Journal of Machine Learning Research 12 (2011) 1923-1953 Submitted 9/09; Revised 1/11; Published 6/11

Dirichlet Process Mixtures of Generalized Linear Models

Lauren A. Hannah LH140@DUKE.EDU
Department of Statistical Science
Duke University
Durham, NC 27708, USA

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

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

Editor: Carl Edward Rasmussen

Abstract

We propose Dirichlet Process mixtures of Generalized Linear Models (DP-GLM), a new class of methods for nonparametric regression. Given a data set of input-response pairs, the DP-GLM produces a global model of the joint distribution through a mixture of local generalized linear models. DP-GLMs allow both continuous and categorical inputs, and can model the same class of responses that can be modeled with a generalized linear model. We study the properties of the DP-GLM, and show why it provides better predictions and density estimates than existing Dirichlet process mixture regression models. We give conditions for weak consistency of the joint distribution and pointwise consistency of the regression estimate.

Keywords: Bayesian nonparametrics, generalized linear models, posterior consistency

1. Introduction

In this paper, we examine the general regression problem. The general regression problem models a response variable Y as dependent on a set of covariates x,

$$Y \,|\, x \sim f(m(x)).$$

The function $m(x)$ is the *mean function*, which maps the covariates to the conditional mean of the response; the distribution f characterizes the deviation of the response from its conditional mean. The simplest example is linear regression, where $m(x)$ is a linear function of x, and f is a Gaussian distribution with mean $m(x)$ and fixed variance.

Generalized linear models (GLMs) extend linear regression to many types of response variables (McCullagh and Nelder, 1989). In their canonical form, a GLM assumes that the conditional mean of the response is a linear function of the covariates, and that the response distribution is in an exponential family. Many classical regression and classification methods are GLMs, including logistic regression, multinomial regression, and Poisson regression.

The GLM framework makes two assumptions about the relationship between the covariates and the response. First, the covariates enter the distribution of the response through a linear function; a non-linear function may be applied to the output of the linear function, but only one that does not depend on the covariates. Second, the variance of the response cannot depend on the covariates. Both these assumptions can be limiting—there are many applications where we would like the response to be a non-linear function of the covariates or where our uncertainty around the response might depend on the covariates. In this paper, we develop a general regression algorithm that relaxes both of these assumptions. Our method captures arbitrarily shaped response functions and heteroscedasticity, that is, the property of the response distribution where both its mean and variance change with the covariates, while still retaining the flexibility of GLMs.

Our idea is to model the mean function $m(x)$ by a mixture of simpler "local" response distributions $f_i(m_i(x))$, each one applicable in a region of the covariates that exhibits similar response patterns. To handle multiple types of responses, each local regression is a GLM. This means that each $m_i(x)$ is a linear function, but a non-linear mean function arises when we marginalize out the uncertainty about which local response distribution is in play. (See Figure 1 for an example with one covariate and a continuous response function.) Furthermore, our method captures heteroscedasticity: the variance of the response function can vary across mixture components and, consequently, varies as a function of the covariates.

Finally, we use a Bayesian nonparametric mixture model to let the data determine both the number and form of the local mean functions. This is critical for modeling arbitrary response distributions: complex response functions can be constructed with many local functions, while simple response functions need only a small number. Unlike frequentist nonparametric regression methods, for example, those that create a mean function for each data point, the Bayesian nonparametric approach uses only as complex a model as the data require. Moreover, it produces a generative model. It can be used to infer properties other than the mean function, such as the conditional variance or response quantiles.

Thus, we develop *Dirichlet process mixtures of generalized linear models* (DP-GLMs), a regression tool that can model many response types and many response shapes. DP-GLMs generalize several existing Bayesian nonparametric regression models (Müller et al., 1996; Shahbaba and Neal, 2009) to a variety of response distributions. We derive Gibbs sampling algorithms for fitting and predicting with DP-GLMs. We investigate some asymptotic properties, including weak consistency of the joint density estimate and consistency of the regression estimate. We study DP-GLMs with several types of data.

The paper is organized as follows. In Section 2, we review the current research on Bayesian nonparametric regression and discuss how the DP-GLM extends this field. In Section 3, we review Dirichlet process mixture models and generalized linear models. In Section 4, we construct the DP-GLM and derive algorithms for posterior computation. In Section 5 we give general conditions for weak consistency of the joint density model and consistency of the regression estimate; we give several models where the conditions hold. In Section 6 we study the DP-GLM and other methods on three data sets; our study illustrates that the DP-GLM provides a powerful nonparametric regression model that can be used in many types of data analysis.

2. Related Work

Existing methods for Bayesian nonparametric regression include Gaussian processes (GP), Bayesian regression trees, and Dirichlet process mixtures.

GP priors assume that the observations arise from a Gaussian process model with known covariance function form (Rasmussen and Williams, 2006). GPs are can model many response types, including continuous, categorical, and count data (Rasmussen and Williams, 2006; Adams et al., 2009). With the proper choice of covariance function, GPs can handle continuous and discrete covariates (Rasmussen and Williams, 2006; Qian et al., 2008). GPs assume that the response exhibits a constant covariance; this assumption is relaxed with Dirichlet process mixtures of GPs (Rasmussen and Ghahramani) or treed GPs (Gramacy and Lee, 2008).

Regression tree models, such as classification and regression trees (CART) (Brieman et al., 1984), are a natural way to handle regression with continuous, categorical or mixed data. They split the data into a fixed, tree-based partitioning and fit a regression model within each leaf of the tree. Bayesian regression trees place a prior over the size of the tree and can be viewed as an automatic bandwidth selection method for CART (Chipman et al., 1998). Bayesian trees have been expanded to include linear models (Chipman et al., 2002) and GPs (Gramacy and Lee, 2008) in the leaf nodes.

The Dirichlet process has been applied to regression problems. West et al. (1994), Escobar and West (1995) and Müller et al. (1996) used joint Gaussian mixtures for continuous covariates and response. Rodriguez et al. (2009) generalized this method using dependent DPs, that is, Dirichlet processes with a Dirichlet process prior on their base measures, in a setting with a response defined as a set of functionals. However, regression by a joint density estimate poses certain challenges. The balance between fitting the response and the covariates, which often outnumber the response, can be slanted toward fitting the covariates at the cost of fitting the response.

To avoid these issues—which amount to over-fitting the covariate distribution and under-fitting the response—some researchers have developed methods that use local weights on the covariates to produce local response DPs. This has been achieved with kernels and basis functions (Griffin and Steel, 2010; Dunson et al., 2007), GPs (Gelfand et al., 2005) and general spatial-based weights (Griffin and Steel, 2006, 2010; Duan et al., 2007). Still other methods, again based on dependent DPs, capture similarities between clusters, covariates or groups of outcomes, including in non-continuous settings (De Iorio et al., 2004; Rodriguez et al., 2009). The method presented here is equally applicable to the continuous response setting and tries to balance its fit of the covariate and response distributions by introducing local GLMs—the clustering structure is based on both the covariates and how the response varies with them.

There is less research about Bayesian nonparametric models for other response types. Mukhopadhyay and Gelfand (1997) and Ibrahim and Kleinman (1998) used a DP prior for the random effects portion of a GLM. Likewise, Amewou-Atisso et al. (2003) used a DP prior to model arbitrary symmetric error distributions in a semi-parametric linear regression model. These methods still maintain the assumption that the covariates enter the model linearly and in the same way. Our work is closest to Shahbaba and Neal (2009). They proposed a model that mixes over both the covariates and response, where the response is drawn from a multinomial logistic model. The DP-GLM is a generalization of their idea.

Asymptotic properties of Dirichlet process mixture models have been studied mostly in the context of density estimation, specifically consistency of the posterior density 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). Recently, the posterior properties of DP regression estimators have been studied. Rodriguez et al. (2009) showed point-wise consistency (asymptotic unbiasedness) for the regression estimate produced by their model assuming continuous covariates under different treatments with a continuous responses and a conjugate base measure (normal-inverse Wishart). In Section 5 we show weak consistency of the joint density estimate produced by the DP-GLM. This is used to show pointwise consistency of the regression estimate in both the continuous and categorical response settings. In the continuous response setting, our results generalize those of Rodriguez et al. (2009) and Rodrıguez (2009). In the categorical response setting, our theory provides results for the classification model of Shahbaba and Neal (2009).

3. Mathematical Background

In this section we provide mathematical background. We review Dirichlet process mixture models and generalized linear models.

3.1 Dirichlet Process Mixture Models

The *Dirichlet process* (DP) is a distribution over distributions (Ferguson, 1973). It is denoted,

$$G \sim \mathrm{DP}(\alpha G_0),$$

where G is a random distribution. There are two parameters. The base distribution G_0 is a distribution over the same space as G. For example, if G is a distribution on reals then G_0 must be a distribution on reals too. The concentration parameter α is a positive scalar. One property of the DP is that random distributions G are discrete, and each places its mass on a countably infinite collection of atoms drawn from G_0.

Consider the model

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha G_0), \\ \theta_i &\sim G. \end{aligned}$$

Marginalizing out the random distribution, the joint distribution of n replicates of θ_i is

$$p(\theta_{1:n} \,|\, \alpha G_0) = \int \left(\prod_{i=1}^{n} G(\theta_i) \right) P(G) dG.$$

This joint distribution has a simpler form. The conditional distribution of θ_n given $\theta_{1:(n-1)}$ follows a Polya urn distribution (Blackwell and MacQueen, 1973),

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

With this conditional distribution, we use the chain rule to specify the joint distribution.

Equation (1) reveals the *clustering property* of the joint distribution of $\theta_{1:n}$: there is a positive probability that each θ_i will take on the value of another θ_j, leading some of the variables to share values. This equation also reveals the roles of scaling parameter α and base distribution $\mathbb{G}_0$. The

unique values contained in $\theta_{1:n}$ are drawn independently from $\mathbb{G}_0$, and 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 $\theta_{1:n}$.

In a DP mixture, θ_i is a latent variable that parameterizes the distribution of an observed data point, point (Antoniak, 1974),

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

Consider the posterior distribution of $\theta_{1:n}$ given $x_{1:n}$. Because of the clustering property, observations group according to their shared parameters. Unlike finite clustering models, however, the number of groups is not assumed known in advance of seeing the data. For this reason, DP mixtures are sometimes called "infinite clustering" models.

