Dirichlet Process Mixtures of Generalized Linear Models

Lauren A. Hannah
Department of Operations Research and Financial Engineering, Princeton University

David M. Blei
Department of Computer Science, Princeton University

Warren B. Powell
Department of Operations Research and Financial Engineering, Princeton University

Abstract

We propose Dirichlet Process mixtures of Generalized Linear Models (DP-GLMs), a new method of nonparametric regression that accommodates continuous and categorical inputs, models a response variable locally by a generalized linear model. We give conditions for the existence and asymptotic unbiasedness of the DP-GLM regression mean function estimate; we then give a practical example for when those conditions hold. We evaluate DP-GLM on several data sets, comparing it to modern methods of nonparametric regression including regression trees and Gaussian processes.

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,

$$Y \mid X \sim f(m(X)).$$

Here, $m(\cdot)$ 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.

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

Regression models are a central focus of statistics and machine learning. Our goal is to develop a model that can be used in many settings. Generalized linear models (GLMs) serve this purpose in a parametric setting. They pass a linear transformation of covariates through a possibly non-linear link function to generate a response. Inherent to GLMs, however, is an assumption that the response varies according to a linear transformation of the covariates. The method that we develop relieves this assumption.

We develop 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 nonparametric regression. A DP-GLM produces a regression model by modeling the joint distribution of the covariates and the response. This is done using a Dirichlet process (DP) mixture model: for each observation a hidden parameter θ is drawn, covariates are generated from a parametric distribution conditioned on θ, and then the response is drawn from a GLM conditioned on the covariates and θ. The clustering effect of the DP mixture leads to an "infinite mixture" of GLMs, a model which effectively identifies local regions of covariate space in which the covariates exhibit a consistent relationship to the response. In combination, these local GLMs represent a complex global response function. 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.

Like the Dirichlet process regression models of Muller et al. (1996) and Rodriguez et al. (2009), we model the joint distribution of the covariates and response to generate an implicit conditional response distribution. This method is conceptually similar to the double-kernel method of Fan et al. (1996), except that the kernel generated by the Dirichlet process is determined by the imparted distribution over partition structures rather than a traditional distance metric. The DP-GLM is a generalization of several existing DP-based regression models (Muller et al., 1996; Shahbaba and Neal, 2009) to a variety of covariate types and response

distributions. Bayesian nonparametric models have previously been proposed, but they lacked generality and/or theoretical guarantees, such as the existence of a mean function estimator or its asymptotic unbiasedness.

In this paper, we present the DP-GLM and give conditions under which a mean function estimate exists and is asymptotically unbiased. We review the Bayesian nonparametric regression literature in Section 2, review Dirichlet processes and GLMs in Section 3, present the DP-GLM in Section 4, and give theoretical properties in Section 5. In Section 6, we report on a study of the DP-GLM to several data sets, illustrating its flexibility to multiple types of covariates and response variables.

2 Bayesian Regression Literature

Nonparametric regression is a field that has received considerable study, but less so in a Bayesian context. Gaussian process (GP) and Dirichlet process mixtures are the most common prior choices for Bayesian nonparametric regression. GP priors assume that the observations arise from a Gaussian process model with known covariance function form (see Rasmussen and Williams (2006) for a review). Without heavy modification, however, the GP model is only applicable to problems with continuous covariates.

Dirichlet process priors have been used previously for regression. West et al. (1994); Escobar and West (1995) and Muller et al. (1996) used joint Gaussian mixtures for the covariates and response; Rodriguez et al. (2009) generalized this method using dependent DPs for multiple response functionals. This method can be slow if a fully populated covariance matrix is used and is potentially inaccurate if only a diagonal matrix is used. To avoid these over-fitting the covariate distribution and under-fitting the response, local weights on the covariates have been used to produce local response DPs, with kernels and basis functions (Griffin and Steel, 2007; Dunson et al., 2007), GPs (Gelfand et al., 2005) or general spatial-based weights (Griffin and Steel, 2006, 2007; Duan et al., 2007). Other methods, such as dependent DPs, have been introduced to capture similarities between clusters, covariates or groups of outcomes (De Iorio et al., 2004; Rodriguez et al., 2009). The DP-GLM tries to balance fitting the covariate and response distributions by introducing local GLMs–the clustering structure is heavily based on the covariates, but within each cluster response fit is better because it is represented by a GLM rather than a constant. This method is simple and flexible.

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. 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 of this idea.

