E[Y \mid x]$.

The rest of this paper is organized as follows. In Section 2, we review Dirichlet process mixture models and generalized linear models. In Section 3, we construct the DP-GLM, describe its properties, and derive algorithms for posterior computation; in Section 4 we give general conditions for unbiasedness and prove it in a specific case with conjugate priors. In Section 5 we compare DP-GLM and existing methods on three data sets with both continuous covariates and response variables, one with continuous and categorical covariates and a continuous response, and one with categorical covariates and a count response.

1.1 Related work

Nonparametric regression is a field that has received considerable study. Approaches to nonparametric regression can be broken into two general categories: frequentist and Bayesian. Bayesian methods assume a prior distribution on the set of all mean functions, whereas frequentist methods do not. Frequentist nonparametric methods have comprised the bulk of the literature. Brieman et al. (1984) proposed classification and regression trees (CART), which partition the data and fit constant responses within the partitions. CART can be used with any type of data but produces a non-continuous mean function and is prone to over-fitting, particularly in situations with heteroscedasticity. Friedman (1991) proposed multivariate adaptive regression splines (MARS), which selects basis functions with which it builds a continuous mean function model. Locally polynomial methods (Fan and Gij-

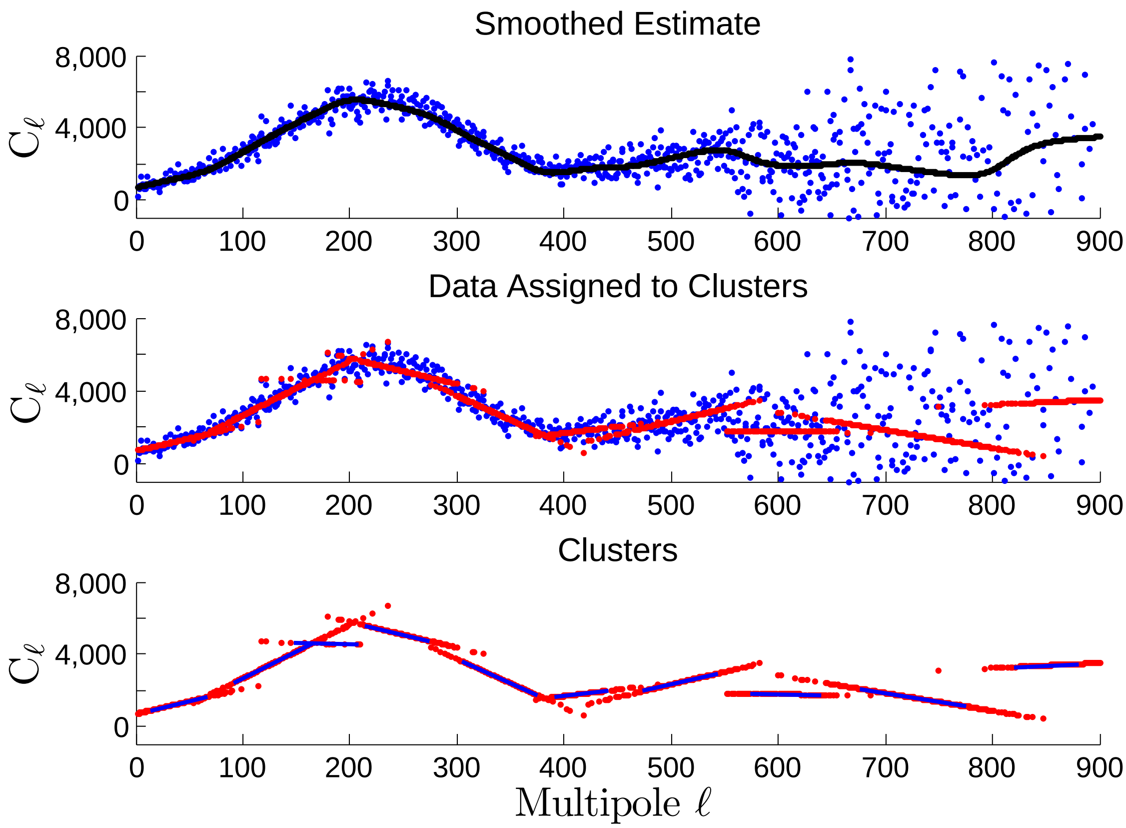

Figure 1: The underlying DP-GLM is the linear model described in equation (3). The top figure shows the smoothed regression estimate. The center figure shows the training data (blue) fitted into clusters, with the prediction given a single sample from the posterior, $\theta^{(i)}$ (red). The bottom figure shows the underlying clusters (blue) centered around the mean $\pm$ one standard deviation; the fitted data (red) are plotted with each cluster. Data are from the cosmic microwave background data set of Bennett et al. (2003) that plots multipole moments against power spectrum C_ℓ.

bels, 1996) calculate local polynomial coefficients via kernel weighting of the observations. Locally polynomial methods have attractive theoretical properties, such as unbiasedness at the covariate boundaries, and often perform well. However, they are sensitive to kernel bandwidth and require dense sampling of the covariate space, which is often unrealistic in higher dimensions. A number of techniques have been proposed for automatic bandwidth selection, in either a local or global form (Fan and Gijbels, 1995; Ruppert et al., 1995), sometimes combined with dimension reduction (Lafferty and Wasserman, 2008). Support vector regression is an active field of research that uses kernels to map a high-dimensional input space into a lower dimensional feature space and fitting support vectors within the feature space (see Smola and Schölkopf (2004) for a review). It can be sensitive to kernel choice.

Gaussian process and Dirichlet process mixtures are the most common prior choices for Bayesian nonparametric regression. Gaussian process priors assume that the observations arise from a Gaussian process model with known covariance function form (see Rasmussen and Williams (2006) for a review). Since the size of the covariance function grows on the order of the number of observations squared, much research has been devoted to sparse approximations (Lawrence et al., 2003; Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006). Although Gaussian processes have been quite successful for regression with continuous covariates, continuous response and i.i.d. errors, other problem settings often require extensive modeling in order to use a GP prior.

We use a Dirichlet process mixture for regression; this in itself is not new. Escobar and West (1995) used a Gaussian kernel with a DP prior on the location parameter for Bayesian density estimation; extension to the regression setting is trivial. West et al. (1994) and Muller et al. (1996) developed a model that uses a jointly Gaussian mixture over both the covariates and response. The model is limited to a continuous covariate/response setting; the use of a fully populated covariance matrix makes the method impractical in high dimensions. More current methods have focused on mixing only over the response with a DP prior and correlating the distributions of the hidden response parameters according to some function of the covariates. This has been achieved by using spatial processes or kernel functions of the covariates to link the local DPs (Griffin and Steel, 2006, 2007; Dunson et al., 2007), or by using a dependent Dirichlet process prior (De Iorio et al., 2004; Gelfand et al., 2005; Duan et al., 2007; Rodriguez et al., 2009). These models can involve a large number of hidden variables or distributions and their complexity can pose difficulties for theoretical claims about the resulting response estimators.

Dirichlet process priors have also been used in conjunction with GLMs. Mukhopadhyay and Gelfand (1997) and Ibrahim and Kleinman (1998) used a DP prior for the random effects portion of the the GLM; the model retains the linear relationship between the covariates and response while allowing over dispersion to be modeled. Likewise, Amewou-Atisso et al. (2003) used a DP prior to model arbitrary symmetric error distributions in a semi-parametric linear regression model. Shahbaba and Neal (2009) proposed a model that mixes over both the covariates and response, which are linked by a multinomial logistic model. The DP-GLM studied here is a generalization.

The asymptotic properties of models using Dirichlet process priors have not been well studied. Most current literature centers around consistency of the posterior density and its rate of convergence to the true density. Weak, L1 (strong) and Hellinger consistency have been shown in instances for DP Gaussian mixture models (Barron et al., 1999; Ghosal et al., 1999; Ghosh and Ramamoorthi, 2003; Walker, 2004; Tokdar, 2006) and semi-parametric linear regression models (Amewou-Atisso et al., 2003; Tokdar, 2006). By decreasing the radius of the balls used to show consistency, consistency results are used to give asymptotic rates of posterior convergence in the models where L1 or Hellinger consistency has been shown (Ghosal et al., 2000; Shen and Wasserman, 2001; Walker et al., 2007; Ghosal and van der Vaart, 2007). More broadly, Shen (2002) gave conditions for asymptotic normality of nonparametric functionals, but required fairly strong conditions. Only recently have the posterior properties of DP regression estimators been studied. Rodriguez et al. (2009) showed point-wise asymptotic unbiasedness for their model, which uses a dependent Dirichlet process prior, assuming continuous covariates under different treatments with a continuous responses and a conjugate base measure (normal-inverse Wishart). We give general conditions for asymptotic unbiasedness for DP-GLM.

2. Mathematical Review

We review Dirichlet process mixture models and generalized linear models.

Dirichlet Process Mixture Models. In a Dirichlet process mixture model, we assume that the true density can be written as a mixture of parametric densities conditioned on a hidden parameter, θ. Due to our model formulation, θ can be split into two parts, θ_x, which is associated only with the covariates X, and θ_y, which is associated only with the response, Y. Set $\theta = (\theta_x, \theta_y)$. Because a Dirichlet process mixture is a Bayesian model, the parameter θ is endowed with a prior, and the marginal probability of an observation is given by a continuous mixture,

$$f_0(x, y) = \int_{\mathcal{T}} f(x, y|\theta) P(d\theta).$$

In this equation, $\mathcal{T}$ is the set of all possible parameters and the prior P is a measure on that space.

We can use a *Dirichlet process* (DP) to model uncertainty about the prior density P (Ferguson, 1973; Antoniak, 1974). If P is drawn from a Dirichlet process with base measure $\mathbb{G}_0$ and scaling parameter α then for any finite partition $A_1, \ldots, A_k$ of $\mathcal{T}$,

$$(P(A_1), \ldots, P(A_k)) \sim \text{Dir}(\alpha\mathbb{G}_0(A_1), \ldots, \alpha\mathbb{G}_0(A_k)).$$

Here $\text{Dir}(a_1, \ldots, a_k)$ denotes the Dirichlet distribution with positive parameters $(a_1, \ldots, a_k)$. The base measure G_0 is a measure on $\mathcal{T}$, and the scaling parameter is a positive scalar value.

The random variable P is a *random measure*. Such random variables are usually hard to manipulate. However, if P is drawn from a Dirichlet process then it can be analytically integrated out of the conditional distribution of θ_n given $\theta_{1:(n-1)}$. Specifically, the random variable Θ_n has a Polya urn distribution (Blackwell and MacQueen, 1973),

$$\Theta_n | \theta_{1:(n-1)} \sim \frac{1}{\alpha + n - 1} \sum_{i=1}^{n-1} \delta_{\theta_i} + \frac{\alpha}{\alpha + n - 1} \mathbb{G}_0. \tag{2}$$

(In this paper, lower case values refer to observed or fixed values, while upper case refer to random variables.)

Equation (2) reveals the *clustering property* of the joint distribution of $\theta_{1:n}$: There is positive probability that each θ_i will take on the value of another θ_j. This equation also makes clear the roles of α and $\mathbb{G}_0$. The unique values of $\theta_{1:n}$ are drawn independently from $\mathbb{G}_0$; the parameter α determines how likely Θ_{n+1} is to be a newly drawn value from $\mathbb{G}_0$ rather than take on one of the values from $\theta_{1:n}$. $\mathbb{G}_0$ controls the distribution of the new component.

In a DP mixture, θ is a latent parameter to an observed data point x,

$$\begin{aligned} P &\sim \mathrm{DP}(\alpha\mathbb{G}_0) \\ \Theta_i &\sim P \\ x_i|\theta_i &\sim f(\cdot\,|\,\theta_i). \end{aligned}$$

Examining the posterior distribution of $\theta_{1:n}$ given $x_{1:n}$ brings out its interpretation as an "infinite clustering" model. Because of the clustering property, observations are grouped by their shared parameters. Unlike finite clustering models, however, the number of groups is random and unknown and, moreover, a new data point can be assigned to a new cluster that was not previously seen in the data. This model is amenable to efficient Gibbs sampling and algorithms (Neal, 2000; Blei and Jordan, 2005), and has emerged a powerful technique for flexible data analysis.

Generalized Linear Models. Generalized linear models (GLMs) build on linear regression to provide a flexible suite of predictive models. While linear regression assumes a Gaussian response, GLMs allow for any exponential family response by mapping a linear function of the covariates to the natural parameter of the response distribution. This embodies familiar models like logistic regression, Poisson regression, and multinomial regression. (Our review here is incomplete; See McCullagh and Nelder (1989) for a full discussion.)

A GLM assumes that a response variable Y is dependent on covariates X, with an exponential family distribution,

$$f(y|\eta) = \exp\left(\frac{y\eta - b(\eta)}{a(\phi)} + c(y, \phi)\right),$$

Here the canonical form of the exponential family is given, where a, b, and c are known functions specific to the exponential family, ϕ is an arbitrary scale (dispersion) parameter, and η is the canonical parameter. The form of the exponential family (e.g. multinomial, Poisson, beta, Gaussian), called a random component, determines the form of the response, Y. A systematic component is included by setting the covariates to be a linear predictor of the canonical parameter,

$$\eta = X\beta.$$

The systematic and random components are combined via a link function g of the mean μ, where $\eta = g(\mu)$. It can be shown that $b'(\eta) = \mu$, thereby connecting the linear predictor to the canonical parameter, $X\beta = g^{-1}(b'(\eta))$.

The canonical form is useful for discussion of GLM properties, but we use the mean form in the rest of this paper. For example, the canonical form of a Gaussian distribution, setting $\eta = \mu$ is

$$f(y|\eta) = \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi})\right],$$

but we use the more familiar mean notation,

$$f(y|\eta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right].$$

3. Dirichlet Process Mixtures of Generalized Linear Models

We now turn to Dirichlet process mixtures of generalized linear models (DP-GLMs), a flexible Bayesian predictive model that places prior mass on a large class of response densities. Given a data set of covariate-response pairs, we describe Gibbs sampling algorithms for approximate posterior inference and prediction. Theoretical properties of the DP-GLM are developed in Section 4.