3.2 Generalized Linear Models

Generalized linear models (GLMs) build on linear regression to provide a flexible suite of predictive models. GLMs relate a linear model to a response via a link function; examples include familiar models like logistic regression, Poisson regression, and multinomial regression. See McCullagh and Nelder (1989).

GLMs have three components: the conditional probability model of response Y given covariates x, the linear predictor, and the link function. GLMs assume that the response distribution is in the exponential family,

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

Here we give the canonical form of the exponential family, where a, b, and c are known functions specific to the exponential family, ϕ is a scale parameter (sometimes called a dispersion parameter), and η is the canonical parameter. A linear predictor, $X\beta$, is used to determine the canonical parameter through a set of transformations. The mean response is $b'(\eta) = \mu = \mathbb{E}[Y|X]$ (Brown, 1986). However, we can choose a link function g such that $\mu = g^{-1}(X\beta)$, which defines η equal to $X\beta$.

4. Dirichlet Process Mixtures of Generalized Linear Models

We now turn to Dirichlet process mixtures of generalized linear models (DP-GLMs), a 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. We derive theoretical properties of the DP-GLM in Section 5.

4.1 Model 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 covariates, and that these models are connected by associating a set of GLM coefficients with each exponential family mixture component. Let $\theta = (\theta_x, \theta_y)$ be the bundle of parameters over X and $Y\,|\,X$, and let $\mathbb{G}_0$ be a base measure on the space of both. For example, θ_x might be a set of d-dimensional multivariate Gaussian

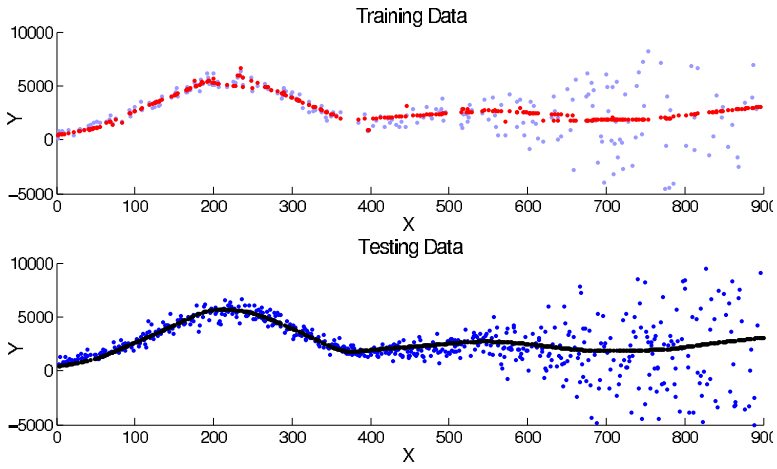

Figure 1: The top figure shows the training data (gray) fitted into clusters, with the prediction given a single sample from the posterior, $\theta^{(i)}$ (red). The bottom figure shows the smoothed regression estimate (black) for the Gaussian model of Equation (2) with the testing data (blue). Data plot multipole moments (X) against power spectrum C_ℓ (Y) for cosmic microwave background radiation (Bennett et al., 2003).

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 GLM dispersion parameter. The full model is

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

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 clusters the covariate-response pairs (x,y). When both are observed, that is, in "training," the posterior distribution of this model will cluster data points according to nearby 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. The DP prior acts as a kernel for the covariates; instead of being a Euclidean metric, the DP measures the distance between two points by the probability that the hidden parameter is shared. See Figure 1 for a demonstration of the DP-GLM.

We now give a few examples of the DP-GLM that will be used throughout this paper.

4.1.1 EXAMPLE: GAUSSIAN MODEL

We now give an example of the DP-GLM for continuous covariates/response that will be used throughout the rest of the paper. For continuous covariates/response in $\mathbb{R}$, we model locally with a Gaussian distribution for the covariates and a linear regression model for the response. The covariates have mean $\mu_{i,j}$ and variance $\sigma^2_{i,j}$ for the j^{th} dimension of the i^{th} observation; the covariance matrix is diagonal in this example. The GLM parameters are the linear predictor $\beta_{i,0},\ldots,\beta_{i,d}$ and the response variance $\sigma^2_{i,y}$. Here, $\theta_{x,i} = (\mu_{i,1:d}, \sigma_{i,1:d})$ and $\theta_{y,i} = (\beta_{i,0:d}, \sigma_{i,y})$. This produces a mixture of multivariate Gaussians. The full model is,

$$\begin{aligned} P &\sim DP(\alpha\mathbb{G}_0), \\ \theta_i|P &\sim P, \\ X_{i,j}|\theta_{i,x} &\sim N\left(\mu_{ij}, \sigma^2_{ij}\right), \qquad j = 1,\ldots,d, \\ Y_i|X_i,\theta_{i,y} &\sim N\left(\beta_{i0} + \sum_{j=1}^{d} \beta_{ij}X_{ij}, \sigma^2_{iy}\right). \end{aligned} \tag{2}$$

This model has been proposed by West et al. (1994), Escobar and West (1995) and Müller et al. (1996). However, they use a fully populated covariance matrix that gives *de facto* β parameters. This is computationally expensive for larger problems and adds posterior likelihood associated with the covariates, rather than the response. A discussion of the problems associated with the latter issue is given in Section 4.4.

4.1.2 EXAMPLE: MULTINOMIAL MODEL (SHAHBABA AND NEAL, 2009)

This model was proposed by Shahbaba and Neal (2009) for nonlinear classification, using a Gaussian mixture to model continuous covariates and a multinomial logistic model for a categorical response with K categories. The covariates have mean $\mu_{i,j}$ and variance $\sigma^2_{i,j}$ for the j^{th} dimension of the i^{th} observation; the covariance matrix is diagonal for simplicity. The GLM parameters are the K linear predictor $\beta_{i,0,k},\ldots,\beta_{i,d,k}$, $k = 1,\ldots,K$. The full model is,

$$\begin{aligned} P &\sim DP(\alpha\mathbb{G}_0), \\ \theta_i|P &\sim P, \\ X_{i,j}|\theta_{i,x} &\sim N\left(\mu_{ij}, \sigma^2_{ij}\right), \qquad j = 1,\ldots,d, \\ \mathbb{P}(Y_i = k|X_i,\theta_{i,y}) &= \frac{\exp\left(\beta_{i,0,k} + \sum_{j=1}^{d}\beta_{i,j,k}X_{i,j}\right)}{\sum_{\ell=1}^{K}\exp\left(\beta_{i,0,\ell} + \sum_{j=1}^{d}\beta_{i,j,\ell}X_{i,j}\right)}, \qquad k = 1,\ldots,K. \end{aligned} \tag{3}$$

4.1.3 EXAMPLE: POISSON MODEL WITH CATEGORICAL COVARIATES

We model the categorical covariates by a mixture of multinomial distributions and the count response by a Poisson distribution. If covariate j has K categories, let $(p_{i,j,1},\ldots,p_{i,j,K})$ be the probabilities for categories $1,\ldots,K$. The covariates are then coded by indicator variables, $\mathbf{1}_{\{X_{i,j}=k\}}$, which

are used with the linear predictor, $\beta_i, 0, \beta_{i,1,1:K}, \ldots, \beta_{i,d,1:K}$. The full model is,

$$\begin{aligned} P &\sim DP(\alpha \mathbb{G}_0), \\ \theta_i | P &\sim P, \\ \mathbb{P}(X_{i,j} = k | \theta_{i,x}) &= p_{i,j,k}, \qquad j = 1, \ldots, d,\ k = 1, \ldots, K, \\ \lambda_i | X_i, \theta_{i,y} &= \exp\left(\beta_{i,0} + \sum_{j=1}^{d} \sum_{k=1}^{K} \beta_{i,j,k} \mathbf{1}_{\{X_{i,j}=k\}} \right), \\ \mathbb{P}(Y_i = k | X_i, \theta_{i,y}) &= \frac{e^{-\lambda_i} \lambda_i^k}{\ell!}, \qquad k = 0, 1, 2, \ldots. \end{aligned} \tag{4}$$

We apply Model (4) to data in Section 6.

4.2 Heteroscedasticity and Overdispersion

One advantage of the DP-GLM is that it provides a strategy for handling common problems in predictive modeling. Many models, such as GLMs and Gaussian processes, make assumptions about data dispersion and homoscedasticity. Overdispersion occurs in single parameter GLMs when the data variance is larger than the variance predicted by the model mean. Mukhopadhyay and Gelfand (1997) have successfully used DP mixtures over GLM intercept parameters to create classes of models that include overdispersion. The DP-GLM retains this property, but is not limited to linearity in the covariates.

A model is *homoscedastic* when the response variance is across constant all covariates; a model is *heteroscedastic* when the response variance changes with the covariates. Models like GLMs are homoscedastic and give poor fits when that assumption is violated in the data. In contrast, the DP-GLM captures heteroscedasticity when mixtures of GLMs are used. The mixture model setting allows variance to be modeled by a separate parameter in each cluster or by a collection of clusters in a single covariate location. This leads to smoothly transitioning heteroscedastic posterior response distributions.

This property is shown in Figure 2, where we compare a DP-GLM to a homoscedastic model (Gaussian processes) and heteroscedastic modifications of homoscedastic models (treed Gaussian processes and treed linear models). The DP-GLM is robust to heteroscedastic data—it provides a smooth mean function estimate, while the other models are not as robust or provide non-smooth estimates.

4.3 Posterior Prediction With a DP-GLM

The DP-GLM is used in prediction problems. Given a collection of covariate-response pairs $D = (X_i, Y_i)_{i=1}^n$, we estimate the joint distribution of $(X, Y) \,|\, D$. For a new set of covariates x, we use the joint to compute the conditional distribution, $Y \,|\, x, D$ and the conditional expectation, $\mathbb{E}[Y \,|\, x, D]$. We give the step-by-step process for formulating specific DP-GLM models and computing the conditional distribution of the response.