The asymptotic properties of Dirichlet process regression models have not been well studied. Most current literature centers around 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).Only recently have the posterior properties of DP regression estimators been studied. Rodriguez et al. (2009) showed pointwise 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).

3 Mathematical Background

Dirichlet Process Mixture Models. In a Bayesian mixture model, we assume that the true density of the covariates X and response Y can be written as a mixture of parametric densities, such as Gaussians or multinomials, conditioned on a hidden parameter θ. For example, in a Gaussian mixture, θ includes the mean μ and variance σ^2. Due to our model formulation, θ is 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 $\theta = (\theta_x, \theta_y)$. 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.

The Dirichlet process models uncertainty about the prior density P (Ferguson, 1973; Antoniak, 1974). 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),

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

(Lower case values refer to observed or fixed values, while upper case refer to random variables.)

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 parameters to share values. 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 a new component.

In a DP mixture, θ is a latent parameter to an observed data point x (Antoniak, 1974),

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

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. Moreover, a new data point can be assigned to a new cluster that was not previously seen in the data.

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) for a full discussion.)

GLMs have three components: the conditional probability model for response Y, the linear predictor and the link function. The probability model for Y, dependent on covariates X, is

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

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. A linear predictor, $X\beta$, is used to determine the canonical parameter through a set of transformations. It can be shown that $b'(\eta) = \mu = \mathbb{E}[Y|X]$. However, we can choose a link function g such that $\mu = g^{-1}(X\beta)$, which defines η in terms of $X\beta$. The canonical form is useful for discussion of GLM properties, but we use the mean form in the rest of this paper. The flexible nature of GLMs allows us to use them as a local approximation for a global response function.

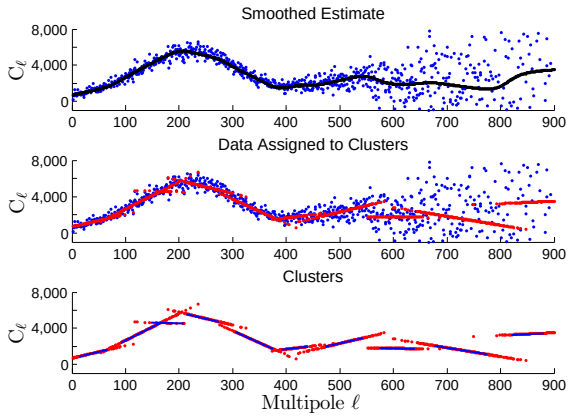

Figure 1: The top figure shows the smoothed regression estimate for the Gaussian model of equation (2). The center figure shows the training data (blue) fitted into clusters, with the prediction given a single sample from the posterior, $\theta^{(i)}$ (red). The bottom figure shows the underlying clusters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

4 Dirichlet Process Mixtures of Generalized Linear Models

We now develop 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. We present theoretical properties of the DP-GLM in Section 5.

4.1 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 with 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 GLM

dispersion parameter. The full model is

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

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, i.e., in "training," 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. 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 an example of the DP-GLM for continuous covariates/response that will be used throughout the rest of the paper.

Example: Gaussian Model. 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; covariance matrix is diagonal for simplicity. 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,

$$\begin{aligned} P &\sim DP(\alpha \mathbb{G}_0), \\ \Theta_i|P &\sim P, \\ X_{ij}|\mu_{ij}, \sigma_{ij} &\sim N(\mu_{ij}, \sigma^2_{ij}), \quad j = 1, \ldots, d, \\ Y_i|x_i, \beta_i, \sigma_{iy} &\sim N(\beta_{i0} + \sum_{j=1}^{d} \beta_{ij} x_{ij}, \sigma^2_{iy}). \end{aligned} \tag{2}$$

4.2 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 that generated the observed data, $\theta_{1:n}$, the expectation

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 $\theta^{(m)}_{1:n}|(X_j, Y_j)_{1:n}$;
  Compute $\mathbb{E}[Y|x, \theta^{(m)}_{1:n}]$;
end
Set $\bar{m}(x) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[Y|x, \theta^{(m)}_{1:n}]$;

of the response is

$$\begin{aligned} \mathbb{E}[Y|x, \theta_{1:n}] &= \frac{\alpha}{b} \int_{\mathcal{T}} \mathbb{E}\left[Y|x, \theta\right] f_x(x|\theta) \mathbb{G}_0(d\theta) \\ &\quad + \frac{1}{b} \sum_{i=1}^{n} \mathbb{E}\left[Y|x, \theta_i\right] f_x(x|\theta_i), \\ b &= \alpha \int_{\mathcal{T}} f_x(x|\theta) \mathbb{G}_0(d\theta) + \sum_{i=1}^{n} f_x(x|\theta_i). \end{aligned} \tag{3}$$

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.