3.1 DP-GLM Formulation

In a DP-GLM, we assume that the covariates X are modeled by a mixture of exponential-family distributions, the response Y is modeled by a GLM conditioned on the inputs, and that these models are connected by associating a set of GLM coefficients to each exponential family mixture component. Let $\theta = (\theta_x, \theta_y)$ denote the bundle of parameters over X and $Y \,|\, X$, and let $\mathbb{G}_0$ denote a base measure on the space of both. For example, θ_x might be a set of d-dimensional multivariate Gaussian location and scale parameters for a vector of continuous covariates; θ_y might be a $d+2$-vector of reals for their corresponding GLM linear prediction coefficients, along with a nuisance term and a noise term for the GLM. The full model is

$$\begin{aligned} P &\sim DP(\alpha\mathbb{G}_0), \\ (\theta_{x,i}, \theta_{y,i})|P &\sim P, \\ X_i|\theta_{x,i} &\sim f_x(\cdot|\theta_{x,i}), \\ Y_i|x_i, \theta_{y,i} &\sim GLM(\cdot|x_i, \theta_{y,i}). \end{aligned}$$

The density f_x describes the covariate distribution; the GLM for y depends on the form of the response (continuous, count, category, or others) and how the response relates to the covariates (i.e., the link function).

The Dirichlet process prior on the parameters clusters the covariate-response pairs (x, y). When both are observed, i.e., in "training," then the posterior distribution of this model will cluster data points according to near-by covariates that exhibit the same kind of relationship to their response. When the response is not observed, its predictive expectation can be understood by clustering the covariates based on the training data, and then predicting the response according to the GLM associated with the covariates' cluster.

Modeling the covariates with a mixture is not technically needed to specify a mixture of GLMs. Many regression models use a DP mixture without explicitly mixing over the covariates (De Iorio et al., 2004; Gelfand et al., 2005; Griffin and Steel, 2006, 2007; Dunson et al., 2007; Duan et al., 2007; Rodriguez et al., 2009). The resulting models, however, are often inflexible with respect to response/covariate type and have a high degree of complexity. Conversely, a full linear predictor required if we are modeling the covariates with a mixture. The inclusion of linear predictors in the DP-GLM greatly enhances the accuracy while adding a minimal number of hidden variables.

Shahbaba and Neal (2009) use a specific version of DP-GLM for classification with a multinomial logit response function and continuous inputs, which they call the "Dirichlet Process Multinomial Logistic Model" (dpMNL). We have fully generalized this model for use in a variety of settings, including continuous multivariate regression, classification and Poisson regression.

We now give some examples of the DP-GLM that will be used throughout the rest of the paper.

Gaussian Model. In a simple case, the response has a Gaussian distribution, resulting in a linear regression within each cluster. That is,

$$f_y(y|x,\theta_i) = \frac{1}{\sigma_{i,y}\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_{i,y}^2}(y-\beta_{i,0}-\sum_{j=1}^{d}\beta_{i,j}x_j)^2\right].$$

We assume continuous covariates modeled by Gaussian mixtures with mean $\mu_{i,j}$ and variance $\sigma i, j^2$ for the j^{th} dimension of the i^{th} observation. The GLM parameters are the linear predictor $\beta_{i,0},\ldots,\beta_{i,d}$ and the noise coefficient $\sigma_{i,y}^2$. Here, $\theta_{x,i} = (\mu_{i,1:d},\sigma_{i,1:d})$ and $\theta_{y,i} = (\beta_{i,0:d},\sigma_{i,y})$. The full model is,

$$\begin{aligned} P &\sim DP(\alpha\mathbb{G}_0), \qquad (3)\\ \Theta_i = (\mu_{i,1:d},\sigma_{i,1:d},\beta_{i,0:d},\sigma_{i,y})|P &\sim P,\\ X_{i,j}|\mu_{i,j},\sigma_{i,j} &\sim N(\mu_{i,j},\sigma_{i,j}^2), \qquad j=1,\ldots,d,\\ Y_i|x_i,\beta_i,\sigma_{i,y} &\sim N(\beta_{i,0}+\sum_{j=1}^{d}\beta_{i,j}x_{i,j},\sigma_{i,y}^2). \end{aligned}$$

Poisson Model. In a more complicated case, the response is count data. We model this with a Poisson distribution. We assume that the covariates are categorical; these are modeled with multinomial mixtures with k_j levels and parameters $p_{i,j,1},\ldots,p_{i,j,k_j-1}$ in the j^{th} dimension. Because the covariates are categorical, the linear predictor does not use the covariates themselves, but instead the indicators of covariate level. Define $z_{i,j,\ell} = \mathbf{1}_{\{x_{i,j}=\text{ level }\ell\}}$ for $\ell = 1,\ldots,k_j-1$. The hidden parameter is partitioned as follows:

$$\theta_{x,i} = ((p_{i,1,1},\ldots,p_{i,1,k_1-1}),\ldots,(p_{i,d,1},\ldots,p_{i,d,k_d-1}))$$

and

$$\theta_{y,i} = (\beta_{i,0},\beta_{i,1,1:j_1-1},\ldots,\beta_{i,d,k_d-1}),$$

where $\beta_{i,j,1:k_j-1}$ is the set of linear predictors associated with $z_{i,j}$. The full model is,

$$
\begin{aligned}
P &\sim DP(\alpha \mathbb{G}_0), \\
(p_{i,1:d}, \beta_{i,0:d}) &\sim P, \\
X_{i,j} | p_{i,j,1:k_j-1} &\sim Multinomial(1, p_{i,j,1}, \dots, p_{i,j,k_j}), \qquad j = 1, \dots, d \\
Y_i | x_i, \beta_i &\sim \frac{1}{y!} \exp \left[y \left(\beta_{i,0} + \sum_{j=1}^{d} \sum_{\ell=1}^{k_j-1} \beta_{i,j,\ell} z_{i,j,\ell} \right) - \exp \left[\beta_{i,0} + \sum_{j=1}^{d} \sum_{\ell=1}^{k_j-1} \beta_{i,j,\ell} z_{i,j,\ell} \right] \right].
\end{aligned}
$$

The Role of $\mathbb{G}_0$. The choice of $\mathbb{G}_0$ is how prior knowledge about the hidden components is imparted, including the center and spread of their distribution. For example, suppose that the covariates X are as in the Gaussian model. The base measure on μ may be specified $\mu \sim N(m_\mu, s_\mu^2)$. This would produce an accumulation of hidden parameters around m_μ, with large deviations unlikely. However, if $\mu \sim \text{Cauchy}(m_\mu, b)$, the hidden parameters would still be centered around m_μ, but the spread would be much greater.

Aside from prior knowledge considerations, the choice of base measure has computational implications. A conjugate base measure allows the analytical computation of many integrals that involve the hidden parameters integrated with respect to the base measure. This can greatly increase the efficiency of posterior sampling methods.

3.2 DP-GLM Regression

The DP-GLM is used in prediction problems. Given a collection of covariate-response pairs $(x_i, y_i)_{i=1}^n$, our goal is to compute the expected response for a new set of covariates x. Conditional on the latent parameters $\theta_{1:n}$ that generated the observed data, the expectation of the response is

$$
\mathbb{E}[Y | x, \theta_{1:n}] = \frac{\alpha \int_{\mathcal{T}} \mathbb{E}\left[Y | x, \theta\right] f_x(x|\theta) \mathbb{G}_0(d\theta) + \sum_{i=1}^n \mathbb{E}\left[Y | x, \theta_i\right] f_x(x|\theta_i)}{\alpha \int_{\mathcal{T}} f_x(x|\theta) \mathbb{G}_0(d\theta) + \sum_{i=1}^n f_x(x|\theta_i)}. \tag{4}
$$

Since Y is assumed to be a GLM, the quantity $\mathbb{E}\left[Y|x, \theta\right]$ is analytically available as a function of x and θ.

The unobserved random variables $\Theta_{1:n}$ are integrated out using their posterior distribution given the observed data. Following the notation of Ghosh and Ramamoorthi (2003), let Π^P denote the DP prior on the set of hidden parameter measures, P. Let $\mathcal{M}_\mathcal{T}$ be the space of all distributions over the hidden parameters. Since $\int_\mathcal{T} f_y(y|x, \theta) f_x(x|\theta) P(d\theta)$ is a density for (x, y), Π^P induces a prior on $\mathcal{F}$, the set of all densities f on (x, y). Denote this prior by Π^f and define the posterior distribution,

$$
\Pi_n^f(A) = \frac{\int_A \prod_{i=1}^n f(X_i, Y_i) \Pi^f(df)}{\int_\mathcal{F} \prod_{i=1}^n f(X_i, Y_i) \Pi^f(df)},
$$

where $A \subseteq \mathcal{F}$. Define Π_n^P similarly. Then the regression becomes

$$
\begin{aligned}
\mathbb{E}[Y | x, (X_i, Y_i)_{1:n}] = &\frac{1}{b} \sum_{i=1}^n \int_{\mathcal{M}_\mathcal{T}} \int_\mathcal{T} \mathbb{E}[Y|x, \theta_i] f_x(x|\theta_i) P(d\theta_i) \Pi_n^P(dP) \\
&+ \frac{\alpha}{b} \int_\mathcal{T} \mathbb{E}\left[Y|x, \theta\right] f_x(x|\theta) \mathbb{G}_0(d\theta),
\end{aligned} \tag{5}
$$

Algorithm 1: DP-GLM Regression

Data: Observations $(X_i, Y_i)_{1:n}$, functions f_x, f_y, number of posterior samples M, query x
Result: Mean function estimate at x, $\bar{m}(x)$
initialization;
for $m = 1$ *to* M **do**
  Obtain posterior sample $\theta_{1:n}^{(m)} | (X_j, Y_j)_{1:n}$;
  Compute $\mathbb{E}[Y|x, \theta_{1:n}^{(m)}]$;
end
Set $\bar{m}(x) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[Y|x, \theta_{1:n}^{(m)}]$;

where b normalizes the probability of Y being associated with the parameter θ_i,

$$b = \alpha \int_{\mathcal{T}} f_x(x|\theta) \mathbb{G}_0(d\theta) + \sum_{i=1}^{n} \int_{\mathcal{M}_{\mathcal{T}}} \int_{\mathcal{T}} f_x(x|\theta_i) P(d\theta_i) \Pi_n^P(dP).$$

Equation (5) can also be written as

$$\mathbb{E}\left[Y|x, (X_i, Y_i)_{1:n}\right] = \int_{-\infty}^{\infty} y \hat{f}_n(y|x) dy, \tag{6}$$

where $\hat{f}_n(y|x)$ is the posterior predictive distribution after n observations,

$$\hat{f}_n(y|x) = \frac{\int_{\mathcal{F}} f(y|x) \prod_{i=1}^{n} f(Y_i|X_i) \Pi^f(df)}{\int_{\mathcal{F}} \prod_{i=1}^{n} f(Y_i|X_i) \Pi^f(df)}. \tag{7}$$

In this case, let $f(y|x) = f(x,y)/(\int f(x,y)dy)$ and restrict Π^f to place positive measure only on the set $\{f : f(x,y) > 0 \text{ for some } y\}$. Equation (5) is useful for implementation of the DP-GLM. The characterization in equation (6) is used to show theoretical properties, such as asymptotic unbiasedness.

Equations (5) and (6) are difficult to compute because they require integration over a random measure. To avoid this problem, we approximate equation (5) by an average of M Monte Carlo samples of the expectation conditioned on $\theta_{1:n}$. This simpler computation is given in equation (8),

$$\mathbb{E}[Y|x, (X_i, Y_i)_{1:n}] \approx \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\left[Y|x, \theta_{1:n}^{(m)}\right]. \tag{8}$$

The regression procedure is given in Algorithm 1. We describe how to generate posterior samples $(\theta_{1:n}^{(m)})_{m=1}^{M}$ in Section 3.3.

Example: Gaussian Model. If we have the Gaussian model, equation (4) becomes

$$\begin{aligned}\mathbb{E}[Y|x,\theta_{1:n}] &= \frac{\alpha}{b}\int_{\mathcal{T}}(\beta_0+\sum_{j=1}^{d}\beta_j x_j)\prod_{j=1}^{d}\phi_{\sigma_j}(x_j-\mu_j)\mathbb{G}_0(d\theta)\\ &+\frac{1}{b}\sum_{i=1}^{n}(\beta_{i,0}+\sum_{j=1}^{d}\beta_{i,j}x_j)\prod_{j=1}^{d}\phi_{\sigma_{i,j}}(x_j-\mu_{i,j}),\end{aligned}$$

where

$$b=\alpha\int_{\mathcal{T}}\prod_{j=1}^{d}\phi_{\sigma_j}(x_j-\mu_j)\mathbb{G}_0(d\theta)+\sum_{i=1}^{n}\prod_{j=1}^{d}\phi_{\sigma_{i,j}}(x_j-\mu_{i,j}),$$

and

$$\phi_{\sigma}(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}x^2\right]$$

is the Gaussian density at x with variance σ^2.

An example of regression for a single covariate Gaussian model is shown earlier, in Figure 1. The bottom figure shows the mean and standard deviation for each cluster from one posterior sample. The middle figure shows how the training data are placed into clusters and the top figure gives a smoothed estimate of the mean function for testing data. Clusters act locally to give an estimate of the mean function; in areas where there are multiple clusters present, the mean function is generated by an average of the clusters.

Remark on Choice of $\mathbb{G}_0$. Equation (4) demonstrates one of the aspects of prior choice; it includes the term

$$\alpha\int_{\mathcal{T}}f_x(x|\theta)\mathbb{G}_0(d\theta),$$

which can be computed analytically if the base measure of θ_x is conjugate to $f_x(\cdot|\theta)$. If this is not the case, integration must be performed numerically, which does not usually work well in high dimensional spaces. Fortunately, if the covariate space is well-populated, the base measure term is often inconsequential.

3.3 Posterior Sampling Methods

The above algorithm relies on samples of $\theta_{1:n}|(X_i,Y_i)_{1:n}$. We use Markov chain Monte Carlo (MCMC), specifically Gibbs sampling, to obtain $\left\{\theta_{1:n}^{(m)}|(X_i,Y_i)_{1:n}\right\}_{m=1}^{M}$. Gibbs sampling has a long history of being used for DP mixture posterior inference (see Escobar (1994); MacEachern (1994); Escobar and West (1995) and MacEachern and Muller (1998) for foundational work; Neal (2000) provides a modern treatment and state of the art algorithms). While DP mixtures were proposed in the 1970's, they did not become computationally feasible until the mid-1990's, with the advent of these MCMC methods.

Gibbs sampling uses the known conditional distributions of the hidden variables in a sequential manner to generate a Markov chain that has the unknown joint distribution of the hidden parameters as its ergodic distribution. That is, the state variable ϕ_i associated with observation (X_i,Y_i) is drawn given that the rest of the state variable,

$(\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_n)$, is fixed. θ_i is selected according to the Polya urn scheme of equation (2).

The composition of the state variable is determined by the base measure, $\mathbb{G}_0$. If $\mathbb{G}_0$ is conjugate to the model, the hidden parameters can be integrated out of the posterior, leaving only a cluster number as a state variable. Non-conjugate base measures require that the state variable for (X_i, Y_i) is the hidden mixture parameter, θ_i.