4.3.1 Choosing the Mixture Component and GLM

We begin by choosing f_x and the GLM. The Dirichlet process mixture model and GLM provide flexibility in both the covariates and the response. Dirichlet process mixture models allow many

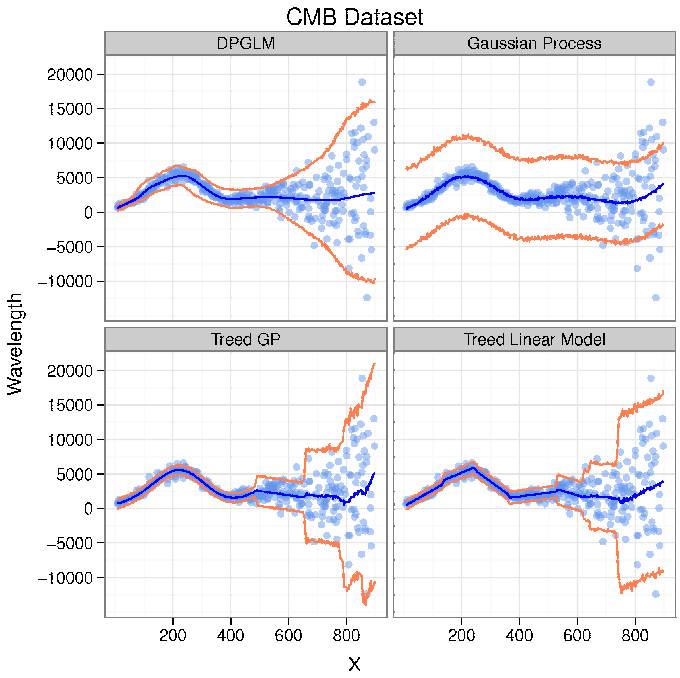

Figure 2: Modeling heteroscedasticity with the DP-GLM and other Bayesian nonparametric methods. The estimated mean function is given along with a 90% predicted confidence interval for the estimated underlying distribution. DP-GLM produces a smooth mean function and confidence interval.

types of variables to be modeled by the covariate mixture and subsequently transformed for use as a covariate in the GLM. Note that certain mixture distributions support certain types of covariates but may not necessarily be a good fit. The same care that goes into choosing distributions and GLMs in a parametric setting is required here.

4.3.2 CHOOSING THE BASE MEASURE AND OTHER HYPERPARAMETERS

The choice of the base measure $\mathbb{G}_0$ affects how expressive the DP-GLM is, the computational efficiency of the prediction and whether some theoretical properties, such as asymptotic unbiasedness, hold. For example, $\mathbb{G}_0$ for the Gaussian model is a distribution over $(\mu_i, \sigma_i, \beta_{i,0:d}, \sigma_{i,y})$. A conjugate base measure is normal-inverse-gamma for each covariate dimension and multivariate normal inverse-gamma for the response parameters. This $\mathbb{G}_0$ allows all continuous, integrable distributions to be supported, retains theoretical properties, such as asymptotic unbiasedness, and yields efficient posterior approximation by collapsed Gibbs sampling (Neal, 2000). In summary, the base measure is chosen in line with data size, distribution type, distribution features (such as heterogeneity, and others) and computational constraints.

Hyperparameters for the DP-GLM include the DP scaling parameter α and hyperparameters parameters for the base measure $\mathbb{G}_0$. We can place a gamma prior on α (Escobar and West, 1995); the parameters of $\mathbb{G}_0$ may also have a prior. Each level of prior reduces the influence of the hyperparameters, but adds computational complexity to posterior inference (Escobar and West, 1995).

4.3.3 APPROXIMATING THE POSTERIOR AND FORMING PREDICTIONS

We derive all quantities of interest—that is, conditional distributions and expectations—from the posterior of the joint distribution of (x,y). Define $f(x,y\,|\,D)$ as the joint posterior distribution given data D and $f(x,y\,|\,\theta_{1:n})$ as the joint distribution given parameters $\theta_{1:n}$ that are associated with data $D=(X_i,Y_i)_{i=1}^n$. The posterior can be expressed through a conditional expectation,

$$f(x,y\,|\,D)=\mathbb{E}\left[f(x,y\,|\,\theta_{1:n})\,|\,D\right]. \tag{5}$$

While the true posterior distribution, $f(x,y\,|\,D)$, may be impossible to compute, the joint distribution conditioned on $\theta_{1:n}$ has the form

$$f(x,y\,|\,\theta_{1:n})=\frac{\alpha}{\alpha+n}\int_{\mathcal{T}} f_y(y|x,\theta)f_x(x|\theta)\mathbb{G}_0(d\theta)+\frac{1}{\alpha+n}\sum_{i=1}^{n} f_y(y|x,\theta_i)f_x(x|\theta_i).$$

We approximate the expectation in Equation (5) by Monte Carlo integration using M posterior samples of $\theta_{1:n}$,

$$f(x,y\,|\,D)\approx\frac{1}{M}\sum_{m=1}^{M} f(x,y\,|\,\theta_{1:n}^{(m)}).$$

We use Markov chain Monte Carlo (MCMC), specifically Gibbs sampling, to obtain M i.i.d. samples from this distribution. (See Escobar, 1994, MacEachern, 1994, Escobar and West, 1995 and MacEachern and Müller, 1998 for foundational work; Neal, 2000 provides a review and state of the art algorithms.) We construct a Markov chain on the hidden variables $\theta_{1:n}$ such that its limiting distribution is the posterior. We give implementation details in Appendix A.

We use a similar strategy to construct the conditional distribution of $Y\,|\,X=x,D$. The conditional distribution is

$$f(Y\,|\,X=x,D)=\frac{f(Y,x\,|\,D)}{\int f(y,x\,|\,D)dy}.$$

Again using M i.i.d. samples from the posterior of $\theta_{1:n}\,|\,D$,

$$\begin{aligned} f(Y\,|\,X=x,D)&\approx\frac{1}{M}\sum_{m=1}^{M} f(Y\,|\,X=x,\theta_{1:n}^{(m)}),\\ &=\frac{1}{M}\sum_{m=1}^{M}\frac{\alpha\int_{\mathcal{T}} f_y(Y|X=x,\theta)f_x(x|\theta)\mathbb{G}_0(d\theta)+\sum_{i=1}^{n} f_y(Y|X=x,\theta_i^{(m)})f_x(x|\theta_i^{(m)})}{\alpha\int_{\mathcal{T}} f_x(x|\theta)\mathbb{G}_0(d\theta)+\sum_{i=1}^{n} f_x(x|\theta_i^{(m)})}. \end{aligned}$$

We use the same methodology to compute the conditional expectation of the response given a new set of covariates x and the observed data D, $\mathbb{E}[Y\,|\,X=x,D]$. Again using iterated expectation, we condition on the latent variables,

$$\mathbb{E}\left[Y\,|\,X=x,D\right]=\mathbb{E}\left[\mathbb{E}\left[Y\,|\,X=x,\theta_{1:n}\right]\,|\,D\right]. \tag{6}$$

Conditional on the latent parameters $\theta_{1:n}$ that generated the observed data, the inner expectation is

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

Since we assume Y is a GLM, $\mathbb{E}\left[Y|X=x,\theta\right]$ is available in closed form as a function of x and θ.

The outer expectation of Equation (6) is usually intractable. We approximate it by Monte Carlo integration with M posterior samples of $\theta_{1:n}$,

$$\mathbb{E}\left[Y \mid X = x, D\right] \approx \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\left[Y \mid X = x, \theta_{1:n}^{(m)}\right].$$

4.4 Comparison to the Dirichlet Process Mixture Model Regression

The DP-GLM models the response Y conditioned on the covariates X. An alternative is one where we model (X,Y) from a common mixture component in a classical DP mixture (see Section 3), and then form the conditional distribution of the response from this joint. We investigate the mathematical differences between these approaches and the consequences of those differences. (They are compared empirically in Section 6.)

A Dirichlet process mixture model (DPMM) has the form,

$$\begin{aligned} P &\sim DP(\alpha \mathbb{G}_0), \\ \theta_i | P &\sim P, \\ X_i | \theta_{i,x} &\sim f_x(x|\theta_{i,x}), \\ Y_i | \theta_{i,y} &\sim f_y(y|\theta_{i,y}). \end{aligned} \tag{7}$$

This model has been studied in Escobar and West (1995) where (X_i, Y_i) are assumed to have a joint Gaussian distribution. When the covariance matrix is assumed to be diagonal, the regression estimate is generally poor. However, when the covariance matrix is assumed to be fully populated, computation becomes difficult with more than a few covariate dimensions. We focus on the case with diagonal covariance. We study why it performs poorly and how the DP-GLM improves on it with minimal increase in computational difficulty. The difference between Model (7) and the DP-GLM is that the distribution of Y given θ is conditionally independent of the covariates X. This difference has consequences on the posterior distribution and, thus, the posterior predictions.

One consequence is that the GLM response component acts to remove boundary bias for samples near the boundary of the covariates in the training data set. The GLM fits a linear predictor through the training data; all predictions for boundary and out-of-sample covariates follow the local predictors. The traditional DP model, however, only fits a local mean; all boundary and out-of-sample predictions center around that mean. The boundary effects are compared in Figure 3. The DP-GLM can be viewed as a Bayesian analogy of a locally linear kernel estimator while the regular DP is similar to the Nadaraya-Watson kernel estimator (Nadaraya, 1964; Watson, 1964).

Another consequence is that the proportion of the posterior likelihood devoted to the response differs between the two methods. Consider the log of the posterior of the DPMM given in Model (7). Assume that f_y is a single parameter exponential, where $\theta_y = \beta$,

$$\ell(\theta^{dp} \mid D) \propto \sum_{i=1}^{K} \left[\ell(\beta_{C_i}) + \sum_{c \in C_i} \ell(y_c \mid \beta_{C_i}) + \sum_{j=1}^{d} \ell(\theta_{C_i,x_j} \mid D) \right]. \tag{8}$$

Here, ℓ denotes log likelihood and "$\propto$" means "proportional in the log space." The log of the DP-GLM posterior for a single parameter exponential family GLM, where $\theta_y = (\beta_0, \ldots, \beta_d)$, has the form,

$$\ell(\theta^{dpglm} \mid D) \propto \sum_{i=1}^{K} \left[\sum_{j=0}^{d} \ell(\beta_{C_i,j}) + \sum_{c \in C_i} \ell(y_c \mid \beta_{C_i}^T x_c) + \sum_{j=1}^{d} \ell(\theta_{C_i,x_j} \mid D) \right]. \tag{9}$$

As the number of covariates grows, the likelihood associated with the covariates grows in both equations. However, the likelihood associated with the response also grows with the extra response parameters in Equation (9), whereas it is fixed in Equation (8).