As $\theta_{1:n}$ is not actually known, the unobserved random variables $\Theta_{1:n}$ are integrated out of equation (3) using the posterior distribution given the observed data. 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,

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

where $A \subseteq \mathcal{F}$. Define Π^P_n similarly. The regression is

$$\begin{aligned} &\mathbb{E}[Y|x, (X_i, Y_i)_{1:n}] \\ &= \frac{1}{b} \sum_{i=1}^{n} \int_{\mathcal{M}_{\mathcal{T}}} \int_{\mathcal{T}} \mathbb{E}[Y|x, \theta_i] f_x(x|\theta_i) P(d\theta_i) \Pi^P_n(dP) \\ &\quad + \frac{\alpha}{b} \int_{\mathcal{T}} \mathbb{E}\left[Y|x, \theta\right] f_x(x|\theta) \mathbb{G}_0(d\theta), \end{aligned} \tag{4}$$

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

Equation (4) is difficult to compute because it requires integration over a hidden random measure. To avoid this problem, we approximate equation (4) by an average of M Monte Carlo samples of the expectation

conditioned on $\theta_{1:n}^{(m)}$,

$$\mathbb{E}[Y|x,(X_i,Y_i)_{1:n}] \approx \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[Y|x,\theta_{1:n}^{(m)}\right]. \quad (5)$$

The regression procedure is given in Algorithm 1. We describe how to generate posterior samples $\{\theta_{1:n}^{(m)}\}_{m=1}^{M}$ in Section 4.3.

Example: Gaussian Model. Continuing our example of the Gaussian model, equation (3) becomes

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

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

and $\phi_\sigma(x)$ is the Gaussian density at x with variance σ^2. An example of regression for a single covariate Gaussian model is shown in Figure 1.

4.3 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 Neal (2000) for state of the art algorithms). We use Algorithm 8 of Neal (2000), but conjugate base measures allow the use of the more efficient collapsed sampler, Algorithm 3.

5 Theoretical Properties

Two pitfalls might arise when using the DP-GLM: an inability to compute the mean function and mean function bias that does not tend toward 0. A desirable property of any estimator is that it should be unbiased. If it holds in the limit, this property is called asymptotic unbiasedness. 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 avoids these problems; we sketch the ideas needed to show asymptotic unbiasedness and mean function existence and then give theorems for asymptotic unbiasedness and mean function existence.

5.1 Consistency and Theoretical Properties

Consistency, the notion that as the number of observations goes to infinity the posterior distribution accumulates in neighborhoods arbitrarily "close" to the true distribution, is tightly related to both asymptotic unbiasedness and mean function estimate existence. 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. Note that an expectation is not bounded; in addition to weak consistency, uniform integrability is needed to guarantee that the posterior expectation converges to the true expectation, giving asymptotic unbiasedness. Uniform integrability also ensures that the posterior expectation almost surely exists with every additional observation. Therefore we need to show weak consistency and uniform integrability.

With the inclusion of covariates, the observations of covariate response pairs are not identically distributed, so we use a modified variant of the Schwartz (1965) theorem for weak posterior consistency, given in Amewou-Atisso et al. (2003). We assume that covariates are observed from the entire distribution; we give conditions for consistency for a predictor x in a compact subset $\mathcal{C}$ of the covariate domain. In practice, this is not particularly restrictive as $\mathcal{C}$ can be made arbitrarily large.

5.2 Asymptotic Unbiasedness

We are now ready to state the main theorem.

Theorem 5.1. *Let x be in a compact set $\mathcal{C}$ and Π^f be a prior on $\mathcal{F}$. If,*

(i) for every $\delta > 0$, Π^f puts positive measure on

$$\left\{f : \int f_0(x,y)\log\frac{f_0(x,y)}{f(x,y)}dxdy < \delta, \int f_0(x,y)\left(\log\frac{f_0(x,y)}{f(x,y)}\right)^2 dxdy < \delta\right\},$$

(ii) $\int |y|^2 f_0(y|x)dy < \infty$ for every $x \in \mathcal{C}$, and

(iii) there exists an $\epsilon > 0$ such that for every $x \in \mathcal{C}$,

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

then for every $n \geq 0$, $\mathbb{E}_\Pi[Y|x,(X_i,Y_i)_{1:n}]$ exists and has the limit $\mathbb{E}_{f_0}[Y|x]$, almost surely $\mathbb{P}_{F_0^\infty}$.