To obtain posterior samples, the state variables are updated in a sequential manner over many iterations. Samples are the observed state variables. Two main problems arise because the samples are drawn from a Markov chain. First, successive observations are not independent. Second, observations should be drawn from the posterior distribution only once the Markov chain has reached its ergodic distribution. The first problem is usually approached by sampling in a pulsed manner; e.g., after sampling, the next sample will occur 10 iterations later. The second problem is much more difficult.

Since we are not able to compute the full posterior distribution, we must rely on samples drawn from the MCMC to infer whether convergence has been reached. First, we try to obtain a relatively "good" starting point for the state variable to speed mixing; this is often done through *a priori* information or a mode estimate. Then to minimize the impact of starting position, we discard the early iterations of the sampler in a burn-in period. Further precautions include checking whether the log likelihood of the posterior has "leveled off" or computing parameter variation estimates based on between- and within- sequence variances (Brooks and Gelman, 1998). For a full discussion of convergence assessment, see Gelman et al. (2004).

A simple Gibbs sampler for a non-conjugate base measure is given in Algorithm 2.

Other Posterior Sampling Methods. MCMC is not the only way to sample from the posterior. Blei and Jordan (2005) apply variational inference methods to Dirichlet process mixture models. Variational inference methods use optimization to create a parametric distribution close to the posterior from which samples can be drawn. Variational inference avoids the issue of determining when the stationary distribution has been reached, but lacks some of the convergence properties of MCMC. Fearnhead (2004) uses a particle filter to generate posterior samples for a Dirichlet process mixture. Particle filters also avoid having to reach a stationary distribution, but the large state vectors created by non-conjugate base measures and large data sets can lead to collapse.

Hyperparameter Sampling. Hyperparameters, such as α and the parametric components of $\mathbb{G}_0$, can be sampled via Gibbs sampling. For our numerical experiments below (see Section 5), we placed a gamma prior on α and resampled every five iterations. Sampling for hyperparameters associated with $\mathbb{G}_0$ varied with the choice of $\mathbb{G}_0$.

3.4 The Conditional Distribution of DP-GLM

We have formulated the DP-GLM and given an algorithm for posterior inference and prediction. Finally, we explore precisely the type of model that the DP-GLM produces. Consider the distribution $f(y|x)$ for a fixed x under DP-GLM. This is also a mixture model with

Algorithm 2: MCMC Gibbs Sampling for the DP-GLM Posterior with Non-Conjugate Base Measures

Data: Observations $(X_i, Y_i)_{1:n}$, initial MCMC state $\hat{\theta}^{(1)}_{1:n}$, number of posterior samples M, lag time L, MC convergence criteria
Result: Posterior sample $(\theta^{(m)}_{1:n})_{m=1}^{M}$
initialization;
Set $\ell = 0$, $i = 1$, $m = 1$;
while $m \leq M$ **do**
  for $j = 1$ *to* n **do**
    Draw $\hat{\theta}^{(m+1)}_i | \hat{\theta}^{(m)}_1, \dots, \hat{\theta}^{(m)}_{j-1}, \hat{\theta}^{(m)}_{j+1}, \dots, \hat{\theta}^{(m)}_n, (X_i, Y_i)_{1:n}$;
    Set $m = m + 1$;
  end
  if *Convergence criteria satisfied* **then**
    if remainder$(\ell, L) = 0$ **then**
      Set $\theta^{(m)}_{1:n} = \hat{\theta}^{(m)}_{1:n}$;
      Set $i = i + 1$;
    end
    Set $\ell = \ell + 1$;
  end
end

hidden parameters,

$$f(y|x) = \sum_{i=1}^{\infty} f_y(y|x, \theta_y) p(\theta_y|x).$$

Usually the full set of hidden parameters can be collapsed into a smaller set only associated with y, θ_y. In general, for a fixed x, θ_y alone does not have a Dirichlet process prior.

To see why we lose the Dirichlet process prior for the conditional distribution, we work with the Poisson random measure construction of a Dirichlet process (Pitman, 1996). First, we construct a Poisson random measure representation for the hidden parameters of the full model. We then fix x and use a transition kernel generated by the DP-GLM model to move the process from (θ, p) to (θ_y, q). This induces a new Poisson random measure. The parameter θ_y is usually just a subset of θ, but q is generated by weighting p by some function of θ and x. It is in this change of weights that the Dirichlet process prior is lost. We compute a necessary and sufficient condition under which the resulting Poisson random measure also characterizes a Dirichlet process.

Theorem 1 (Pitman (1996)) *Let $\Gamma_{(1)} > \Gamma_{(2)} > \dots$ be the points of a Poisson random measure on $(0, \infty)$ with mean measure $\alpha \frac{e^{-p}}{p} dp$. Put*

$$G_i = \Gamma_{(i)} / \Lambda,$$

where $\Lambda = \sum_i \Gamma_{(i)}$ and define

$$F = \sum_{i=1}^{\infty} G_i \delta_{\Theta_i}, \tag{9}$$

where Θ_i are i.i.d. $\mu(d\theta)g_0(d\theta) = \mathbb{G}_0(d\theta)$, independent also of the $\Gamma_{(i)}$. Then, F is a Dirichlet process with base measure $(\alpha\mathbb{G}_0)$.

The pairs $(\Theta_i, \Gamma_{(i)})$ form a Poisson random measure on $\mathcal{T} \times (0, \infty)$ with mean measure

$$\nu(d\theta, dp) = g_0(\theta)\alpha\frac{e^{-p}}{p}\mu(d\theta)dp. \tag{10}$$

Now consider the induced mixture model on the y component. If the joint model has a Dirichlet process prior, then

$$\begin{aligned}\sum_{i=1}^{\infty} f_y(y|\theta_i, x)p(\theta_i|x) &= \sum_{i=1}^{\infty} f_y(y|\theta_i, x)\frac{p(x|\theta_i)p(\theta_i)}{p(x)} \\ &= \sum_{i=1}^{\infty} f_y(y|\theta_{(i)}, x)\frac{p(x|\theta_{(i)})\Gamma_{(i)}}{\Lambda\int p(x|\theta)F(d\theta)} \\ &= \sum_{i=1}^{\infty} f_y(y|\theta_{(i)}, x)\frac{p(x|\theta_{(i)})\Gamma_{(i)}}{\sum_{i=1}^{\infty} p(x|\theta_{(i)})\Gamma_{(i)}}.\end{aligned} \tag{11}$$

The conditional measure on the hidden parameters in equation (11) has the same form as the measure in equation (9) where $\Gamma_{(i)}$ is replaced by $p(x|\theta_{(i)})\Gamma_{(i)}$ and Λ by $\sum_{i=1}^{\infty} p(x|\theta_{(i)})\Gamma_{(i)}$.

When we shift from the overall model to the conditional model, two mappings occur. The first maps θ to θ_y, where θ_y is usually just a subset of the hidden parameter vector θ. The second maps p to q, and usually depends on both p and θ. We will represent these by a transition kernel, $K_x(\theta, p; d\theta_y, dq)$. For instance, if we have the Gaussian model, then

$$K_x(\theta, p; d\theta_y, dq) = \delta_\beta(\beta)\delta_{\sigma_y}(\sigma_y)\delta_{p\prod_{i=1}^{d}\phi_{\sigma_i}(x_i-\mu_i)}(q)d\beta d\sigma_y dq, \tag{12}$$

where $\phi_\sigma(x)$ is the Gaussian pdf with standard deviation σ evaluated at x and $\delta_y(x)$ is the Dirac measure with mass at y. The transition is deterministic.

Using the transition kernel, we can calculate a mean measure $\tilde{\nu}(d\theta_y, dq)$ for the induced Poisson random measure,

$$\tilde{\nu}(d\theta_y, dq) = \int_{\mathcal{T}}\int_{(0,\infty)} \mu(d\theta)\, dp g_0(\theta)\, \alpha\frac{e^{-p}}{p}\, K_x(\theta, p; d\theta_y, dq).$$

To remain as a Dirichlet process prior, $\tilde{\nu}$ must have the same form as equation (10), where $g_0(\theta)$ is allowed to be flexible, but the measure on the weights must have the form $\tilde{\alpha}\, q^{-1}e^{-q}dq$. In the following Theorem, we give a necessary and sufficient condition for this.

Theorem 2 *Fix x. Let (Θ, Γ) have a Dirichlet process prior with base measure $\alpha\mathbb{G}_0$ and let (Θ_y, Υ) be the parameters and weights associated with the conditional mixture. Then, the prior on (Θ_y, Υ) is a Dirichlet process if and only if for every $\lambda > 0$,*

$$\int_{\mathcal{T}} \mu(d\theta)g_0(\theta)\int_{(0,\infty)} dp\frac{e^{-p}}{p}\int_{(0,\infty)} K_x(\theta, p; d\theta_y, dq)\left(1 - e^{-\lambda q}\right) = \tilde{\nu}_\theta(d\theta_y)\log\left(1 + \lambda\right).$$

Proof Let $h(q) = 1 - e^{-\lambda q}$, then set

$$\int_{\mathcal{T}} \int_{(0,\infty)} \mu(d\theta)\, dp g_0(\theta)\, \alpha \frac{e^{-p}}{p}\, K_x(\theta, p; d\theta_y, dq) h(q) = \tilde{\nu}(d\theta_y, dq) h(q),$$

and integrate over q. Like a Laplace transform, the expectation of h uniquely characterizes the distribution. ■

Example: Gaussian Model. We use Theorem 2 to check the mean measure of the conditional distribution. For the transition kernel given in equation (12),

$$\begin{aligned}
&\int_{\mathcal{T}} \mu(d\theta) g_0(\theta) \int_{(0,\infty)} dp \frac{e^{-p}}{p} \int_{(0,\infty)} K_x(\theta, p; d\theta_y, dq) \left(1 - e^{-\lambda q}\right) \\
&\quad = \int_{\mathcal{T}} \mu(d\theta) g_0(\theta) \int_{(0,\infty)} dq \left(\frac{\exp\left(-q/f(\theta)\right) - \exp\left(-q/f(\theta)(1 + \lambda f(\theta))\right)}{q/f(\theta)} \right), \\
&\quad = d\beta d\sigma_y g_0(\beta, \sigma_y) \int_{\mathcal{T}} f(\theta) \log(1 + \lambda f(\theta)) \mu(d\theta) g_0(\mu_{1:d}, \sigma_{1:d}),
\end{aligned}$$

where $f(\theta) = \prod_{i=1}^{d} \phi_{\sigma_i}(x_i - \mu_i)$. Notice that the $f(\theta)$ term has slipped into the log function, making q dependent on θ and, therefore, the prior on the conditional parameters not a Dirichlet process.

Practically, this construction tells us that the most efficient means of sampling the conditional posterior distribution is via the joint posterior distribution. Sampling the joint distribution allows us to use the structure of the Dirichlet process (either the stick-breaking construction of Sethuraman (1994) or the Polya urn posterior). The conditional posterior may have fewer hidden variables, but it lacks the DP mathematical conveniences.

4. Asymptotic Unbiasedness of the DP-GLM Regression Model

We cannot assume that the mean function estimate computed with the DP-GLM will be close to the true mean function, even in the limit. Diaconis and Freedman (1986) give an example of a location model with a Dirichlet process prior where the estimated location can be bounded away from the true location, even when the number of observations approaches infinity. We want to assure that DP-GLM does not end up in a similar position.

Traditionally, an estimator is called unbiased if the expectation of the estimator over the observations is the true value of the quantity being estimated. In the case of DP-GLM, that would mean for every $x \in \mathcal{A}$ and every $n > 0$,

$$\mathbb{E}_{f_0}\left[\mathbb{E}_{\Pi}[Y|x, (X_i, Y_i)_{i=1}^n]\right] = \mathbb{E}_{f_0}[Y|x],$$

where $\mathcal{A}$ is some fixed domain, $\mathbb{E}_{\Pi}$ is the expectation with respect to the prior Π and $\mathbb{E}_{f_0}$ is the expectation with respect to the true distribution.

Since we use Bayesian priors in DP-GLM, we will have bias in almost all cases (Gelman et al., 2004). The best we can hope for is asymptotic unbiasedness, where as the number of observations grows to infinity, the mean function estimate converges to the true mean function. That is, for every $x \in \mathcal{A}$,

$$\mathbb{E}_{\Pi}[Y|x, (X_i, Y_i)_{i=1}^n] \rightarrow \mathbb{E}[Y|x] \qquad \text{as } n \rightarrow \infty.$$

Diaconis and Freedman (1986) give an example for a location problem with a DP prior where the posterior estimate was not asymptotically unbiased. Extending that example, it follows that estimators with DP priors do not automatically receive asymptotic unbiasedness.

The question is under which circumstances does the DP-GLM escape the Diaconis and Freedman (1986) trap and have asymptotic unbiasedness. To show that the DP-GLM is asymptotically unbiased, we use equation (6) and show that

$$\left|\mathbb{E}_{f_0}[Y|x] - \mathbb{E}_{\hat{f}_n}[Y|x]\right| \rightarrow 0$$

in an appropriate manner as $n \rightarrow \infty$, where $\mathbb{E}_f[\cdot]$ means "expectation under f" and $\hat{f}_n$ is the posterior predictive density, given in equation (7).

We show asymptotic unbiasedness through consistency. It is the idea that, given an appropriate prior, as the number of observations goes to infinity the posterior distribution accumulates in neighborhoods arbitrarily "close" to the true distribution. Consistency depends on the model, the prior and the true distribution. Consistency gives the condition that, if the posterior distribution accumulates in weak neighborhoods (sets of densities under which the integral of all bounded, continuous functions is close to the integral under the true density), the posterior predictive distribution converges weakly to the true distribution (meaning that the posterior predictive distribution is in all weak neighborhoods of the true distribution). Weak convergence of measure is not sufficient to show convergence in expectation; uniform integrability of the random variables generated by the posterior predictive density is also required. When weak consistency is combined with uniform integrability, asymptotic unbiasedness is achieved.

In this section, we first review consistency and related topics in subsection 4.1. Using this basis, in subsection 4.2 we state and prove Theorem 9, which gives general conditions under which the mean function estimate given by DP-GLM is asymptotically unbiased. In subsection 4.3, we use Theorem 9 to show DP-GLM with conjugate priors produces an asymptotically unbiased estimate for the Gaussian model and related models.

4.1 Review of Consistency and Related Topics

Let F_0 be the true distribution. For convenience in the integral notation, we assume that all densities are absolutely continuous with respect to the Lebesgue measure, but that need not be the case (for example, a density may be absolutely continuous with respect to a counting measure). Let $\mathbb{E}_f[Y]$ denote $\int y f(y) dy$, the expectation of Y under the density f, let Ω be the outcome space for $(X_i, Y_i)_{i=1}^{\infty}$, let F_0^{∞} be the infinite product measure on that space, and let $\hat{f}_n(x)$ be the posterior predictive density given observations $(X_i, Y_i)_{i=1}^n$.