These posterior differences lead to two predictive differences. First, the DP-GLM is much more resistant to dimensionality than the DPMM. Since the number of response related parameters grows with the number of covariate dimensions in the DP-GLM, the relative posterior weight of the response does not shrink as quickly in the DP-GLM as it does in the DPMM. This keeps the response variable important in the selection of the mixture components and makes the DP-GLM a better predictor than the DPMM as the number of dimensions grows.

As the dimensionality grows, however, the DP-GLM produces less stable predictions than the DPMM. While the additional GLM parameters help maintain the relevance of the response, they also add noise to the prediction. This is seen in Figure 3. The GLM parameters in this figure have a Gaussian base measure, effectively creating a local ridge regression.[1] In lower dimensions, the DP-GLM produced more stable results than the DPMM because a smaller number of larger clusters were required to fit the data well. The DPMM, however, consistently produced stable results in higher dimensions as the response became more of a sample average than a local average. The DPMM has the potential to predict well if changes in the mean function coincide with underlying local modes of the covariate density. However, the DP-GLM forces the covariates into clusters that coincide more with the response variable due to the inclusion of the slope parameters.

We now discuss the theoretical properties of the DP-GLM.

5. Asymptotic Properties of the DP-GLM Model

In this section, we study the asymptotic properties of the DP-GLM model, namely weak consistency of the joint density estimate and pointwise consistency (asymptotic unbiasedness) of the regression estimate. Consistency is the notion that posterior distribution accumulates in regions close to the true distribution. Weak consistency assures that the posterior distribution accumulates in regions of densities where "properly behaved" functions (i.e., bounded and continuous) integrated with respect to the densities in the region are arbitrarily close to the integral with respect to the true density. We then use the weak consistency results to give conditions for asymptotic unbiasedness of the regression estimate. Both consistency and asymptotic unbiasedness act as frequentist justification of Bayesian methods; more observations lead to models that tend toward the "correct" value. Neither weak consistency nor asymptotic unbiasedness are guaranteed for Dirichlet process mixture models.

Notation for this section is more complicated than the notation for the model. Let $f_0(x,y)$ be the true joint distribution of (x,y); in this case, we will assume that f_0 is a density. Let $\mathcal{F}$ be the set of all density functions over (x,y). Let Π^f be the prior over $\mathcal{F}$ induced by the DP-GLM model. Let $\mathbb{E}_{f_0}[\cdot]$ denote the expectation under the true distribution and $\mathbb{E}_{\Pi^f}[\cdot]$ be the expectation under the prior Π^f.

In general, an estimator is a function of observations. Assuming a true distribution of those observations, an estimator is called unbiased if its expectation under that distribution is equal to the value that it estimates. If an estimator has this property, it is called consistent. In the case of

1. In unpublished results, we tried other base measures, such as a Laplacian distribution. They produced less stable results than the Gaussian base measure.

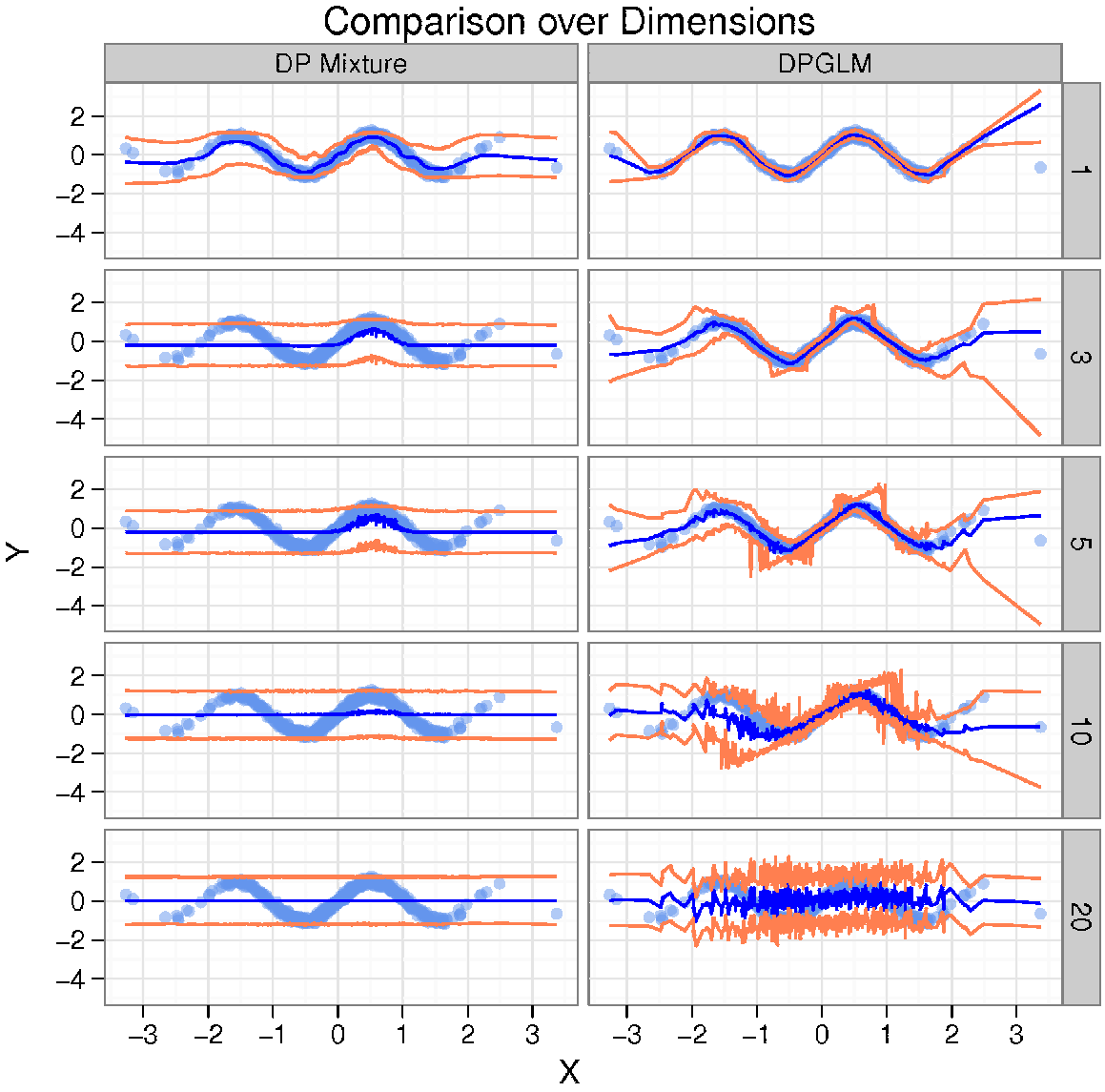

Figure 3: A plain Dirichlet process mixture model regression (left) versus DP-GLM, plotted against the number of spurious dimensions (vertical plots). We give the estimated mean function along with a 90% predicted confidence interval for the estimated underlying distribution. Data have one predictive covariate and a varying number of spurious covariates. The covariate data were generated by a mixture model. DP-GLM produces a smoother mean function and is much more resistant to spurious dimensionality.

DP-GLM, that would mean for every x in a fixed domain $\mathcal{A}$ and every $n > 0$,

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

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

$$\mathbb{E}_{\Pi^f}[Y|x,(X_i,Y_i)_{i=1}^n] \to \mathbb{E}_{f_0}[Y|x] \quad \text{as } n \to \infty.$$

5.1 Weak Consistency of the Joint Posterior Distribution

Weak consistency is the idea that the posterior distribution, $\Pi^f(f \,|\, (X_i,Y_i)_{i=1}^n)$ collects in weak neighborhoods of the true distribution, $f_0(x,y)$. A weak neighborhood of f_0 of radius ε, $\mathcal{W}_\varepsilon(f_0)$, is defined as follows,

$$\mathcal{W}_\varepsilon(f_0) = \left\{ f : \left| \int f_0(x,y)g(x,y)dxdy - \int f(x,y)g(x,y)dxdy \right| < \varepsilon \right\}$$

for every bounded, continuous function g. Aside from guaranteeing that the posterior collects in regions close to the true distribution, weak consistency can be used to show consistency of the regression estimate under certain conditions. We give conditions for weak consistency for joint posterior distribution of the Gaussian and multinomial models and use these results to show consistency of the regression estimate for these same models.

We now give a theorem for the asymptotic unbiasedness of the Gaussian model.

Theorem 1 *Let Π^f be the prior induced by the Gaussian model of Equation (2). If $f_0(x,y)$ has compact support, is absolutely continuous over that domain and $\mathbb{G}_0$ has support $\mathbb{R}^d \times \mathbb{R}_+^d \times \mathbb{R}^{d+1} \times \mathbb{R}_+$, then*

$$\Pi^f(\mathcal{W}_\varepsilon(f_0) \,|\, (X_i,Y_i)_{i=1}^n) \to 1$$

as $n \to \infty$ for every $\varepsilon > 0$.

Posterior consistency of similar models, namely Dirichlet process mixtures of Gaussians, has been extensively studied by Ghosal et al. (1999), Ghosh and Ramamoorthi (2003), and Tokdar (2006) and convergence rates in Walker et al. (2007). The compact support condition for f_0 allows for broad array of base measures to produce weakly consistent posteriors. See Tokdar (2006) for results on non-compactly supported f_0.

We now give an analogous theorem for the multinomial model.

Theorem 2 *Let Π^f be the prior induced by the multinomial model of Equation (3). If $f_0(x)$ has compact support, is absolutely continuous, $\mathbb{G}_0$ has support $\mathbb{R}^d \times \mathbb{R}_+^d \times \mathbb{R}^{d+1}$, and $\mathbb{P}_{f_0}[Y = k \,|\, X = x]$ is absolutely continuous in x for $k = 1,\ldots,K$, then*

$$\Pi^f(\mathcal{W}_\varepsilon(f_0) \,|\, (X_i,Y_i)_{i=1}^n) \to 1$$

as $n \to \infty$ for every $\varepsilon > 0$.