The conditions of Theorem 5.1 must be checked for the problem (f_0) and prior (Π^f) pair, and can be difficult

to show. Condition (i) assures weak consistency of the posterior, condition (ii) guarantees a mean function exists in the limit and condition (iii) guarantees that positive probability is only placed on densities that yield a finite mean function estimate. See Hannah et al. (2009) for a discussion and proof.

5.3 Example: Gaussian Model with a Conjugate Base Measure

The next theorem gives an example of when Theorem 5.1 holds for the Gaussian model. Discussion, proof and extensions are given in Hannah et al. (2009).

Theorem 5.2. *Let (X, Y) have the joint Gaussian model. If, for a compact covariate set $\mathcal{C}$,*

(i) $\int f_0(x,y)(\log f_0(x,y))^2 dxdy < \infty$,

(ii) $\int |y|^2 f_0(y|x)dy < \infty$ *for every* $x \in \mathcal{C}$, *and*

(iii) $\mathbb{G}_0$ *is conjugate to the Gaussian model, that is,*

$$\beta_{0:d}, \sigma_y \sim N - Inv - Gamma(\nu_y, \Xi_y, r_y, \lambda_y),$$
$$\mu_i, \sigma_i \sim N - Inv - Gamma(\nu_i, \xi_i, r_i, \lambda_i),$$

and $r_y, r_{1:d} \in (1/2, 1)$, then the conditions of Theorem 5.1 are satisfied.

6 Empirical Study

We compare the performance of DP-GLM regression to other regression methods. We chose data sets to illustrate the strengths of the DP-GLM, including ability to model different response/covariate types, and robustness with respect to heteroscedasticity and moderate dimensionality.

We compare to the following algorithms:

Naive Ordinary Least Squares (OLS). A parametric method that often provides a reasonable fit when there are few observations.

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). 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 μ_y, rather than $\beta_0 + \sum \beta_i x_i$. That is,

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

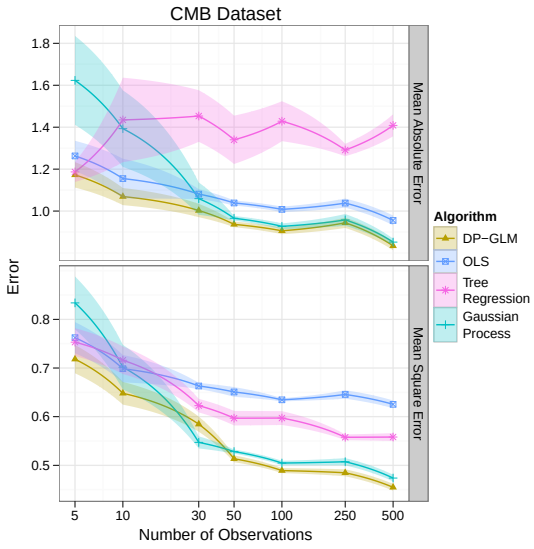

Figure 2: The average mean absolute error (top) and mean squared error (bottom) +/− one standard deviation for ordinary least squares (OLS), tree regression, Gaussian processes and DP-GLM on the CMB data set. The data were normalized.

Without the GLM response, the model cannot interpolate well in higher dimensions, leading to poor predictive performance.

Poisson GLM (GLM). A Poisson generalized linear model (count responses), used on the Solar Flare data set.

With these methods, we examined three data sets:

Cosmic Microwave Background (CMB) Results. The CMB dataset Bennett et al. (2003) 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 are highly nonlinear and heteroscedastic. Competitors were OLS, regression trees and Gaussian processes. 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. Results are given in Figure 2.