Weak consistency, and uniform integrability form the basis of convergence for this problem. Weak consistency, uniform integrability and related topics are outlined below.

First, we define weak consistency and show that weak consistency of the prior implies weak convergence of the posterior predictive distribution to the true distribution.

Definition 3 *A weak neighborhood U of f_0 is a subset of*

$$V = \left\{ f : \left| \int g(x)f(x)dx - \int g(x)f_0(x)dx \right| < \epsilon \text{ for all bounded, continuous } g \right\}.$$

A weak neighborhood of f_0 is a set of densities defined by the fact that integrals of all "well behaving functions," i.e. bounded and continuous, taken with respect to the densities in the set are within ϵ of those taken with respect to f_0. This definition of "closeness" can include some densities that vary greatly from f_0 in sufficiently small areas.

Definition 4 $\{\Pi(\cdot|(X_i, Y_i)_{1:n})\}$ *is said to be weakly consistent at f_0 if there is a $\Omega_0 \subset \Omega$ such that $P^{\infty}_{F_0}(\Omega_0) = 1$ and $\forall \omega \in \Omega_0$, for every weak neighborhood U,*

$$\Pi(U|(X_i, Y_i)_{1:n}(\omega)) \to 1$$

for all weak neighborhoods of f_0.

Weak consistency means that the posterior distribution gathers in the weak neighborhoods of f_0.

Let $f_n \Rightarrow f$ denote that f_n converges to f weakly. That is,

$$\int g(x)f_n(x)dx \to \int g(x)f(x)dx$$

for all bounded, uniformly continuous functions g.

The following proposition shows that weak consistency of the prior induces weak convergence of the posterior predictive distribution, $\hat{f}_n$, as described in equation (7), to the true distribution, f_0.

Proposition 5 *If $\{\Pi^f(\cdot|(X_i, Y_i)_{1:n})\}$ is weakly consistent at f_0, then $\hat{f}_n \Rightarrow f_0$, almost surely $\mathbb{P}^{\infty}_{F_0}$.*

Proof Let U be a weak neighborhood of f_0. Let g be bounded and continuous. Then,

$$\begin{aligned}
\left| \int g(y)\hat{f}_n(y|x)dy - \int g(y)f_0(y|x)dy \right| &= \left| \int g(y) \int_{\mathcal{F}} f(y|x)\Pi^f_n(df) - \int g(y)f_0(y|x)dy \right|, \\
&\leq \left| \int_U \int g(y)f(y|x)dy - \int g(y)f_0(y|x)dy \right| \\
&\quad + \left| \int_{U^C} \int g(y)f(y|x)dy \right| \\
&\leq \epsilon + o(1).
\end{aligned}$$

■

Since weak consistency implies weak convergence of the posterior predictive distribution to the true distribution, we need to show weak consistency. To do this, we first use the Kullback-Leibler (KL) divergence between two densities f and g, $K(f, g)$, where

$$K(f, g) = \int_{\mathbb{R}^d} f(x) \log \left(\frac{f(x)}{g(x)} \right) dx.$$

Next, consider neighborhoods of densities where the KL divergence from f_0 is less than ϵ and define $K_\epsilon = \{g : K(f_0, g) < \epsilon\}$. If Π^f puts positive measure on these neighborhoods for every $\epsilon > 0$, then we say that Π^f is in the K-L support of f_0.

Definition 6 *For a given prior Π^f, f_0 is in the K-L support of Π^f if $\forall \epsilon > 0$, $\Pi^f\left(K_\epsilon(f_0)\right) > 0$; this is denoted by $f_0 \in KL(\Pi^f)$.*

A theorem by Schwartz (1965), as modified by Ghosh and Ramamoorthi (2003), provides a simple way to show weak posterior consistency using the idea of K-L support.

Theorem 7 (Schwartz (1965)) *Let Π^f be a prior on $\mathcal{F}$. If f_0 is in the K-L support of Π^f, then the posterior is weakly consistent at f_0.*

Weak consistency alone does not prove convergence of $\mathbb{E}_{\hat{f}_n}[Y|x]$ to $\mathbb{E}_{f_0}[Y|x]$. Let $(\mathfrak{F}_n)_{n=1}^\infty$ be the filtration generated by the observations where $\mathfrak{F}_n = \sigma((X_i, Y_i)_{i=1}^n)$ and σ is the minimal σ-algebra. The sequence of random variables $(Y|x, \mathfrak{F}_n)_{n=1}^\infty$ needs to be uniformly integrable to ensure convergence of $\mathbb{E}_{\hat{f}_n}[Y|x]$ to $\mathbb{E}_{f_0}[Y|x]$ for a specific x. This gives pointwise convergence of the conditional expectation.

Definition 8 *A set of random variables $X_1, X_2, \ldots$ is uniformly integrable if*

$$\rho(\alpha) = \sup_n \mathbb{E}\left[|X_n|\, \mathbf{1}_{|X_n|>\alpha}\right] \to 0,$$

as $\alpha \to \infty$.

This definition is often hard to check, so we rely on a well known sufficient condition for uniform integrability (Billingsley, 2008). If there exists an $\epsilon > 0$ such that,

$$\sup_n \mathbb{E}\left[|X_n|^{1+\epsilon}\right] < \infty,$$

then the sequence $(X_n)_{n\geq 1}$ is uniformly integrable.

We now have the necessary concepts to discuss the asymptotic unbiasedness of DP-GLM regression.

4.2 Asymptotic Unbiasedness of the DP-GLM Regression Model

In this subsection we give a general result for the asymptotic unbiasedness of the DP-GLM regression model. We do this by showing uniform integrability of the conditional expectations in Lemma 11. We show uniform integrability by demonstrating that the posterior predictive densities, $\hat{f}_n(y|x)$ are martingales with respect to $\mathfrak{F}_n$ in Lemma 10. The martingale property used to create the positive martingale $\mathbb{E}_{\hat{f}_n}\left[|Y|^{1+\epsilon}|x\right]$, which is used to show uniform integrability of $(Y|x, \mathfrak{F}_n)_{n=1}^\infty$. We are now ready to state the main theorem.

Theorem 9 *Fix x. Let Π^f be a prior on $\mathcal{F}$. If*

(i) Π^f *is in the K-L support of $f_0(y|x)$,*

(ii) $\int |y| f_0(y|x) dy < \infty$, *and*

(iii) there exists an $\epsilon > 0$ *such that*

$$\int \int |y|^{1+\epsilon} f_y(y|x,\theta)\mathbb{G}_0(d\theta) < \infty,$$

then $\mathbb{E}_{\hat{f}_n}[Y|x] \to \mathbb{E}_{f_0}[Y|x]$ *almost surely* $\mathbb{P}_{F_0^\infty}$*.*

The conditions of Theorem 9 must be checked for the problem (f_0) and prior (Π^f) pair, which we do for the Gaussian model in subsection 4.3. Condition (i) assures weak consistency of the posterior. Condition (iii) ensures that $\Pi^P\left\{P : \int\int |y|^{1+\epsilon} f_y(y|x,\theta)P(d\theta)\right\} = 1$, meaning that the random variable with the density f is almost surely uniformly integrable. Together, conditions (ii) and (iii) assure that there will always be a finite conditional expectation and that it is uniformly integrable.

To prove Theorem 9, we first show the integrability of the random variables generated by the posterior predictive distribution via martingale methods. While traditional mixtures of Gaussian densities are always uniformly integrable, this is not always the case when the location parameters are chosen according to an underlying distribution. That is, a prior with sufficiently heavy tails may not produce a uniformly integrable mixture. We begin by showing that the distribution $\hat{f}_n(y|x)$ is a martingale.

The following Lemma is found in Nicoleris and Walker (2006):

Lemma 10 *The predictive density* $\hat{f}_n(y|x)$ *at the point* (y, x) *is a martingale with respect to the filtration* $(\mathfrak{F}_n)_{n\geq 0}$*.*

Lemma 10 is useful for constructing other martingales, such as $\mathbb{E}_{\hat{f}_n}[|Y|^{1+\epsilon}|x]$. We use these martingales to show uniform integrability of the random variables generated by the posterior predictive distribution.

Lemma 11 *Suppose that conditions* (ii) *and* (iii) *of Theorem 9 are satisfied. Then,*

$$\sup_n \mathbb{E}_{\hat{f}_n}[|Y|^{1+\epsilon}|x] < \infty.$$

Proof $\mathbb{E}_{\hat{f}_n}[|Y|^{1+\epsilon}|x]$ is a positive martingale with respect to $(\mathfrak{F}_n)_{n\geq 0}$ and conditions (ii) – (iii) assure that this quantity will always exist. Since it is a positive martingale, it will almost surely converge to a finite limit. Therefore,

$$\sup_n \mathbb{E}_{\hat{f}_n}[|Y|^{1+\epsilon}|x] < \infty$$

almost surely $\mathbb{P}_{F_0^\infty}$. ■

Using Lemma 11, the proof of Theorem 9 follows.

Proof [Proof of Theorem 9] By Theorem 7, condition (i) implies weak consistency of the prior. Proposition 5 states that weak consistency implies weak convergence of the posterior predictive density. Conditions (ii) and (iii) imply uniform integrability of that density, which taken along with weak convergence give convergence of the expectation.

■

Showing that the conditions of Theorem 9 are satisfied requires some effort. In the next subsection we demonstrate that they are satisfied for the Gaussian model with a conjugate base measure.

4.3 Asymptotic Unbiasedness of DP-GLM for a Gaussian Model with a Conjugate Base Measure

In Theorem 9, the most difficult condition to satisfy is condition (i), which requires that the prior of the conditional distribution be weakly consistent. This subsection shows that the Gaussian model with a conjugate base measure satisfies condition (i). Additionally, they also satisfy condition (iii), resulting in asymptotic unbiasedness so long as the true mean function actually exists.

To reach this conclusion, we first explore how the prior on the joint distribution is related to the prior on the conditional distribution. It turns out that KL divergence for the conditional prior is very similar to the KL divergence of the joint prior. Next, we modify the work of Tokdar (2006), which gives weak consistency conditions for a single-dimensional location-scale mixture of Gaussians where the hidden parameters have a Dirichlet process prior. We do this by mapping the DP-GLM Gaussian model to a model that only has a location-scale response and show that the prior on these parameters is still a Dirichlet process in Lemma 12. Then we extend the work of Tokdar (2006) to give weak consistency criteria for the prior on a multidimensional location-scale mixture in Theorem 14. Finally, we show that a conjugate base measure satisfies Theorem 14.

Weak Consistency Theorems. Fix x. Using Theorem 7, weak consistency for a given f_0 and prior Π^f can be proven by showing that for an arbitrary $\epsilon > 0$, the set $\{f : f \text{ satisfies Equation 13}\}$ has strictly positive measure under Π^f, where

$$\int_{-\infty}^{\infty} f_0(y|x) \log \frac{f_0(y|x)}{f(y|x)} dy < \epsilon. \tag{13}$$

Since Π^P is defined on the hidden parameters of the joint distribution of $f(x, y)$ and not the conditional $f(y|x)$, we need to express equation (13) using the joint distribution. Let $f(x) = \int f(x, y)dy$. Then,

$$\begin{aligned}\int_{-\infty}^{\infty} f_0(y|x) \log \frac{f_0(y|x)}{f(y|x)} dy &= \int_{-\infty}^{\infty} \frac{f_0(x, y)}{f_0(x)} \left(\log \frac{f_0(x, y)}{f(x, y)} + \log \frac{f(x)}{f_0(x)} \right) dy,\\ &= \log \frac{f(x)}{f_0(x)} + \frac{1}{f_0(x)} \int_{-\infty}^{\infty} f_0(x, y) \log \frac{f_0(x, y)}{f(x, y)} dy.\end{aligned}$$

To give conditions for weak consistency, we build a series of Lemmas that are similar to those of Tokdar (2006). That work gives conditions for the weak consistency of a one-dimensional continuous density with a Dirichlet process prior on a location-scale mixture. We extend these results to our setting. First, we show that for a fixed x the DP prior on the location, scale and slopes of DP-GLM induces a DP prior on a location-scale mixture. Then we approach the problem in the same manner as Tokdar (2006) and modify the Lemmas used for a multidimensional setting. Lemma 13 gives a tail condition that, if satisfied, guarantees weak consistency. Theorem 14 constructs a prior that satisfies Lemma 13 by placing restrictions on the DP base measure, $\mathbb{G}_0$.

DP-GLM does not have the traditional location-scale format used in Muller et al. (1996), Ghosal et al. (1999) and Tokdar (2006), but instead has conditional response density $\phi_{\sigma_y}(y - \beta_0 - \sum_{i=1}^d \beta_i x_i)$, where $\phi_\sigma(x)$ is the Gaussian probability density function with standard deviation σ evaluated at x. For fixed x, however, we can introduce the location parameter $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$. Lemma 12 shows the parameters $(\eta, \sigma_y, \mu_{1:d}, \sigma_{1:d})$ also have a Dirichlet process prior. Similar mappings also have Dirichlet process priors because they change only the base measure of the hidden parameters, $\mathbb{G}_0$, not the weights associated with each component. Let Π^P be the original prior and $\tilde{\Pi}^P$ be the induced prior on $P(\eta, \sigma_y, \mu_{1:d}, \sigma_{1:d})$.

Lemma 12 *For a fixed x, let $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$. If $\Pi^P \sim DP(\alpha, \mathbb{G}_0)$, then $\tilde{\Pi}^P \sim DP(\alpha, \tilde{\mathbb{G}}_0)$, where*

$$\tilde{\mathbb{G}}_0(A_1 \times A_2 \times A_3) = \int_{\mathbb{R}^{d+1}} \mathbf{1}_{A_1}(\eta)\, \mathbb{G}_0(d\beta_0, \dots, d\beta_d, A_2, A_3).$$

The proof is in the Appendix. This lemma allows us to follow the methods of Tokdar (2006) closely by changing our parameters into the same location-scale mixtures. We need only extend the theorems of Tokdar (2006) to multiple dimensional settings.

The following lemma uses a tail condition on the KL divergence that, if satisfied, shows that the true density is in the KL support of the prior. In order to state the density f in terms of P, we use $f = \phi^{d+1} * P$ to denote the convolution of the $d+1$ dimensional Gaussian pdf with the measure $P(\eta, \sigma_y, \mu_{1:d}, \sigma_{1:d})$.