The proofs of Theorems 1 and 2 are given in the Appendix.

5.2 Consistency of the Regression Estimate

We approach consistency of the regression estimate by using weak consistency for the posterior of the *joint* distribution and then placing additional integrability constraints on the base measure $\mathbb{G}_0$. We now give results for the Gaussian and multinomial models.

Theorem 3 *Let Π^f be the prior induced by the Gaussian model of Equation (2). If*

(i) $\mathbb{G}_0$ and f_0 satisfy the conditions of Theorem 1, and

(ii) $\int(\beta_0+\sum_{i=1}^d \beta_i x_i)\mathbb{G}_0(d\beta) < \infty$ for every $x \in C$,

then

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

almost surely $\mathbb{P}_{f_0}^\infty$.

Similarly, we give a theorem for the multinomial model.

Theorem 4 *Let Π^f be the prior induced by the multinomial model of Equation (3). If*

(i) $\mathbb{G}_0$ and f_0 satisfy the conditions of Theorem 2, and

(ii) $\mathbb{P}_{f_0}[Y=k\,|\,X=x]$ is continuous in x for $k=1,\ldots,K$,

then

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

almost surely $\mathbb{P}_{f_0}^\infty$ for $k=1,\ldots,K$.

See Appendix B for proofs of Theorems 3 and 4.

5.3 Consistency Example: Gaussian Model

Examples of prior distributions that satisfy Theorems 1 and 3 are as follows.

5.3.1 NORMAL-INVERSE-WISHART

Note that in the Gaussian case, slope parameters can be generated by a full covariance matrix: using a conjugate prior, a Normal-Inverse-Wishart, will produce an instance of the DP-GLM. Define the following model, which was used by Müller et al. (1996),

$$\begin{aligned} P &\sim DP(\alpha\mathbb{G}_0), \\ \theta_i\,|\,P &\sim P, \\ (X_i,Y_i)\,|\,\theta_i &\sim N(\mu,\Sigma). \end{aligned} \tag{10}$$

The last line of Model (10) can be broken down in the following manner,

$$\begin{aligned} X_i\,|\,\theta_i &\sim N\left(\mu_x,\Sigma_x\right), \\ Y_i\,|\,\theta_i &\sim N\left(\mu_y + b^T\Sigma_x^{-1}b(X_i-\mu_x), \sigma_y^2 - b^T\Sigma_x^{-1}b\right), \end{aligned}$$

where

$$\mu = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_y^2 & b^T \\ b & \Sigma_x \end{bmatrix}.$$

We can then define β as,

$$\beta_0 = \mu_y - b^T \Sigma_x^{-1} \mu_x, \qquad \beta_{1:d} = b^T \Sigma_x^{-1}.$$

The base measure $\mathbb{G}_0$ is defined as,

$$(\mu, \Sigma) \sim \textit{Normal Inverse Wishart}(\lambda, \nu, a, B).$$

Here λ is a mean vector, ν is a scaling parameter for the mean, a is a scaling parameter for the covariance, and B is a covariance matrix.

5.3.2 DIAGONAL NORMAL-INVERSE-GAMMA

It is often more computationally efficient to specify that Σ_x is a diagonal matrix. In this case, we can specify a conjugate base measure component by component:

$$\begin{aligned}
\sigma_{i,j} &\sim \textit{Inverse Gamma}(a_j, b_j), && j = 1, \ldots, d, \\
\mu_{i,j} \,|\, \sigma_{i,j} &\sim N(\lambda_j, \sigma_{i,j}/\nu_j), && j = 1, \ldots, d, \\
\sigma_{i,y} &\sim \textit{Inverse Gamma}(a_y, b_y), \\
\beta_{i,j} \,|\, \sigma_{i,y} &\sim N_{d+1}(\lambda_y, \sigma_y/\nu_y).
\end{aligned}$$

The Gibbs sampler can still be collapsed, but the computational cost is much lower than the full Normal-Inverse-Wishart.

5.3.3 NORMAL MEAN, LOG NORMAL VARIANCE

Conjugate base measures tie the mean to the variance and can be a poor fit for small, heteroscedastic data sets. The following base measure was proposed by Shahbaba and Neal (2009),

$$\begin{aligned}
\log(\sigma_{i,j}) &\sim N(m_{j,\sigma}, s^2_{j,\sigma}), && j = y, 1, \ldots, d, \\
\mu_{i,j} &\sim N(m_{j,\mu}, s^2_{j,\mu}), && j = 1, \ldots, d, \\
\beta_{i,j} &\sim N(m_{j,\beta}, s^2_{j,\beta}) && j = 0, \ldots, d.
\end{aligned}$$

5.4 Consistency Example: Multinomial Model

Now consider the multinomial model of Shahbaba and Neal (2009), given in Model (3),

$$\begin{aligned}
P &\sim DP(\alpha \mathbb{G}_0), \\
\theta_i | P &\sim P, \\
X_{i,j} | \theta_{i,x} &\sim N\left(\mu_{ij}, \sigma_{ij}^2\right), && j = 1, \ldots, d, \\
\mathbb{P}(Y_i = k | X_i, \theta_{i,y}) &= \frac{\exp\left(\beta_{i,0,k} + \sum_{j=1}^d \beta_{i,j,k} X_{i,j}\right)}{\sum_{\ell=1}^K \exp\left(\beta_{i,0,\ell} + \sum_{j=1}^d \beta_{i,j,\ell} X_{i,j}\right)}, && k = 1, \ldots, K.
\end{aligned}$$

Examples of prior distributions that satisfy Theorems 2 and 4 are as follows.

5.4.1 NORMAL-INVERSE-WISHART

The covariates have a Normal-Inverse-Wishart base measure while the GLM parameters have a Gaussian base measure,

$$\begin{aligned}(\mu_{i,x}, \Sigma_{i,x}) &\sim \textit{Normal Inverse Wishart}(\lambda, \nu, a, B),\\ \beta_{i,j,k} &\sim N(m_{j,k}, s^2_{j,k}), \qquad\qquad j = 0,\ldots,d,\ \ k = 1,\ldots,K.\end{aligned}$$

5.4.2 DIAGONAL NORMAL-INVERSE-GAMMA

It is often more computationally efficient to specify that Σ_x is a diagonal matrix. Again, we can specify a conjugate base measure component by component while keeping the Gaussian base measure on the GLM components,

$$\begin{aligned}\sigma_{i,j} &\sim \textit{Inverse Gamma}(a_j, b_j), && j = 1,\ldots,d,\\ \mu_{i,j}\,|\,\sigma_{i,j} &\sim N(\lambda_j, \sigma_{i,j}/\nu_j), && j = 1,\ldots,d,\\ \beta_{i,j,k}\,|\,\sigma_{i,y} &\sim N(m_{j,k}, s^2_{j,k}), && j = 0,\ldots,d,\ \ k = 1,\ldots,K.\end{aligned}$$

5.4.3 NORMAL MEAN, LOG NORMAL VARIANCE

Likewise, for heteroscedastic covariates we can use the log normal base measure of Shahbaba and Neal (2009),

$$\begin{aligned}\log(\sigma_{i,j}) &\sim N(m_{j,\sigma}, s^2_{j,\sigma}), && j = 1,\ldots,d,\\ \mu_{i,j} &\sim N(m_{j,\mu}, s^2_{j,\mu}), && j = 1,\ldots,d,\\ \beta_{i,j,k} &\sim N(m_{j,k,\beta}, s^2_{j,k,\beta}) && j = 0,\ldots,d,\ \ k = 1,\ldots,K.\end{aligned}$$

6. Empirical Study

We compare the performance of DP-GLM regression to other regression methods. We studied data sets that illustrate the strengths of the DP-GLM, including robustness with respect to data type, heteroscedasticity and higher dimensionality than can be approached with traditional methods. Shahbaba and Neal (2009) used a similar model on data with categorical covariates and count responses; their numerical results were encouraging. We tested the DP-GLM on the following data sets.

6.1 Data Sets

We selected three data sets with continuous response variables. They highlight various data difficulties within regression, such as error heteroscedasticity, moderate dimensionality (10–12 covariates), various input types and response types.

- **Cosmic Microwave Background (CMB) (Bennett et al., 2003).** The data set consists of 899 observations which map positive integers $\ell = 1, 2, \ldots, 899$, 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 binary and 4 have 3 to 6 classes for a total of 22 categories. 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.

Data set testing sizes ranged from very small (20 observations) to moderate sized (800 observations). Small data set sizes were included due to interests in (future) online applications.

6.2 Competitors

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

- **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 $[1\, X_1 \ldots X_d]^T$ as basis functions. OLS can be modified to accommodate both continuous and categorical inputs, but it requires a continuous response function.

- **CART.** A nonparametric tree regression method (Brieman et al., 1984) generated by the Matlab function classregtree. It accommodates both continuous and categorical inputs and any type of response.

- **Bayesian CART.** A tree regression model with a prior over tree size (Chipman et al., 1998); it was implemented in R with the tgp package.

- **Bayesian Treed Linear Model.** A tree regression model with a prior over tree size and a linear model in each of the leaves (Chipman et al., 2002); it was implemented in R with the tgp package.

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

- **Treed Gaussian Processes.** A tree regression model with a prior over tree size and a GP on each leaf node (Gramacy and Lee, 2008); it was implemented in R with the tgp package.

Method	Mean Absolute Error					Mean Square Error				
Training set size	30	50	100	250	500	30	50	100	250	500
DP-GLM	0.58	0.51	0.49	**0.48**	**0.45**	**1.00**	**0.94**	**0.91**	**0.94**	**0.83**
Linear Regression	0.66	0.65	0.63	0.65	0.63	1.08	1.04	1.01	1.04	0.96
CART	0.62	0.60	0.60	0.56	0.56	1.45	1.34	1.43	1.29	1.41
Bayesian CART	0.66	0.64	0.54	0.50	0.47	1.04	1.01	0.93	**0.94**	0.84
Treed Linear Model	0.64	0.52	0.49	**0.48**	0.46	1.10	0.95	0.93	0.95	0.85
Gaussian Process	0.55	0.53	0.50	0.51	0.47	1.06	0.97	0.93	0.96	0.85
Treed GP	**0.52**	**0.49**	**0.48**	**0.48**	0.46	1.03	0.95	0.95	0.96	0.89

Table 1: Mean absolute and square errors for methods on the CMB data set by training data size. The best results for each size of training data are in bold.