Concrete Compressive Strength (CCS) Results. The CCS Yeh (1998) dataset has 8 continuous covariates. The response is the compressive strength of the resulting concrete, also continuous. There are 1,030 observations. The data have relatively little noise. Competitors were OLS, GPs and regression

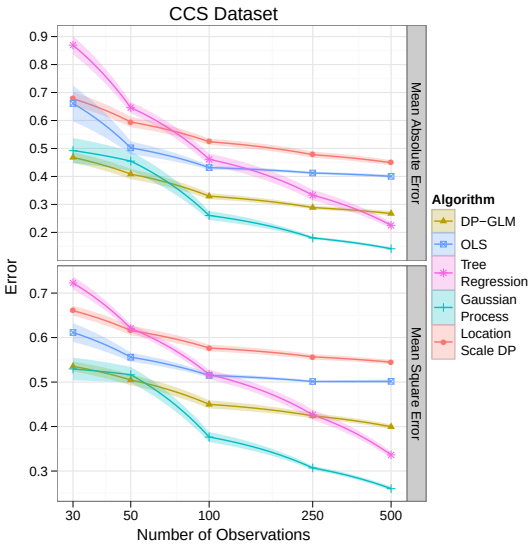

Figure 3: The average mean absolute error (top) and mean squared error (bottom) +/− one standard deviation for ordinary least squares (OLS), tree regression, Gaussian processes, location/scale DP and the DP-GLM Poisson model on the CCS data set. The data were normalized.

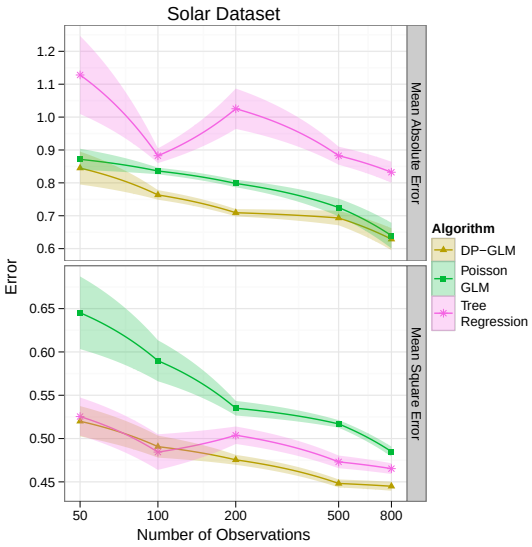

Figure 4: The average mean absolute error (top) and mean squared error (bottom) +/− one standard deviation for tree regression, a Poisson GLM (GLM) and DP-GLM on the Solar data set.

trees. We also included a basic DP regression technique (location/scale DP) on this data set. 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 mean and log-Gaussian scale base measures were used. Results are given in Figure 3.

Solar Flare Results. The Solar Bradshaw (1989) dataset was chosen to demonstrate the flexibility of DP-GLM. The response is the number of solar flares in a 24 hour period in a given area. There 1,389 observations and 11 categorical covariates. Competitors were tree regression and a Poisson GLM. GPs and other methods cannot be used for count/categorical data. 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. A Dirichlet covariate and Gaussian slope base measure was used with a Poisson response distribution. Results are given in Figure 4.

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 (σ_{iy}) 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, and the DP-GLM has many desirable properties that tree regression does not, such as a smooth mean function estimate and less sensitivity to bandwidth/pruning level.

References

Amewou-Atisso, M., Ghosal, S., Ghosh, J. and Ramamoorthi, R. (2003), 'Posterior consistency for semi-parametric regression problems', *Bernoulli* **9**(2), 291–312.

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

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

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

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

Bradshaw, G. (1989), 'UCI machine learning repository'.

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

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

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

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

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

Fan, J., Yao, Q. and Tong, H. (1996), 'Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems', *Biometrika***83**(1), 189–204.

Ferguson, T. (1973), 'A Bayesian analysis of some nonparametric problems', *The Annals of Statistics* **1**(2), 209–230.

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

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

Ghosh, J. and Ramamoorthi, R. (2003), *Bayesian Nonparametrics*, Springer.

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

Griffin, J. and Steel, M. (2007), Bayesian nonparametric modelling with the Dirichlet process regression smoother, Technical report, Institute of Mathematics, Statistics and Actuarial Science, University of Kent.

Hannah, L., Blei, D. and Powell, W. (2009), Dirichlet Process Mixtures of Generalized Linear Models. arxiv:0909.5194v1.

Ibrahim, J. and Kleinman, K. (1998), Semiparametric Bayesian methods for random effects models, *in* 'Practical Nonparametric and Semiparametric Bayesian Statistics', chapter 5, pp. 89–114.

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

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

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

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

Rasmussen, C. and Williams, C. (2006), *Gaussian processes for machine learning*, Springer.

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

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

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

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

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

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

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