Lemma 13 *Suppose that $f_0 \in \mathcal{F}$ and $\tilde{\Pi}^P$ satisfies the following properties: for any $0 < \tau < 1$ and any $\epsilon > 0$, there exists a set $\mathcal{A}$ and a $y_0 > 0$ such that,*

i) $f_0(x) > 0$,

ii) there exists a closed hypercube $\mathcal{D} \subset \mathbb{R}^d$ such that $x \in \text{int}(\mathcal{D})$ and

$$\int_{\mathcal{D}\times\mathbb{R}} f_0(x,y) \log f_0(x,y) dx dy < \infty,$$

iii) $\tilde{\Pi}^P(\mathcal{A}) > 1 - \tau$, and

iv) for any $f = \phi^{d+1} * P$ with $P \in \mathcal{A}$,

$$\int_{|y|>y_0} f_0(x,y) \log \frac{f_0(x,y)}{f(x,y)} dy < \epsilon.$$

Then, $f_0(y|x) \in KL(\Pi^f)$.

The proof is in the Appendix. Now, we construct conditions for which Lemma 13 holds. This is done mainly by constructing tail conditions on the base measure for the response location-scale parameters, $\tilde{\mathbb{G}}_{0,y}$. For convenience, we assume that the base measure on the covariate parameters, $\tilde{\mathbb{G}}_{0,x}$ is independent of $\tilde{\mathbb{G}}_{0,y}$.

Theorem 14 *Fix x. Let f_0 be a density on $\mathbb{R}^{d+1}$ satisfying,*

i) $f_0(x) > 0$,

ii) there exists a closed hypercube $\mathcal{D} \subset \mathbb{R}^d$ such that $x \in \mathrm{int}(\mathcal{D})$ and

$$\int_{\mathcal{D}\times\mathbb{R}} f_0(x,y)\log f_0(x,y)dxdy < \infty,$$

iii) there exists $\gamma \in (0,1)$ such that $\int |y|^\gamma f_0(x,y)dy < \infty$.

Further assume that $\tilde{\mathbb{G}}_{0,y}$ is independent of $\tilde{\mathbb{G}}_{0,x}$. Assume that $\tilde{\mathbb{G}}_{0,x}$ puts positive density on all possible hidden parameter values. Assume that there exist $\sigma_0 > 0$, $0 < \nu < \gamma$, $\psi > \nu$ and $b_1, b_2 > 0$ such that for large $y > 0$,

iv) $\max\left\{\tilde{\mathbb{G}}_{0,y}\left([y - \sigma_0 y^{\gamma/2}, \infty) \times [\sigma_0, \infty)\right), \tilde{\mathbb{G}}_{0,y}\left([0,\infty) \times (y^{1-\gamma/2}, \infty)\right)\right\} \geq b_1 y^{-\nu}$,

v) $\tilde{\mathbb{G}}_{0,y}\left((-\infty, y) \times (0, \exp[|y|^\gamma - 1/2])\right) > 1 - b_2|y|^{-\psi}$,

and for large $y < 0$,

iv′) $\max\left\{\tilde{\mathbb{G}}_{0,y}\left((-\infty, y + \sigma_0|y|^{\gamma/2}] \times [\sigma_0, \infty)\right), \tilde{\mathbb{G}}_{0,y}\left((-\infty, 0] \times (|y|^{1-\gamma/2}, \infty)\right)\right\} \geq b_1|y|^{-\nu}$,

v′) $\tilde{\mathbb{G}}_{0,y}\left((y, \infty) \times (0, \exp[|y|^\gamma - 1/2])\right) > 1 - b_2|y|^{-\psi}$,

then $f_0(y|x) \in KL(\tilde{\Pi}^f)$.

The proof is in the Appendix. The conditions $v)$ and $v')$ require that the tails of $\mathbb{G}_{0,y}$ not decay faster than a polynomial rate for σ_y. Now that we have conditions under which the KL support condition is satisfied, we can apply them to the Gaussian model with a conjugate base measure.

Asymptotic Unbiasedness for Conjugate Base Measures. We use the results of Theorem 9 and Theorem 14 to show asymptotic unbiasedness of conjugate base measures for the Gaussian model. The conjugate base measure $\mathbb{G}_0$ for the covariates of the location-scale Gaussian mixture model is

$$\begin{aligned} \sigma_i^{-2} &\sim \mathrm{Gamma}(r_i, \lambda_i), & i &= 1, \dots, d, \\ \mu_i|\sigma_i &\sim N(\nu_i, \xi_i\, \sigma_i^2), & i &= 1, \dots, d. \end{aligned}$$

Define both $\mathbb{G}_{0,x}$ and $\tilde{\mathbb{G}}_{0,x}$ in that manner. In either of these cases, for a fixed x we can satisfy conditions $i)$ and $ii)$ of Theorem 14.

The matter of $\mathbb{G}_{0,y}$ is somewhat more complicated; Tokdar (2006) shows that if $\tilde{\mathbb{G}}_{0,y}$, the base distribution for the location-scale mixture of the y component satisfies

$$\begin{aligned} \sigma_y^{-2} &\sim \mathrm{Gamma}(r_y, \lambda_y), \\ \eta|\sigma_y &\sim N(0, \xi_y\, \sigma_y^2), \end{aligned} \tag{14}$$

and $r_y \in (1/2, 1)$, then Theorem 14 is satisfied and the mixture is weakly consistent at x. Using the fact that the sum of Gaussian random variables is also Gaussian, if we let $\mathbb{G}_{0,y}$ be defined as

$$\sigma_y^{-2} \sim \text{Gamma}(r_y, \lambda_y),$$
$$\beta_0, \dots, \beta_d | \sigma_y \sim N(0, \sigma_y^2 \Sigma_\beta), \tag{15}$$

then the location parameter η has the distribution

$$\eta | \sigma_y \sim N(0, \sigma_y^2 \tilde{x}^T \Sigma_\beta \tilde{x}),$$

where $\tilde{x}^T = [1, x_1, \dots, x_d]$. Therefore, a conjugate prior for $\mathbb{G}_{0,y}$ also satisfies Theorem 14, provided that $r_y \in (1/2, 1)$.

Let $\mathcal{D}$ be the closed hypercube of the previous paragraph. If we extend this to all $x \in \mathcal{D}$, we see that DP-GLM is asymptotically unbiased for all $x \in \text{int}(\mathcal{D})$. Therefore, if the covariates have a compact true distribution, asymptotic unbiasedness holds for all x except those on the border of the sampling region.

Extensions of Gaussian Model Results. Lemma 13 and Theorem 14 rely heavily on assumptions about the distribution of Y, but not the distribution of X. Lemmas 13, A-1 and A-2, along with Theorem 14 can easily be modified for covariates modeled with categorical and count data, along with a conjugate base measure choice for $\mathbb{G}_{0,x}$. An example would be the categorical covariates of the Poisson model with a Dirichlet distribution prior. See the Appendix for a sketch of the proof extending the Gaussian model to categorical and count covariates.

Extensions to different response types remain an open question.

5. Numerical Results

We compare the performance of DP-GLM regression to other regression methods. We chose data sets to illustrate the properties of DP-GLM.

Data Sets. We selected three data sets with continuous response variables. They highlight various difficulties within regression.

- **Cosmic Microwave Background (CMB) Bennett et al. (2003).** The data set consists of 899 observations which map positive integers $\ell = 1, 2, \dots, 899$, called 'multipole moments,' to the power spectrum C_ℓ. Both the covariate and response are considered continuous. The data pose challenges because they are highly nonlinear and heteroscedastic. Since this data set is only two dimensions, it allows us to easily demonstrate how the various methods approach estimating a mean function while dealing with non-linearity and heteroscedasticity.

- **Concrete Compressive Strength (CCS) Yeh (1998).** The data set has eight covariates: the components cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate and fine aggregate, all measured in kg per m^3, and the age of the mixture in days; all are continuous. The response is the compressive strength of the resulting concrete, also continuous. There are 1,030 observations. The data have relatively little noise. Difficulties arise from the moderate dimensionality of the data.

- **Solar Flare (Solar) Bradshaw (1989).** The response is the number of solar flares in a 24 hour period in a given area; there are 11 categorical covariates. 7 covariates are Bernoulli: time period (1969 or 1978), activity (reduced or unchanged), previous 24 hour activity (nothing as big as M1, one M1), historically complex (Y/N), recent historical complexity (Y/N), area (small or large), area of the largest spot (small or large). 4 covariates are multinomial: class, largest spot size, spot distribution and evolution. The response is the sum of all types of solar flares for the area. There are 1,389 observations. Difficulties are created by the moderately high dimensionality, categorical covariates and count response. Few regression methods can appropriately model this data.

Competitors. The competitors represent a variety of regression methods; some methods are only suitable for certain types of regression problems.

- **Naive Ordinary Least Squares (OLS).** A parametric method that often provides a reasonable fit when there are few observations. Although OLS can be extended for use with any set of basis functions, finding basis functions that span the true function is a difficult task. We naively choose $[1\, X_1\, \dots\, X_d]^T$ as basis functions. OLS can be modified to accommodate both continuous and categorical inputs, but it requires a continuous response function.

- **Regression Trees (Tree).** A nonparametric method generated by the Matlab function classregtree. It accommodates both continuous and categorical inputs and any type of response.

- **Gaussian Processes (GP).** A nonparametric method that can accommodate only continuous inputs and continuous responses. GPs were generated in Matlab by the program gpr of Rasmussen and Williams (2006). It is suitable only for continuous responses and covariates.

- **Basic DP Regression (DP Base).** Similar to DP-GLM, except the response is a function only of β_0, rather than $\beta_0 + \sum \beta_i x_i$. While suitable for any type of covariate or response, it is inferior to DP-GLM.

- **Poisson GLM (GLM).** A Poisson generalized linear model, used on the Solar Flare data set. It is suitable for count responses.

A comparison of the methods suitable for regression on the Gaussian model (continuous covariates and a continuous response) are given in Figure 2 for CMB data. Naive ordinary least squares tries to fit a straight line to the data; this is often inappropriate when the data are not linear. Regression trees fit the data with functions that are locally constant; issues are bandwidth selection, particularly in the presence of heteroscedasticity. Both DP-GLM and Gaussian processes fit a continuous curve to the data; challenges arise in both methods with selection of priors. Additionally, a model must be chosen for DP-GLM and adjustments have to be made to accommodate heteroscedasticity with Gaussian processes.

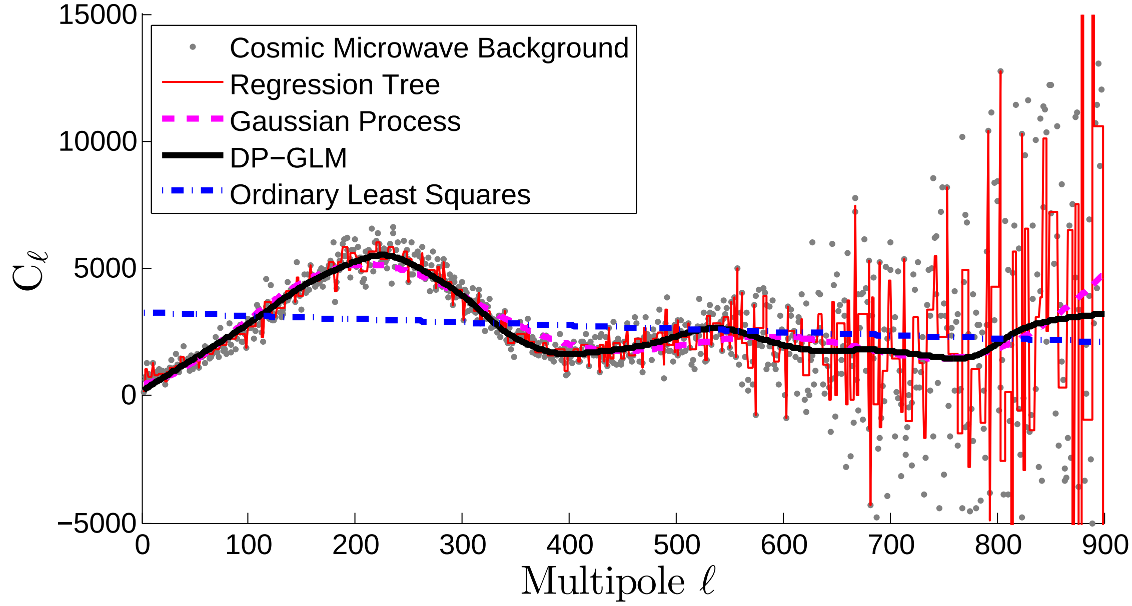

Figure 2: Comparison of DP-GLM, Gaussian process regression, tree regression and ordinary least squares on the CMB dataset. Tree regression over-fits the data much more than the other methods.

Cosmic Microwave Background (CMB) Results. The CMB dataset was chosen to demonstrate the modeling results for typical regression methods because it is both highly non-linear and easily viewed. OLS, regression trees and Gaussian processes were compared to the DP-GLM Gaussian model on this data set. Both DP-GLM and Gaussian processes fit the data well, while OLS ignores the non-linearity of the data set, and tree regression over-fits the noise; see Figure 2.

Mean absolute ($L1$) error and mean squared ($L2$) error for 5, 10, 30, 50, 100, 250, and 500 training data were computed using 10 random subset selections for each amount of data. A conjugate base measure was used for DP-GLM. Results are given in Figure 3. Gaussian processes perform poorly with small amounts of training data; regression trees over-fit, leading to large $L2$ errors. DP-GLM performs well at all levels.