- **Basic DP Regression.** Similar to DP-GLM, except the response is a function only of μ_y, rather than $\beta_0 + \sum \beta_i x_i$. For the Gaussian model,

$$\begin{aligned} P &\sim DP(\alpha \mathbb{G}_0), \\ \theta_i | P &\sim P, \\ X_i | \theta_i &\sim N(\mu_{i,x}, \sigma^2_{i,x}), \\ Y_i | \theta_i &\sim N(\mu_{i,y}, \sigma^2_{i,y}). \end{aligned}$$

  This model was explored in Section 4.4.

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

6.3 Cosmic Microwave Background (CMB) Results

For this data set, we used a Gaussian model with base measure

$$\begin{aligned} \mu_x &\sim N(m_x, s^2_x), & \sigma^2_x &\sim \exp\left\{N(m_{x,s}, s^2_{x,s})\right\}, \\ \beta_{0:d} &\sim N(m_{y,0:d}, s^2_{y,0:d}), & \sigma^2_y &\sim \exp\left\{N(m_{x,s}, s^2_{x,s})\right\}. \end{aligned}$$

This prior was chosen because the variance tails are heavier than an inverse gamma and the mean is not tied to the variance. It is a good choice for heterogeneous data because of those features. Computational details are given in Appendix C.

All non-linear methods except for CART (DP-GLM, Bayesian CART, treed linear models, GPs and treed GPs) did comparably on this data set; CART had difficulty finding an appropriate bandwidth. Linear regression did poorly due to the non-linearity of the data set. Fits for heteroscedasticity for the DP-GLM, GPs, treed GPs and treed linear models on 250 training data points can be seen in Figure 2. See Figure 4 and Table 1 for results.

6.4 Concrete Compressive Strength (CCS) Results

The CCS data set was chosen because of its moderately high dimensionality and continuous covariates and response. For this data set, we used a Gaussian model and a conjugate base measure with

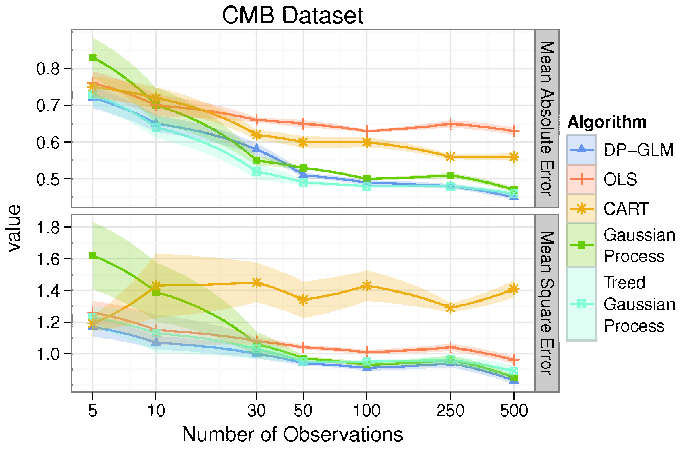

Figure 4: The average mean absolute error (top) and mean squared error (bottom) for ordinary least squares (OLS), tree regression, Gaussian processes and DP-GLM on the CMB data set. The data were normalized. Mean $+/-$ one standard deviation are given for each method.

conditionally independent covariate and response parameters,

$$(\mu_x, \sigma_x^2) \sim Normal - Inverse - Gamma(m_x, s_x, a_x, b_x),$$
$$(\beta_{0:d}, \sigma_y^2) \sim Multivariate\ Normal - Inverse - Gamma(M_y, S_y, a_y, b_y).$$

This base measure allows the sampler to be fully collapsed but has fewer covariate-associated parameters than a full Normal-Inverse-Wishart base measure, giving it a better fit in a moderate dimensional setting. In testing, it also provided better results for this data set than the exponentiated Normal base measure used for the CMB data set; this is likely due to the low noise and variance of the CCS data set. Computational details are given in Appendix C.

Results on this data set were more varied than those for the CMB data set. GPs had the best performance overall; on smaller sets of training data, the DP-GLM outperformed frequentist CART. Linear regression, basic DP regression and Bayesian CART all performed comparatively poorly. Treed linear models and treed GPs performed very well most of the time, but had convergence problems leading to overall higher levels of predictive error. Convergence issues were likely caused by the moderate dimensionality (8 covariates) of the data set. See Figure 5 and Table 2 for results.

6.5 Solar Flare Results

The Solar data set was chosen to demonstrate the flexibility of DP-GLM. Many regression techniques cannot accommodate categorical covariates and most cannot accommodate a count-type re-

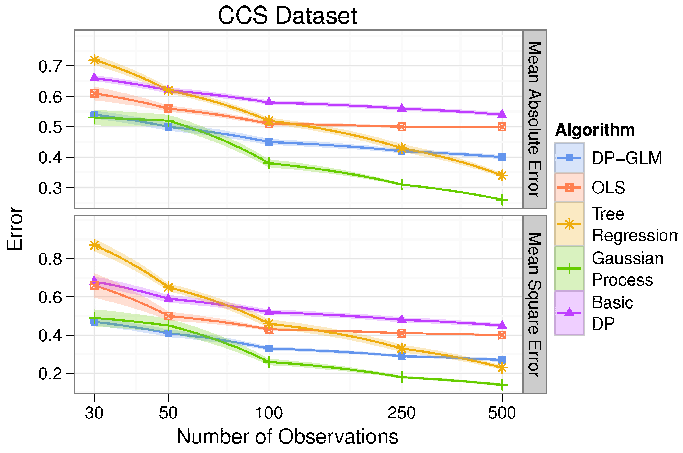

Figure 5: The average mean absolute error (top) and mean squared error (bottom) for ordinary least squares (OLS), tree regression, Gaussian processes, location/scale DP and the DP-GLM Poisson model on the CCS data set. The data were normalized. Mean $+/-$ one standard deviation are given for each method.

Method	Mean Absolute Error					Mean Squared Error				
	30	50	100	250	500	30	50	100	250	500
DP-GLM	0.54	0.50	0.45	0.42	0.40	**0.47**	0.41	0.33	0.28	0.27
Location/Scale DP	0.66	0.62	0.58	0.56	0.54	0.68	0.59	0.52	0.48	0.45
Linear Regression	0.61	0.56	0.51	0.50	0.50	0.66	0.50	0.43	0.41	0.40
CART	0.72	0.62	0.52	0.43	0.34	0.87	0.65	0.46	0.33	0.23
Bayesian CART	0.78	0.72	0.63	0.55	0.54	0.95	0.80	0.61	0.49	0.46
Treed Linear Model	1.08	0.95	0.60	0.35	1.10	7.85	9.56	4.28	0.26	1232
Gaussian Process	**0.53**	0.52	**0.38**	0.31	0.26	0.49	0.45	**0.26**	**0.18**	0.14
Treed GP	0.73	**0.40**	0.47	**0.28**	**0.22**	1.40	**0.30**	3.40	0.20	**0.11**

Table 2: Mean absolute and square errors for methods on the CCS data set by training data size. The best results for each size of training data are in bold.

sponse. For this data set, we used the following DP-GLM,

$$\begin{aligned} P &\sim DP(\alpha \mathbb{G}_0), \\ \theta_i \,|\, P &\sim P, \\ X_{i,j} \,|\, \theta_i &\sim (p_{i,j,1}, \ldots, p_{i,j,K(j)}), \\ Y_i \,|\, \theta_i &\sim Poisson\left(\beta_{i,0} + \sum_{j=1}^{d} \sum_{k=1}^{K(j)} \beta_{i,j,k} \mathbf{1}_{\{X_{i,j}=k\}}\right). \end{aligned}$$

We used a conjugate covariate base measure and a Gaussian base measure for β,

$$(p_{j,1}, \ldots, p_{j,K(j)}) \sim Dirichlet(a_{j,1}, \ldots, a_{j,K(j)}), \qquad \beta_{j,k} \sim N(m_{j,k}, s^2_{j,k}).$$

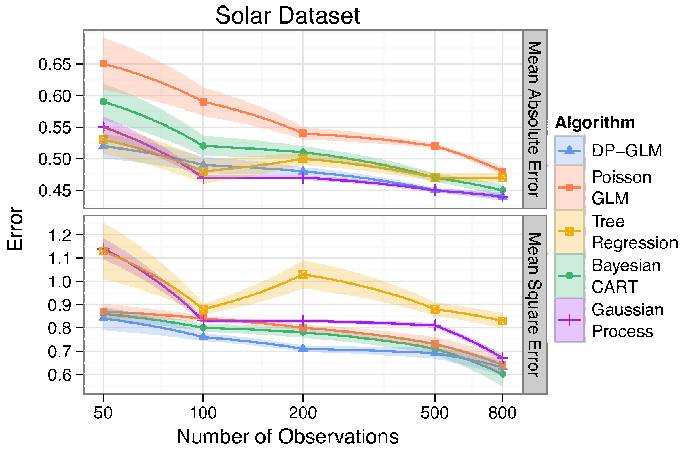

Figure 6: The average mean absolute error (top) and mean squared error (bottom) for tree regression, a Poisson GLM (GLM) and DP-GLM on the Solar data set. Mean $+/-$ one standard deviation are given for each method.

Method	Mean Absolute Error					Mean Squared Error				
	50	100	200	500	800	50	100	200	500	800
DP-GLM	**0.52**	0.49	0.48	**0.45**	**0.44**	**0.84**	**0.76**	**0.71**	**0.69**	0.63
Poisson Regression	0.65	0.59	0.54	0.52	0.48	0.87	0.84	0.80	0.73	0.64
CART	0.53	0.48	0.50	0.47	0.47	1.13	0.88	1.03	0.88	0.83
Bayesian CART	0.59	0.52	0.51	0.47	0.45	0.86	0.80	0.78	0.71	**0.60**
Gaussian Process	0.55	**0.47**	**0.47**	**0.45**	**0.44**	1.14	0.83	0.83	0.81	0.67

Table 3: Mean absolute and square errors for methods on the Solar data set by training data size. The best results for each size of training data are in bold.

Computational details are given in Appendix C.

The only other methods that can handle this data set are CART, Bayesian CART and Poisson regression. GP regression was run with a squared exponential covariance function and Gaussian errors to make use of the ordering in the covariates. The DP-GLM had good performance under both error measures. The high mean squared error values suggests that frequentist CART overfit while the high mean absolute error for Poisson regression suggests that it did not adequately fit nonlinearities. See Figure 6 and Table 3 for results.

6.6 Discussion

The DP-GLM is a relatively strong competitor on all of the data sets. It was more stable than most of its Bayesian competitors (aside from GPs) on the CCS data set. Our results suggest that the DP-GLM would be a good choice for small sample sizes when there is significant prior knowledge; in those cases, it acts as an automatic outlier detector and produces a result that is similar to a Bayesian

GLM. Results from Section 4 suggest that the DP-GLM is not appropriate for problems with high dimensional covariates; in those cases, the covariate posterior swamps the response posterior with poor numerical results.

7. Conclusions and Future Work

We developed the Dirichlet process mixture of generalized linear models (DP-GLM), a flexible Bayesian regression technique. We discussed its statistical and empirical properties; we gave conditions for asymptotic unbiasedness and gave situations in which they hold; finally, we tested the DP-GLM on a variety of data sets against state of the art Bayesian competitors. The DP-GLM was competitive in most setting and provided stable, conservative estimates, even with extremely small sample sizes.

One concern with the DP-GLM is computational efficiency as implemented. All results were generated using MCMC, which does not scale well to large data sets. An alternative implementation using variational inference (Blei and Jordan, 2006), possibly online variational inference (Sato, 2001), would greatly increase computational feasibility for large data sets.

Our empirical analysis of the DP-GLM has implications for regression methods that rely on modeling a joint posterior distribution of the covariates and the response. Our experiments suggest that the covariate posterior can swamp the response posterior, but careful modeling can mitigate the effects for problems with low to moderate dimensionality. A better understanding would allow us to know when and how such modeling problems can be avoided.

Acknowledgments

David M. Blei was supported by ONR 175-6343, NSF CAREER 0745520, AFOSR 09NL202, the Alfred P. Sloan foundation, and a grant from Google. Warren Powell and Lauren Hannah were supported in part by grant AFOSR contract FA9550-08-1-0195 and the National Science Foundation grant CMMI-0856153.

Appendix A.

In the Gibbs sampler, the state is the collection of labels $(z_1, \ldots, z_n)$ and parameters $(\theta^*_1, \ldots, \theta^*_K)$, where θ^*_c is the parameter associated with cluster c and K is the number of unique labels given $z_{1:n}$. In a collapsed Gibbs sampler, all or part of $(\theta^*_1, \ldots, \theta^*_K)$ is eliminated through integration. Let $z_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$. A basic inference algorithm is given in Algorithm 1. Convergence criteria for the Gibbs samplers in our numerical examples are given in Appendix C. See Gelman et al. (2004) for a more complete discussion on convergence criteria.

We can sample from the distribution $p(z_i \,|\, D, z_{-i}, \theta^*_{1:K})$ as follows,

$$p(z_i \,|\, D, z_{-i}, \theta^*_{1:K}) \propto p(z_i \,|\, z_{-i}) p(X_i \,|\, z_{1:n}, D, \theta^*_{1:K}) p(Y_i \,|\, X_i, z_{1:n}, D, \theta^*_{1:K}). \tag{11}$$

The first part of Equation (11) is the Chinese Restaurant Process posterior value,

$$p(z_i \,|\, z_{-i}) = \begin{cases} \frac{n_{z_j}}{n-1+\alpha} & \text{if } z_i = z_j \text{ for some } j \neq i, \\ \frac{\alpha}{n-1+\alpha} & \text{if } z_i \neq z_j \text{ for all } j \neq i. \end{cases}$$

Algorithm 1: Gibbs Sampling Algorithm for the DP-GLM

Require: Starting state $(z_1,\ldots,z_n)$, $(\theta^*_1,\ldots,\theta^*_K)$, convergence criteria.

repeat
  for $i = 1$ to n **do**
    Sample z_i from $p(z_i \mid D, z_{-i}, \theta^*_{1:K})$.
  end for
  for $c = 1$ to K **do**
    Sample θ^*_c given $\{(X_i, Y_i) : z_i = c\}$.
  end for
  if Convergence criteria are met **then**
    Record $(z_1,\ldots,z_n)$ and $(\theta^*_1,\ldots,\theta^*_K)$.
  end if
until M posterior samples obtained.

Here n_{z_j} is the number of elements with the label z_j. The second term of Equation (11) is the same as in other Gibbs sampling algorithms. If possible, the component parameters $\theta^*_{1:K}$ can be integrated out (in the case of conjugate base measures and parameters that pertain strictly to the covariates) and $p(X_i \mid z_{1:n}, D, \theta^*_{1:K})$ can be replaced with

$$\int p(X_i \mid z_{1:n}, D, \theta^*_{1:K}) p(\theta^*_{1:K} \mid z_{1:n}) d\theta^*_{1:K}.$$

The third term of Equation (11) is not found in traditional Dirichlet process mixture model samplers. In some cases, this term can also be collapsed, such as Gaussian model with a Normal-Inverse-Gamma base measure. In that case,

$$\begin{aligned}
p(Y_i \mid X_i, z_c, D_c) &= \frac{\Gamma((n_n+1)/2)}{\Gamma(n_n/2)} (n_n s_n)^{-1/2} \exp\left(-1/2(n_n+1)\log\left(1+\frac{1}{n_n s_n}(Y_i - m_n)^2\right)\right), \\
\tilde{V} &= \left(V^{-1} + \tilde{X}_c^T \tilde{X}_c\right)^{-1}, \\
\hat{m}_n &= \tilde{V}\left(m_0 V^{-1} + \tilde{X}_c^T Y_c\right), \\
m_n &= \tilde{X}_i \hat{m}_n, \\
n_n &= n_{y0} + n_c, \\
s_n^2 &= 4\left(s_{y0}^2 + 1/2\left(m_0 V^{-1} m_0^T + Y_c^T Y_c - \hat{m}_n^T \tilde{V}^{-1} \hat{m}_n\right)\right) / \left((n_{y0}+n_c)\tilde{X}_c \tilde{V} \tilde{X}_c^T\right).
\end{aligned}$$

Here, we define $\tilde{X}_c = \{[1 X_j] : z_j = z_c\}$, $Y_c = \{Y_j : z_j = z_c\}$, $\tilde{X}_i = [1 X_i]$, n_c is the number of data associated with label z_c and the base measure is define as,

$$\begin{aligned}
\sigma_y^2 &\sim Inverse - Gamma(n_{y0}, s_{y0}^2), \\
\beta \mid \sigma_y^2 &\sim N(m_0, \sigma_y^2 V).
\end{aligned}$$

Appendix B.

Proofs for the main theorems.

B.1 Proof of Theorem 1

Both Theorems 1 and 2 rely on a theorem by Schwartz (1965).

Theorem 5 (Schwartz, 1965) *Let Π^f be a prior on $\mathcal{F}$. Then, if Π^f places positive probability on all neighborhoods*

$$\left\{ f : \int f_0(x,y) \log \frac{f_0(x,y)}{f(x,y)} dxdy < \delta \right\}$$

for every $\delta > 0$, then Π^f is weakly consistent at f_0.

The proof for Theorem 1 follows closely both Ghosal et al. (1999) and Tokdar (2006).

Proof Without loss of generality, assume $d = 1$. Since f_0 has compact support, there exists an x_0 and a y_0 such that $f_0(x,y) = 0$ for $|x| > x_0$ or $|y| > y_0$. Fix $\varepsilon > 0$. Following Remark 3 of Ghosal et al. (1999), there exist $\bar{\sigma}_x > 0$ and $\bar{\sigma}_y > 0$ such that

$$\int_{-x_0}^{x_0} \int_{-y_0}^{y_0} f_0(x,y) \log \frac{f_0(x,y)}{\int_{-x_0}^{x_0} \int_{-y_0}^{y_0} \phi(\frac{x-\theta_x}{\bar{\sigma}_x}) \phi(\frac{y-\theta_y}{\bar{\sigma}_y}) f_0(x,y) d\theta_x d\theta_y} < \varepsilon/2.$$

Let P_0 be a measure on $\mathbb{R}^3 \times \mathbb{R}^2_+$, that is, a measure for $(\mu_x, \beta_0, \beta_1, \sigma_x, \sigma_y)$. Define it such that $dP_0 = f_0 \times \delta_0 \times \delta_{\bar{\sigma}_x} \times \delta_{\bar{\sigma}_y}$. Fix a $\lambda > 0$ and $\kappa > 0$. Choose a large compact set K such that $[-x_0, x_0] \times [-y_0, y_0] \times [-y_0, y_0] \times \{\bar{\sigma}_x\} \times \{\bar{\sigma}_y\} \subset K$.Let $\mathcal{B} = \{P : |P(K)/P_0(K) - 1| < \kappa\}$. Since the support of $\mathbb{G}_0$ is $\mathbb{R}^3 \times \mathbb{R}^2_+$, $\Pi(\mathcal{B}) > 0$.