Concrete Compressive Strength (CCS) Results. The CCS dataset was chosen because of its moderately high dimensionality and continuous covariates and response. The continuous variables allowed many common regression techniques to be compared to the DP-GLM Gaussian model. We also included a basic DP regression technique (location/scale DP) on this data set. The location/scale DP models the response with only constant; that is,

$$Y_i | x_i, \theta_i \sim \mu_{i,y}.$$

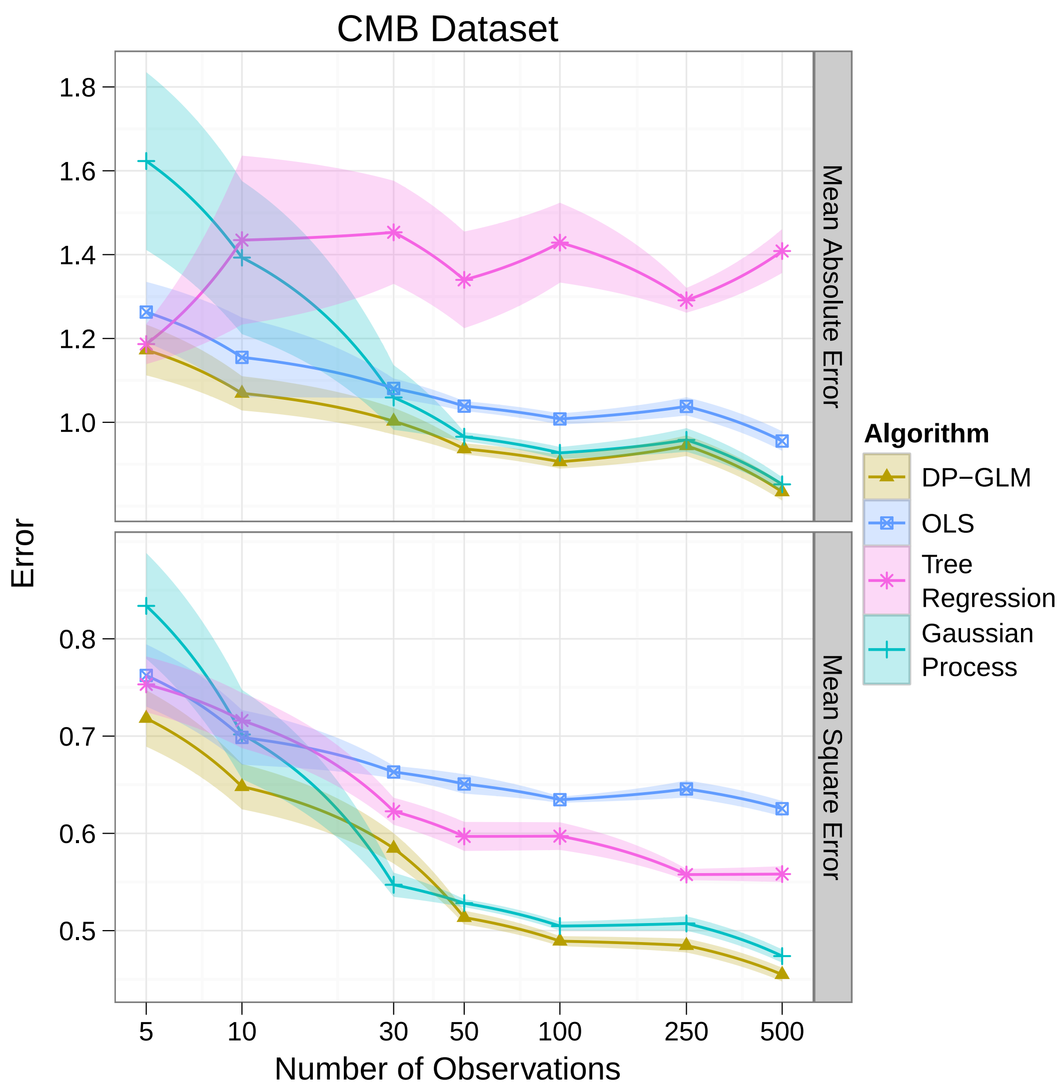

Figure 3: The average mean absolute error (top) and mean squared error (bottom) for ordinary least squares (OLS), tree regression, Gaussian processes and DP-GLM on the CMB data set. The data were normalized. Mean +/− one standard deviation are given for each method.

This model was included on this data set to demonstrate its weakness; without the GLM response, the model cannot interpolate well in higher dimensions, leading to poor predictive performance.

Mean absolute ($L1$) error and mean squared ($L2$) error for 20, 30, 50, 100, 250, and 500 training data were computed using 10 random subset selections for each amount of data. Gaussian base measures were used for the location components of the DPs, while log-Gaussian base measures were used for the scale parameters. Results are given in Figure 4.

As expected, DP Base did quite poorly. DP-GLM did well with few training data, while regression trees and particularly Gaussian processes did well with substantial amounts of training data.

Solar Flare Results. The Solar dataset was chosen to demonstrate the flexibility of DP-GLM. Many regression techniques cannot accommodate categorical covariates and most cannot accommodate a count-type response. Gaussian processes are defined by a mean vector and covariance matrix, which cannot be meaningfully created for categorical data.

The competitors on this dataset are tree regression and a Poisson GLM. Frequentist estimation of the Poisson GLM produced unstable estimates as some of the covariates are colinear. This problem was addressed by estimating the parameters in a Bayesian manner with Gaussian priors. Mean absolute ($L1$) error and mean squared ($L2$) error for 50, 100, 200, 500, and 800 training data were computed using 10 random subset selections for each amount of data. In DP-GLM, a Gaussian base measure was used for the GLM parameters and a Dirichlet for the covariates. Results are given in Figure 5.

Tree regression has a relatively low mean absolute error, while it has a high mean squared error. The Poisson GLM has the opposite, relatively low mean squared error and relatively high mean absolute error. Only the DP-GLM does well under both metrics.

Discussion. DP-GLM has flexibility that is not offered by most regression methods. It does well on data sets with heteroscedastic errors because it fundamentally incorporates them; error parameters ($\sigma_{i,y}$) are included in the DP mixture. DP-GLM is comparatively robust with small amounts of data because in that case it tends to put all (or most) of the observations into one cluster; this effectively produces a linear regression, but eliminates outliers by placing them into their own (low-weighted) clusters.

The comparison between DP-GLM regression and basic DP regression is illustrative. We compared basic DP regression only on the CCS data set because it has a large number of covariates. Like kernel smoothing, basic DP regression struggles in high dimensions because it cannot efficiently interpolate values between observations. The GLM component effectively eliminates this problem.

The diversity of the data sets demonstrates the adaptability of the DP-GLM. Only tree regression was able to work on all of the data sets as well, and the DP-GLM has many desirable properties that tree regression does not, such as a smooth mean function estimate. Moreover, the DP-GLM usually outperformed tree regression.

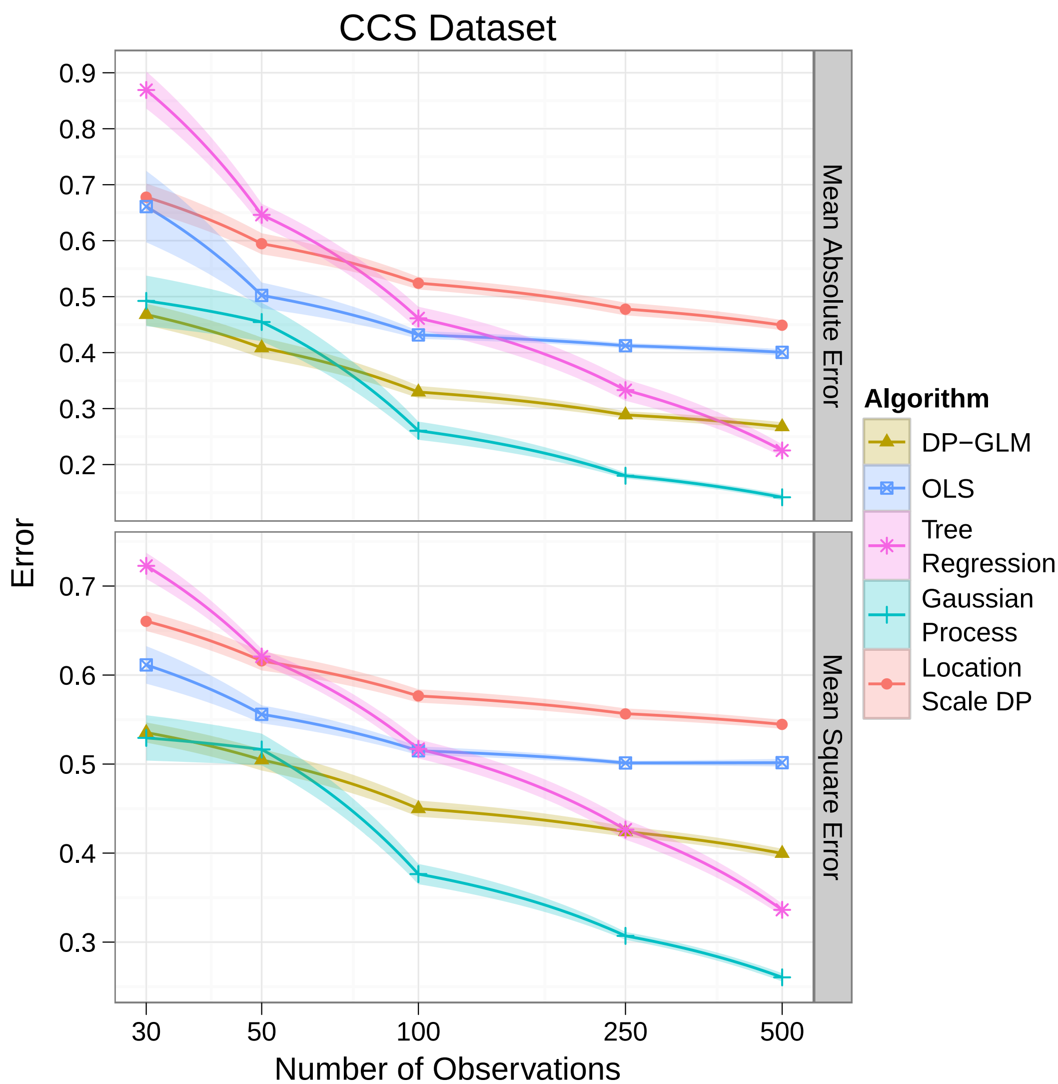

Figure 4: The average mean absolute error (top) and mean squared error (bottom) for ordinary least squares (OLS), tree regression, Gaussian processes, location/scale DP and the DP-GLM Poisson model on the CCS data set. The data were normalized. Mean +/− one standard deviation are given for each method.

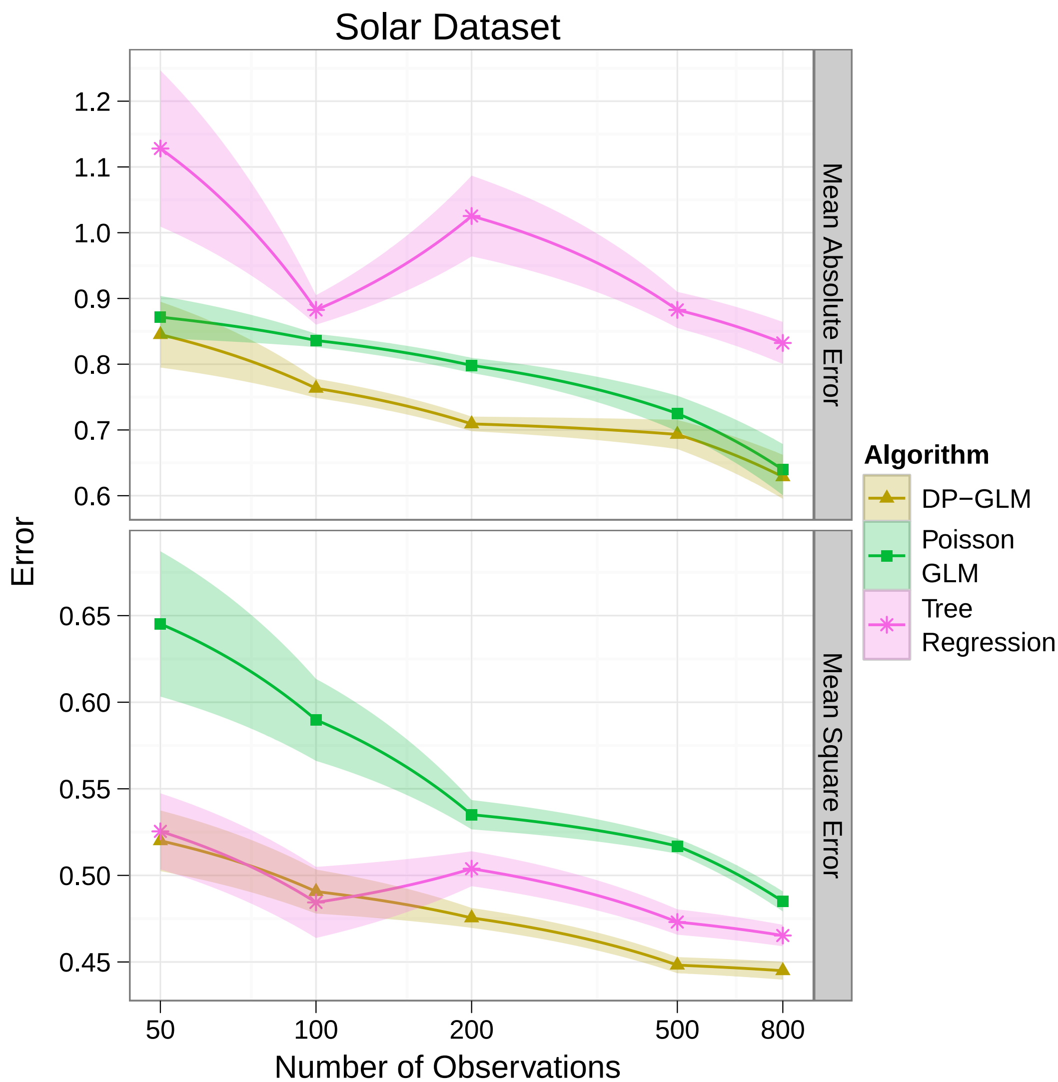

Figure 5: The average mean absolute error (top) and mean squared error (bottom) for tree regression, a Poisson GLM (GLM) and DP-GLM on the Solar data set. Mean +/− one standard deviation are given for each method.

6. Conclusions and Future Work

We have developed the DP-GLM, a flexible model for nonparametric Bayesian regression. We conditions for asymptotic unbiasedness and gave a specific case for when it is asymptotically unbiased with conjugate priors. We then tested the DP-GLM on a variety of data sets to demonstrate its flexibility in the face of common statistical problems such as data type (continuous, count, categorical, etc), heteroscedastic errors and moderately high dimensionality.

The DP-GLM offers a framework for a variety of predictive models. Our future work includes using a shape-restricted function in place of the GLM to create a regression model for functions that are known to be concave in one dimension. Such regression problems often arise in a simulation-optimization setting where one dimension is resource amount, the others are state and the response is resource value. The methods provided in this paper give a way to construct and theoretically analyze many flexible yet complicated models with Dirichlet process priors.

Appendix

A-1 Proofs for Section 4.3

Proof [Proof of Lemma 12.] $\mathbb{G}_0$ is a measure on the space $\mathbb{R}^{d+1} \times \mathbb{R}_+^{d+1} \times \mathbb{R}^d$. The induced parameters are on the space $\mathbb{R} \times \mathbb{R}_+^{d+1} \times \mathbb{R}^d$. Suppose $P \sim DP(\alpha, \mathbb{G}_0)$ and $\tilde{P}$ is the random measure on the induced parameters. Define $\tilde{\mathbb{G}}_0$ as above. Fix a finite partition of $\mathbb{R} \times \mathbb{R}_+^{d+1} \times \mathbb{R}^d$, $B_1, \dots, B_k$. For convenience of notation, break B_i into $B_{i,1} \times B_{i,2} \times B_{i,3}$ along the components of the product space. Define $\tilde{B}_{i,1} = \{\beta_{0:d} \,:\, \beta_0 + \beta_1 x_1 + \dots + \beta_d x_d \in B_{i,1}\}$. Then,

$$\begin{aligned}
\left(\tilde{P}(B_1), \dots, \tilde{P}(B_k)\right) &= \left(P(\tilde{B}_{1,1} \times B_{1,2} \times B_{1,3}), \dots, P(\tilde{B}_{k,1} \times B_{k,2} \times B_{k,3})\right) \\
&\sim \mathrm{Dir}\left(\alpha\mathbb{G}_0(\tilde{B}_{1,1} \times B_{1,2} \times B_{1,3}), \dots, \alpha\mathbb{G}_0(\tilde{B}_{k,1} \times B_{k,2} \times B_{k,3})\right) \\
&= \mathrm{Dir}\left(\tilde{\mathbb{G}}_0(B_1), \dots, \tilde{\mathbb{G}}_0(B_k)\right).
\end{aligned}$$

■

This extends Remark 3 of Ghosal et al. (1999) to our multidimensional setting. It is used in the proof of Lemma 13.

Lemma A-1 *Let $f_0(x, y)$ be a density on $\mathbb{R}^d \times \mathbb{R}$. Suppose that there exists a closed set $\mathcal{D} \subset \mathbb{R}^d \times \mathbb{R}$ such that there exists $c > 0$ where $f_0(x, y) > c$ for every $(x, y) \in \mathcal{D}$ and the x domain of $\mathcal{D}$ is a hypercube. Suppose that $f_0(x, y) = 0$ if x is outside of $\mathcal{D}$, and assume that for y outside of $\mathcal{D}$, $f_0(x, y)$ is increasing in y below $\mathcal{D}$, and decreasing in y above $\mathcal{D}$. Define for $h = (h_1, \dots, h_d, h_y)$,*

$$f_{0,h}(x, y) = \int_{\mathbb{R}^{d+1}} f_0(\mu_1, \dots, \mu_d, \eta) \phi_{h_y}(y - \eta) \prod_{i=1}^{d} \phi_{h_i}(x - \mu_i) d\mu_1 \dots d\mu_d d\eta.$$

If

$$\int_{\mathbb{R}^{d+1}} f_0(x,y)\log f_0(x,y)dx_1\dots dx_d dy < \infty,$$

then for every $\epsilon > 0$ *there exists* $h = (h_1, \dots, h_d, h_y) > \mathbf{0}$ *such that,*

$$\int_{\mathbb{R}^{d+1}} f_0(x,y)\log \frac{f_0(x,y)}{f_{0,h}(x,y)}dxdy < \epsilon.$$

Proof For every $x \in \mathcal{D}$, assume that the minimum width of the interval for which $y \in \mathcal{D}$ is at least $a_1(x)$ and at most $a_2(x)$ for some functions $0 < a_1(x) \leq a_2(x)$. If this is not true, we can simply change the x domain of $\mathcal{D}$ until it is true. Let $\overline{a}$ be the maximum value for $y \in \mathcal{D}$ and $\underline{a}$ be the minimum value. Let $(\overline{x}_1, \dots, \overline{x}_d)$ be the upper left corner of the x hypercube, and $(\underline{x}_1, \dots, \underline{x}_d)$ be the lower right. Now, choose $h_0 = (h_{0,y}, h_{0,1}, \dots, h_{0,d})$ such that $N(0, h_{0,i})$ give probability $1/2$ to the x_i domain of $\mathcal{D}$, and $N(0, h_{0,y})$ gives probability b_1 to $(0, a_1)$ and probability b_2 to $(0, a_2)$, with $0 < b_1 \leq b_2 < 1$. Set $h < h_0$. If $(x, y) \in \mathcal{D}$,

$$\begin{aligned}
f_{0,h}(x,y) &\geq \int_{\mathcal{D}} f_0(\mu_1,\dots,\mu_d,\eta)\phi_{h_y}(y-\eta)\prod_{i=1}^d \phi_{h_i}(x_i-\mu_i)d\mu_1\dots d\mu_d d\eta,\\
&\geq c\left(\Phi((\overline{a}-y)/h_y)+\Phi((y-\underline{a})/h_y)\right)\prod_{i=1}^d \left(\Phi((\overline{x}_i-x_i)/h_i)+\Phi((x_i-\underline{x}_i)/h_i)\right),\\
&\geq c\frac{b_1}{2^d}.
\end{aligned}$$

If $y \notin \mathcal{D}$ and $y > a_2(x)$ given x, then

$$\begin{aligned}
f_{0,h}(x,y) &\geq \int_{\mu\in\mathcal{D}}\int_{\eta>a_2(x)} f_0(\mu_1,\dots,\mu_d,\eta)\phi_{h_y}(y-\eta)\prod_{i=1}^d \phi_{h_i}(x_i-\mu_i)d\mu_1\dots d\mu_d d\eta,\\
&\geq f_0(x,y)\left(1/2+\Phi(\overline{a}/h_y)-1\right)\prod_{i=1}^d \left(\Phi((\overline{x}_i-x_i)/h_i)+\Phi((x_i-\underline{x}_i)/h_i)\right),\\
&\geq f_0(x,y)\frac{b_2}{2^d}.
\end{aligned}$$

Using a similar argument when $y < a_1(x)$ for a given x, we have a function

$$g(x,y) = \begin{cases} \log(2^d f_0(x,y)/(b_1 c)), & \text{if } (x,y)\in\mathcal{D}\\ \log(2^d/b_2), & \text{if } (x,y)\notin\mathcal{D},\end{cases}$$

dominates $\log(f_0(x,y)/f_{0,h}(x,y))$ for $h < h_0$ and $x \in \mathcal{D}$ and is $\mathbb{P}_{f_0}$-integrable (outside $\mathcal{D}$ for x, $f_0(x,y)$ is assumed to have 0 density). For all continuity points (x,y) of f_0, $f_0(x,y)/f_{0,h}(x,y) \to 1$ as $h \to 0$; since $\int f_0 \log(f_0/f_{0,h}) \geq 0$, an application of Fatou's lemma shows that $\int f_0 \log(f_0/f_{0,h}) \to 0$ as $h \to 0$. ■

We can now prove Lemma 13.

Proof [Proof of Lemma 13.] This proof strongly follows the proof of Lemma 3.2 in Tokdar (2006).

Fix x. Fix $0 < \tau < 1$ and $\epsilon > 0$. By assumption, find $\mathcal{A}$ and y_0 such that $\tilde{\Pi}^P(\mathcal{A}) > 1-\tau$ for any $P \in \mathcal{A}$,

$$\int_{|y|>y_0} f_0(x,y) \log \frac{f_0(x,y)}{f(x,y)} dy < \epsilon \frac{f_0(x)}{4}. \tag{A-1}$$

Now consider the region where $f_0(x,y)$ is truncated to the interval $[-y_0, y_0]$. From Lemma A-1, there exist a $h = (h_1, \ldots, h_d, h_y) > 0$ and $\underline{k}_1, \overline{k}_1, \ldots, \underline{k}_d, \overline{k}_d, \underline{k}_y, \overline{k}_y$ such that

$$\int_{-y_0}^{y_0} f_0(x,y) \log \frac{f_0(x,y)}{\int_{\underline{k}_1}^{\overline{k}_1} \int_{\underline{k}_y}^{\overline{k}_y} f_0(\mu_1, \ldots, \mu_d, \eta) \phi_{h_y}(y-\eta) \prod_{i=1}^d \phi_{h_i}(x_i - \mu_i) d\mu_1 \ldots d\eta} dy < \epsilon \frac{f_0(x)}{4}, \tag{A-2}$$

and

$$\left| \log \frac{f_0(x)}{\int_{\underline{k}_1}^{\overline{k}_1} \ldots \int_{\underline{k}_d}^{\overline{k}_d} f_0(\mu_1, \ldots, \mu_d) \prod_{i=1}^d \phi_{h_i}(x_i - \mu_i) d\mu_1 \ldots d\mu_d} \right| < \frac{\epsilon}{8}. \tag{A-3}$$

Let P_0 be a measure over the location and scale for all components on $\mathbb{R}^{d+1} \times \mathbb{R}_+^{d+1}$, such that $dP_0 = f_0(x,y) \times \delta_{h_1} \times \cdots \times \delta_{h_y}$. Fix $\kappa > 0$ and find $\lambda > 0$ such that $1 - \lambda/(\kappa^2(1-\lambda)^2) > \tau$. Choose a compact set K such that $[\underline{k}_1, \overline{k}_1] \times \cdots \times [\underline{k}_y, \overline{k}_y] \times \{h_1\} \times \cdots \times \{h_y\} \subset K$, $\tilde{\mathbb{G}}_0(K) > 1 - \lambda$ and $P_0(K) > 1 - \lambda$. Let $\mathcal{B} = \{P : |P(K)/P_0(K) - 1| < \kappa\}$. Since the prior on P is a Dirichlet process, $P(K) \sim \text{Beta}(\alpha \tilde{\mathbb{G}}_0(K), \alpha \tilde{\mathbb{G}}_0(K^c))$. By applying Markov's inequality,

$$\tilde{\Pi}^P(\mathcal{B}) \geq 1 - \frac{\mathbb{E}(P(K) - P_0(K))^2}{\kappa^2 P_0(K)^2} \geq 1 - \frac{\lambda}{\kappa^2(1-\lambda)^2} > \tau.$$

Therefore, $\tilde{\Pi}^P(\mathcal{A} \cap \mathcal{B}) > 0$.

Following Tokdar (2006) and Ghosal et al. (1999), we can find a set $\mathcal{C}$ such that $\tilde{\Pi}^P(\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}) > 0$ and $P \in \mathcal{B} \cap \mathcal{C}$ implies that for suitable κ,

$$\int_{-y_0}^{y_0} f_0(x,y) \log \frac{\int_K \phi_{\sigma_y}(y-\eta) \prod_{i=1}^d \phi_{\sigma_i}(x_i - \mu_i) dP_0}{\int_K \phi_{\sigma_y}(y-\eta) \prod_{i=1}^d \phi_{\sigma_i}(x_i - \mu_i) dP} dy < \frac{\kappa}{1-\kappa} + 2\kappa < \epsilon \frac{f_0(x)}{4}, \tag{A-4}$$

and

$$\left| \log \frac{\int_{K_x} \prod_{i=1}^d \phi_{\sigma_i}(x_i - \mu_i) dP_{0,x}}{\int_{K_x} \prod_{i=1}^d \phi_{\sigma_i}(x_i - \mu_i) dP_x} \right| < \frac{\kappa}{1-\kappa} + 2\kappa < \frac{\epsilon}{8}. \tag{A-5}$$

Now let $f = \phi^{d+1} * P$, where $P \in \mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$,

$$\int f_0(y|x) \log \frac{f_0(y|x)}{f(y|x)} dy \le \frac{1}{f_0(x)} \int_{|y|>y_0} f_0(x,y) \log \frac{f_0(x,y)}{f(x,y)} dy$$
$$+ \left| \log \frac{f_0(x)}{\int_{\underline{k}_1}^{\bar{k}_1} \dots \int_{\underline{k}_d}^{\bar{k}_d} f_0(\mu_1, \dots, \mu_d) \prod_{i=1}^d \phi_{h_i}(x_i - \mu_i) d\mu_1 \dots d\mu_d} \right|$$
$$+ \left| \log \frac{\int_{K_x} \prod_{i=1}^d \phi_{\sigma_i}(x_i - \mu_i) dP_{0,x}}{\int_{K_x} \prod_{i=1}^d \phi_{\sigma_i}(x_i - \mu_i) dP_x} \right|$$
$$+ \frac{1}{f_0(x)} \left(\int_{-y_0}^{y_0} f_0(x,y) \log \frac{\int_K \phi_{\sigma_y}(y-\eta) \prod_{i=1}^d \phi_{\sigma_i}(x_i - \mu_i) dP_0}{\int_K \phi_{\sigma_y}(y-\eta) \prod_{i=1}^d \phi_{\sigma_i}(x_i - \mu_i) dP} dy \right.$$
$$\left. + \int_{-y_0}^{y_0} f_0(x,y) \log \frac{f_0(x,y)}{\int_{\underline{k}_1}^{\bar{k}_1} \dots \int_{\underline{k}_y}^{\bar{k}_y} f_0(\mu_1, \dots, \mu_d, \eta) \phi_{h_y}(y-\eta) \prod_{i=1}^d \phi_{h_i}(x_i - \mu_i) d\mu_1 \dots d\eta} dy \right)$$
$$< \epsilon,$$

by equations (A-1), (A-2), (A-3), (A-4) and (A-5).

■

Proof [Proof of Theorem 14.] To prove this Theorem, we need to show that such an f_0 and $\tilde{\mathbb{G}}_0$ satisfy Lemma 13. This proof is very similar to the proof of Theorem 3.3 in Tokdar (2006). Let $w(x) = \exp(-x^\gamma)$, $x \ge 0$. Define a class of subsets of $\mathbb{R}^{d+1} \times \mathbb{R}_+^{d+1}$ indexed only by $y \in \mathbb{R}$ (since x is fixed) as follows,

$$K(y)_c = \Big\{ (\mu_1, \sigma_1) \times \cdots \times (\eta, \sigma_y) \in \mathbb{R}^{d+1} \times \mathbb{R}_+^{d+1} \, : \, \phi_{\sigma_i}(x_i - \mu_i) \ge c,$$
$$\phi_{\sigma_y}(y - \eta) \ge (2\pi)^{-1/2} w(|y|) \Big\}.$$

Let $f = \phi^{d+1} * P$,

$$\int_{|y|>y_0} f_0(x,y) \log \frac{f_0(x,y)}{f(x,y)} dy$$
$$\le \int_{|y|>y_0} f_0(x,y) \log \frac{f_0(x,y)}{\int_{K(y)_c} \phi_{\sigma_y}(y-\eta) \prod_{i=1}^d \phi_{\sigma_i}(x_i - \mu_i) dP} dy,$$
$$\le \int_{|y|>y_0} f_0(x,y) \log \frac{f_0(x,y)}{(2\pi)^{-2/2} w(|y|) c^d P(K(y)_c)} dy,$$
$$\le \int_{|y|>y_0} f_0(x,y) \left\{ \log f_0(x,y) + |y|^\gamma + \log \frac{(2\pi)^{1/2}}{c^d P(K(y)_c)} \right\} dy. \qquad \text{(A-6)}$$

We can show that equation (A-6) can be made arbitrarily small for suitably large y_0 if $P(K(y)_c) > a_1 \exp(-a_2 |y|^\gamma)$ for all $|y| > y_0$ for some fixed constants $a_1, a_2 > 0$. To show this, we slightly modify Lemma 3.3 of Tokdar (2006).