Following Ghosal et al. (1999) and Tokdar (2006), it can be shown that there exists a set $\mathcal{C}$ such that $\Pi(\mathcal{B} \cap \mathcal{C}) > 0$ and for every $P \in \mathcal{B} \cap \mathcal{C}$,

$$\int_{-x_0}^{x_0} \int_{-y_0}^{y_0} f_0(x,y) \log \frac{\int_K \phi(\frac{x-\mu_x}{\sigma_x}) \phi(\frac{y-\beta_0-\beta_1 x}{\sigma_y}) dP_0}{\int_K \phi(\frac{x-\mu_x}{\sigma_x}) \phi(\frac{y-\beta_0-\beta_1 x}{\sigma_y}) dP} < \frac{\kappa}{1-\kappa} + 2\kappa < \varepsilon/2$$

for a suitable choice of κ. Therefore, for $f = \phi * P$ for every $P \in \mathcal{B} \cap \mathcal{C}$,

$$\begin{aligned} \int f_0(x,y) \log \frac{f_0(x,y)}{f(x,y)} dxdy \leq & \int_{-x_0}^{x_0} \int_{-y_0}^{y_0} f_0(x,y) \log \frac{f_0(x,y)}{\int_{-x_0}^{x_0} \int_{-y_0}^{y_0} \phi(\frac{x-\theta_x}{\bar{\sigma}_x}) \phi(\frac{y-\theta_y}{\bar{\sigma}_y}) f_0(x,y) d\theta_x d\theta_y} \\ & + \int_{-x_0}^{x_0} \int_{-y_0}^{y_0} f_0(x,y) \log \frac{\int_K \phi(\frac{x-\mu_x}{\sigma_x}) \phi(\frac{y-\beta_0-\beta_1 x}{\sigma_y}) dP_0}{\int_K \phi(\frac{x-\mu_x}{\sigma_x}) \phi(\frac{y-\beta_0-\beta_1 x}{\sigma_y}) dP} \\ & < \varepsilon. \end{aligned}$$

Therefore, Π^f places positive measure on all weak neighborhoods of f_0, and hence satisfies Theorem 5.

■

Proof [Theorem 2] The proof of Theorem 2 follows along the same lines as the proof for Theorem 1. Instead of the continuous response, however, there is a categorical response. The continuity condition on the response probabilities ensures that there exists a $y_0 > 0$ such that there are m continuous functions $b_1(x), \ldots, b_m(x)$ with $|b_i(x)| < y_0$ and

$$\mathbb{P}_{f_0}[Y = i \,|\, X = x] = \frac{\exp(b_i(x))}{\sum_{j=1}^m \exp(b_j(x))}.$$

Using arguments similar to those in the previous proof, there exists $\bar{\sigma}_x > 0$ such that,

$$\int_{-x_0}^{x_0} f_0(x,i) \log \frac{f_0(x,i)}{\int_{-x_0}^{x_0} \phi(\frac{x-\theta_x}{\bar{\sigma}_x}) f_0(x) \frac{\exp(b_i(x))}{\sum_{j=1}^m \exp(b_j(x))} d\theta_x} < \varepsilon/2.$$

Define P_0 such that $dP_0 = f_0(x) \times \{\bar{\sigma}_x\} \times b_1(x) \times \cdots \times b_m(x)$. The rest of the proof follows as previously, with small modifications. ■

B.2 Proof of Theorem 3

We now show pointwise convergence of the conditional densities. The following propositions will be used to prove Theorems 3 and 4.Let $f_n(x,y)$ be the Bayes estimate of the density under Π^f after n observations,

$$f_n(x,y) = \int_{\mathcal{F}} f(x,y) \Pi^f \left(df \,|\, (X_i, Y_i)_{i=1}^n \right).$$

Proposition 6 *Weak consistency of Π^f at f_0 for the Gaussian model and the multinomial model implies that $f_n(x,y)$ converges pointwise to $f_0(x,y)$ and $f_n(x)$ converges pointwise to $f_0(x)$ for (x,y) in the compact support of f_0.*

Proof Both $f_n(x,y)$ and $f_n(x)$ can be written as expectations of bounded functions with respect to the posterior measure. In the Gaussian case, both $f_n(x,y)$ and $f_n(x)$ are absolutely continuous; in the multinomial case, $f_n(x)$ is absolutely continuous while the probability $\mathbb{P}_{f_n}[Y = k \,|\, x]$ is absolutely continuous in x for $k = 1, \ldots, K$. Due to absolute continuity, the result holds. ■

This can be used to show that the conditional density estimate converges pointwise to the true conditional density.

Proposition 7 *Let $f_n(x,y)$ an $f_n(x)$ be as in Proposition 6. Then $f_n(y|x)$ converges pointwise to $f_0(y|x)$ for any (x,y) in the compact support of f_0.*

Proof From Proposition 6, $f_n(x,y)$ converges pointwise to $f_0(x,y)$ and $f_n(x)$ converges pointwise to $f_0(x)$. Then,

$$\lim_{n\to\infty} f_n(y|x) = \lim_{n\to\infty} \frac{f_n(x,y)}{f_n(x)} = \frac{f_0(x,y)}{f_0(x)} = f_0(y|x).$$

The denominator value, $f_n(x)$, is greater than 0 almost surely because it is a mixture of Gaussian densities. ■

Now we proceed to the proof of Theorem 3.

Proof [Theorem 3] The conditions for Theorem 3 assure that Propositions 6 and 7 hold. Because of this and the fact that $\mathbb{G}_0$ places positive measure only on densities with a finite expectation, the results hold. ■

B.3 Proof of Theorem 4

The proof follows in the same manner as that for Theorem 3.

Appendix C.

Implementation details.

C.1 CMB Computational Details

The DP-GLM was run on the largest data size tested several times; log posterior probabilities were evaluated graphically, and in each case the posterior probabilities seem to have stabilized well before 1,000 iterations. Therefore, all runs for each sample size were given a 1,000 iteration burn-in with samples taken every 5 iterations until 2,000 iterations had been observed. The scaling parameter α was given a Gamma prior with shape and scale set to 1. The means and variances of each component and all GLM parameters were also given a log-normal hyper distribution. The model was most sensitive to the hyper-distribution on σ_y, the GLM variance. Small values were used ($\log(m_y) \sim N(-3,2)$) to place greater emphasis on response fit. The non-conjugate parameters were updated using the Hamiltonian dynamics method of Neal (2010). Hyperparameters were chosen based on performance on a subset of 100 data points; values were then held fixed all other data sets. This may produce an overly confident error assessment, but the limited size of the data set did not allow a pure training-validation-testing three way partition. A non-conjugate base measure was used on this data set due to small sample sizes and heteroscedasticity. The conjugate measure, a normal-inverse-gamma, assumes a relationship between the variance and the mean,

$$\mu \,|\, \sigma^2, \lambda, \nu \sim N(\lambda, \sigma^2/\nu).$$

Therefore, smaller variances greatly encourage the mean μ to remain in a small neighborhood around around the prior value, λ. Naturally, this property can be overcome with many observations, but it makes strong statements about the mean in situations with few total samples or few samples per cluster due to heteroscedasticity. This model was implemented in Matlab; a run on the largest data set took about 500 seconds.

C.2 CCS Computational Details

Again, the DP-GLM was run on the largest data size tested several times; log posterior probabilities were evaluated graphically, and in each case the posterior probabilities seem to have stabilized well before 1,000 iterations. Therefore, all runs for each sample size were given a 1,000 iteration burn-in with samples taken every 5 iterations until 2,000 iterations had been observed. The scaling parameter α was given a Gamma prior with shape and scale set to 1. The hyperparameters of the conjugate base measure were set manually by trying different settings over four orders of magnitude for each parameter on a single subset of training data. Again, this may produce an overly confident error assessment, but the limited size of the data set did not allow a pure training-validation-testing three way partition. All base measures were conjugate, so the sampler was fully collapsed. α was updated using Hamiltonian dynamics (Neal, 2010). Original results were generated by Matlab; the longest run times were about 1000 seconds. This method has been re-implemented in Java in a highly efficient manner; the longest run times are now under about 10 seconds. Run times would likely be even faster if variational methods were used for posterior sampling (Blei and Jordan, 2006).

C.3 Solar Computational Details

Again, the DP-GLM was run on the largest data set size tested several times; log posterior probabilities were evaluated graphically, and in each case the posterior probabilities seem to have stabilized well before 1,000 iterations. Therefore, all runs for each sample size were given a 1,000 iteration burn-in with samples taken every 5 iterations until 2,000 iterations had been observed. The scaling parameter α was set to 1 and the Dirichlet priors to $Dir(1,1,\ldots,1)$. The response parameters were given a Gaussian base distribution with a mean set to 0 and a variance chosen after trying parameters with four orders of magnitude on a fixed training data set. This may produce an overly confident error assessment, but the limited size of the data set did not allow a pure training-validation-testing three way partition. All covariate base measures were conjugate and the β base measure was Gaussian, so the sampler was collapsed along the covariate dimensions and used in the auxiliary component setting of Algorithm 8 of Neal (2000). The β parameters were updated using Metropolis-Hastings. Results were in generated by Matlab; run times were substantially faster than the other methods implemented in Matlab (under 200 seconds).

References

R. P. Adams, I. Murray, and D. J. C. MacKay. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16. ACM, 2009.

M. Amewou-Atisso, S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency for semi-parametric regression problems. *Bernoulli*, 9(2):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.

C. L. Bennett, M. Halpern, G. Hinshaw, N. Jarosik, A. Kogut, M. Limon, S. S. Meyer, L. Page, D. N. Spergel, G. S. 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.

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, 2006.

G. Bradshaw. UCI machine learning repository, 1989.

L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman & Hall/CRC, New York, NY, 1984.

L. D. Brown. *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayward, CA, 1986.

H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian treed models. *Machine Learning*, 48 (1):299–320, 2002.

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.

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*, 89(425):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.

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

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, H. S. Stern, and D. S. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, 2004.

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

J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer-Verlag New York, Inc., New York, NY, 2003.

R. B. Gramacy and H. K. H. Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

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

J. E. Griffin and M. F. J. Steel. Bayesian nonparametric modelling with the Dirichlet process regression smoother. *Statistica Sinica*, 20(4):1507–1527, 2010.

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

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

E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964.

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

R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2010.

P. Z. G. Qian, H. Wu, and C. F. J. Wu. Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics*, 50(3):383–396, 2008.

C. E. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. In *Advances in Neural Information Processing Systems, 14*.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

A. Rodrıguez. *Some Advances in Bayesian Nonparametric Modeling*. PhD thesis, Duke University, 2009.

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

M. A. Sato. Online model selection based on the variational Bayes. *Neural Computation*, 13(7): 1649–1681, 2001.

L. Schwartz. On Bayes procedures. *Probability Theory and Related Fields*, 4(1):10–26, 1965.

B. Shahbaba and R. M. Neal. Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10:1829–1850, 2009.

S. Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: 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.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics*, 26(4):359–372, 1964.

M. West, P. Muller, and M. D. Escobar. Hierarchical priors and mixture models, with application in regression and density estimation. In *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.