Lemma A-2 (Modification of Lemma 3.3 of Tokdar (2006)) *For any $\tau > 0$ there exists a $y_0 > 0$, a $c > 0$ and a set $\mathcal{A}$ with $\tilde{\Pi}(\mathcal{A}) > 1 - \tau$ such that $P \in \mathcal{A} \Rightarrow P(K(y)_c) \geq (1/2)b_1 \exp(-2|y|^\gamma/b_2)$ for all $|y| > y_0$.*

The proof follows the lines of the original proof, but relies on the independence of the base measure $\tilde{\mathbb{G}}_0$ for the x and y dimensions. ■

A-2 Extension of Section 4.3 to Categorical Covariates

Assume that X is a set of categorical random variables taking values in the finite set $\mathcal{D}$. To extend Theorem 14 to include categorical covariates, we must define terms, redefine how the random density is constructed and then modify assumptions. Let $\mathcal{M}(x|p)$ denote the multinomial probability mass function with one outcome and probability vector p evaluated at x. Let $\mathcal{P}$ be the probability simplex for p.

For simplicity, we assume that there is only one covariate; extension to multiple covariates is straightforward. We construct the random density f_P in the following manner,

$$f_P(x, y) = \int_{\mathcal{P}\times\mathbb{R}\times\mathbb{R}_+} \phi_{\sigma_y}(y - \eta)\mathcal{M}(x|p)P(dp, d\eta, d\sigma_y).$$

We need the assumptions:

A1) $f_0(x) > 0$ for every $x \in \mathcal{D}$,

A2) $\int_{\mathcal{D}\times\mathbb{R}} f_0(x, y) \log f_0(x, y)dxdy < \infty$,

A3) for any f_P with $P \in \mathcal{A}$,

$$\int_{|y|>y_0} f_0(x, y) \log \frac{f_0(x, y)}{f(x, y)} dy < \epsilon.$$

Now, we sketch how the inclusion of a categorical covariate affects Lemmas A-1, 13 and Theorem 14. We do not need to change Lemma A-2.

Categorical Covariates in Lemma A-1. We modify this Lemma in the following manner:

L1) Suppose that there exists a $c > 0$ and a y_0 such that for every $x \in \mathcal{D}$ and $|y| \leq y_0$, $f_0(x) > c$ and $f_0(y|x) > c$,

L2) Define

$$f_{0,h}(x, y) = \int_{\mathbb{R}} f_0(x, \eta)\phi_h(y - \eta)d\eta.$$

The proof of the Lemma proceeds in the same manner as before.

Categorical Covariates in Lemma 13. First, modify Lemma 13 by replacing conditions *i*), *ii*) and *iv*) with assumptions A1), A2) and A3), respectively. These assumptions are analogous to those in the continuous setting.

Next, because Lemma A-2 is not completely analogous, we need to change equations (A-2) and (A-3). Using Lemma A-2, we can choose an h and $\underline{k}, \overline{k}$ such that,

$$\int_{-y_0}^{y_0} f_0(x,y) \log \frac{f_0(x,y)}{\int_{\underline{k}}^{\overline{k}} f_0(x,\eta)\phi_h(y-\eta)d\eta} dy < \epsilon \frac{f_0(x)}{8}. \tag{A-7}$$

. Using the fact that for every $\delta > 0$ we can choose a closed set $\mathcal{E} \subset \mathcal{P}$ that contains a non-empty interior such that $|f_0(x) - \int_{\mathcal{E}} \mathcal{M}(x|p)dp| < \delta$, we can choose $\mathcal{E}$ such that,

$$\int_{-y_0}^{y_0} f_0(x,y) \log \frac{\int_{\underline{k}}^{\overline{k}} f_0(x,\eta)\phi_h(y-\eta)d\eta}{\int_{\underline{k}}^{\overline{k}} \int_{\mathcal{E}} f_0(\eta|x)\phi_h(y-\eta)\mathcal{M}(x|p)d\eta dp} dy < \epsilon \frac{f_0(x)}{8}, \tag{A-8}$$

and

$$\left| \log \frac{f_0(x)}{\int_{\mathcal{E}} \mathcal{M}(x|p)dp} \right| < \frac{\epsilon}{8}. \tag{A-9}$$

Equations (A-7) and (A-8) take the place of equation (A-2), while equation (A-9) takes the place of equation (A-3).

Let p_0 be the multinomial parameters of $f_0(x)$ (which is multinomial because X is categorical), and P_0 be a measure over $\mathbb{R} \times \mathbb{R}_+ \times \mathcal{P}$. Define $dP_0 = f(y|x) \times \delta_h \times \delta_{p_0}$. Choose compact K such that $[\underline{k}, \overline{k}] \times h \times p_0 \subset K$. The rest of the proof follows.

Categorical Covariates in Theorem 14. Modify Theorem 14 by replacing assumptions *i*) and *ii*) with A1) and A2), respectively. Now we simply need to change the way $K(y)_c$ is defined,

$$K(y)_c = \left\{ p \times (\eta, \sigma_y) \in \mathcal{P} \times \mathbb{R} \times \mathbb{R}_+ : \mathcal{M}(x|p) \geq c, \phi_{\sigma_y}(y-\eta) \geq (2\pi)^{-1/2} w(|y|) \right\}.$$

The rest of the proof follows.

Categorical Covariates with Conjugate Base Measure. The tail properties of $\tilde{G}_{0,y}$ are the limiting factor in Theorem 14; for the covariate base measure, the only major requirement is that every open set that approximates $f_0(x) > c$ arbitrarily well has a positive measure. For every $\delta > 0$, a conjugate base measure (Dirichlet) places positive measure on $\{p : |f_0(x) - \mathcal{M}(x|p)| < \delta\}$ when $f_0(x) > c$.

A-3 Extension of Section 4.3 to Count Covariates

This subsection proceeds in a similar manner to the previous subsection. Assume that X is a count random variable taking values in $\mathbb{Z}_+$. Let $\mathcal{P}(x|\lambda)$ denote the Poisson probability mass function with parameter λ evaluated at x.

For simplicity, we assume that there is only one covariate; extension to multiple covariates is straightforward. We construct the random density f_P in the following manner,

$$f_P(x,y) = \int_{\mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+} \phi_{\sigma_y}(y-\eta)\mathcal{P}(x|\lambda)P(d\lambda, d\eta, d\sigma_y).$$

We need the assumptions:

B1) $f_0(x) > 0$ for every $x \in \mathcal{D}$, where $\mathcal{D}$ is a bounded set,

B2) $\int_{\mathcal{D}\times\mathbb{R}} f_0(x,y)\log f_0(x,y)dxdy < \infty$,

B3) for any f_P with $P \in \mathcal{A}$,

$$\int_{|y|>y_0} f_0(x,y)\log\frac{f_0(x,y)}{f(x,y)}dy < \epsilon.$$

Modify Lemma A-1 as before. We now explain how to modify Lemma 13 and Theorem 14.

Count Covariates in Lemma 13. Modify Lemma 13 by replacing conditions i), ii) and iv) with assumptions B1), B2) and B3), respectively.

Again, because Lemma A-2 is not completely analogous, we need to change equations (A-2) and (A-3). Using Lemma A-2, we can choose an h and $\underline{k}, \overline{k}$ such that,

$$\int_{-y_0}^{y_0} f_0(x,y)\log\frac{f_0(x,y)}{\int_{\underline{k}}^{\overline{k}} f_0(x,\eta)\phi_h(y-\eta)d\eta}dy < \epsilon\frac{f_0(x)}{8}. \tag{A-10}$$

. Using the fact that for every $\delta > 0$ we can choose $\underline{\ell}, \overline{\ell}$ such that $|f_0(x) - \int_{\underline{\ell}}^{\overline{\ell}} \mathcal{P}(x|\lambda)d\lambda| < \delta$,

$$\int_{-y_0}^{y_0} f_0(x,y)\log\frac{\int_{\underline{k}}^{\overline{k}} f_0(x,\eta)\phi_h(y-\eta)d\eta}{\int_{\underline{k}}^{\overline{k}}\int_{\underline{\ell}}^{\overline{\ell}} f_0(\eta|x)\phi_h(y-\eta)\mathcal{P}(x|\lambda)d\eta d\lambda}dy < \epsilon\frac{f_0(x)}{8}, \tag{A-11}$$

and

$$\left|\log\frac{f_0(x)}{\int_{\underline{\ell}}^{\overline{\ell}} \mathcal{P}(x|\lambda)d\lambda}\right| < \frac{\epsilon}{8}. \tag{A-12}$$

Equations (A-10) and (A-11) take the place of equation (A-2), while equation (A-12) takes the place of equation (A-3).

Let λ_0 be the Poisson parameter of $f_0(x)$, were $f_0(x)$ modeled as a Poisson random variable, and P_0 be a measure over $\mathbb{R}\times\mathbb{R}_+\times\mathbb{R}_+$. Define $dP_0 = f(y|x)\times\delta_h\times\delta_{\lambda_0}$. Note that the $\underline{\ell}, \overline{\ell}$ from above can be chosen such that $\underline{\ell} < \lambda_0 < \overline{\ell}$. Choose compact K such that $[\underline{k}, \overline{k}]\times h\times\lambda_0 \subset K$. The rest of the proof follows.

Categorical Covariates in Theorem 14. Modify Theorem 14 by replacing assumptions i) and ii) with B1) and B2), respectively. Now change the way $K(y)_c$ is defined,

$$K(y)_c = \left\{\lambda\times(\eta,\sigma_y) \in \mathcal{P}\times\mathbb{R}\times\mathbb{R}_+ : \mathcal{P}(x|\lambda) \geq c, \phi_{\sigma_y}(y-\eta) \geq (2\pi)^{-1/2}w(|y|)\right\}.$$

The rest of the proof follows.

Count Covariates with Conjugate Base Measure. The conjugate base measure is a Gamma distribution. Like above, the conjugate distribution places positive measure on all weak sets around $f_0(x)$. Unlike above, our conclusions hold only for queries in a bounded region (this is not a problem for categorical covariates).

References

M. Amewou-Atisso, S. Ghosal, J.K. Ghosh, and RV Ramamoorthi. Posterior consistency for semi-parametric regression problems. *Bernoulli*, pages 291–312, 2003.

C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

A. Barron, M.J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

CL Bennett, M. Halpern, G. Hinshaw, N. Jarosik, A. Kogut, M. Limon, SS Meyer, L. Page, DN Spergel, GS Tucker, et al. First-Year Wilkinson Microwave Anisotropy Probe (WMAP) 1 Observations: Preliminary Maps and Basic Results. *The Astrophysical Journal Supplement Series*, 148(1):1–27, 2003.

P. Billingsley. *Probability and measure*. Wiley India Pvt. Ltd., 2008.

D. Blackwell and J.B. MacQueen. Ferguson distributions via Polya urn schemes. *The Annals Statistics*, 1(2):353–355, 1973.

D.M. Blei and M.I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2005.

G. Bradshaw. UCI machine learning repository, 1989. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

L. Brieman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, pages 434–455, 1998.

M. De Iorio, P. Muller, G.L. Rosner, and S.N. MacEachern. An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, 99(465):205–215, 2004.

P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):1–26, 1986.

J.A. Duan, M. Guindani, and A.E. Gelfand. Generalized spatial Dirichlet process models. *Biometrika*, 94(4):809–825, 2007.

D.B. Dunson, N. Pillai, and J.H. Park. Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):163–183, 2007.

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

M.D. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

J. Fan and I. Gijbels. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 371–394, 1995.

J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC, 1996.

P. Fearnhead. Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, 14(1):11–21, 2004.

T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

J.H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1): 1–141, 1991.

A.E. Gelfand, A. Kottas, and S.N. MacEachern. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471): 1021–1035, 2005.

A. Gelman, J.B. Carlin, and H.S. Stern. *Bayesian data analysis*. CRC press, 2004.

S. Ghosal and A. van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Annals of statistics*, 35(2):697, 2007.

S. Ghosal, JK Ghosh, and RV Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158, 1999.

S. Ghosal, J.K. Ghosh, and A.W. van der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000.

JK Ghosh and RV Ramamoorthi. *Bayesian Nonparametrics*. Springer, 2003.

JE Griffin and M.F.J. Steel. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006.

JE Griffin and MFJ Steel. Bayesian nonparametric modelling with the Dirichlet process regression smoother. Technical report, Technical Report, Institute of Mathematics, Statistics and Actuarial Science, University of Kent, 2007.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, 2009.

J.G. Ibrahim and K.P. Kleinman. Semiparametric Bayesian methods for random effects models. *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 89–114, 1998.

J. Lafferty and L. Wasserman. Rodeo: Sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1):28–63, 2008.

N. Lawrence, M. Seeger, and R. Herbrich. Fast sparse Gaussian process methods: The informative vector machine. *Advances in neural information processing systems*, pages 625–632, 2003.

S.N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3):727–741, 1994.

S.N. MacEachern and P. Muller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, 1989.

S. Mukhopadhyay and A.E. Gelfand. Dirichlet Process Mixed Generalized Linear Models. *Journal of the American Statistical Association*, 92(438):633–639, 1997.

P. Muller, A. Erkanli, and M. West. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79, 1996.

R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

T. Nicoleris and S.G. Walker. On a posterior predictive density sample size criterion. *Scandinavian Journal of Statistics*, 33(2):209–218, 2006.

J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267, 1996.

J. Quiñonero-Candela and C.E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*. Springer, 2006.

A. Rodriguez, D.B. Dunson, and A.E. Gelfand. Bayesian nonparametric functional data analysis through density estimation. *Biometrika*, 96(1):149–162, 2009.

D. Ruppert, SJ Sheather, and MP Wand. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, pages 1257–1270, 1995.

L. Schwartz. On Bayes procedures. *Z. Wahrsch. Verw. Gebiete*, 4(1):10–26, 1965.

J. Sethuraman. A constructive definition of the Dirichlet process. *Statist. Sinica*, 4:639–650, 1994.

B. Shahbaba and R.M. Neal. Nonlinear Models Using Dirichlet Process Mixtures. *Journal of Machine Learning Research*, 10:1829–1850, 2009.

X. Shen. Asymptotic normality of semiparametric and nonparametric posterior distributions. *Journal of the American Statistical Association*, 97(457):222–235, 2002.

X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Annals of Statistics*, pages 687–714, 2001.

A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, 18:1257, 2006.

S. Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhya: The Indian Journal of Statistics*, 67:90–110, 2006.

S. Walker. New approaches to Bayesian consistency. *The Annals of Statistics*, 32(5):2028–2043, 2004.

S.G. Walker, A. Lijoi, and I. Prunster. On rates of convergence for posterior distributions in infinite-dimensional models. *Annals of Statistics*, 35(2):738, 2007.

M. West, P. Muller, and M.D. Escobar. Hierarchical priors and mixture models, with application in regression and density estimation. *Aspects of uncertainty: A Tribute to DV Lindley*, pages 363–386, 1994.

I.C. Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808, 1998.