Published in final edited form as:

Biometrics. 2016 September ; 72(3): 707–719. doi:10.1111/biom.12462.

Flexible Link Functions in Nonparametric Binary Regression with Gaussian Process Priors

Dan Li¹, Xia Wang¹, Lizhen Lin², and Dipak K. Dey³

Xia Wang: xia.wang@uc.edu

¹Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221, United States

²Department of Statistics and Data Sciences, University of Texas at Austin, Austin, Texas 78712, United States

³Department of Statistics, University of Connecticut, Storrs, Connecticut 06269, United States

Summary

In many scientific fields, it is a common practice to collect a sequence of 0-1 binary responses from a subject across time, space, or a collection of covariates. Researchers are interested in finding out how the expected binary outcome is related to covariates, and aim at better prediction in the future 0-1 outcomes. Gaussian processes have been widely used to model nonlinear systems; in particular to model the latent structure in a binary regression model allowing nonlinear functional relationship between covariates and the expectation of binary outcomes. A critical issue in modeling binary response data is the appropriate choice of link functions. Commonly adopted link functions such as probit or logit links have fixed skewness and lack the flexibility to allow the data to determine the degree of the skewness. To address this limitation, we propose a flexible binary regression model which combines a generalized extreme value link function with a Gaussian process prior on the latent structure. Bayesian computation is employed in model estimation. Posterior consistency of the resulting posterior distribution is demonstrated. The flexibility and gains of the proposed model are illustrated through detailed simulation studies and two real data examples. Empirical results show that the proposed model outperforms a set of alternative models, which only have either a Gaussian process prior on the latent regression function or a Dirichlet prior on the link function.

Keywords

Flexibility; Gaussian process; Generalized extreme value distribution; Latent variable; Markov chain Monte Carlo; Posterior consistency

Correspondence to: Xia Wang, xia.wang@uc.edu.

Supplementary Materials: Web Appendices, Tables, and Figures referenced in Sections 2-5 are available with this paper at the Biometrics website on Wiley Online Library. A simulated data set analyzed in Section 4 and **MATLAB** codes implementing the proposed method are also available with this paper at the Biometrics website on Wiley Online Library.

1. Introduction

Binary measurements are commonly available in many fields. It is usually of interest to study how these 0-1 measurements are related to different levels of covariates. For example, the researchers are interested in how some type of “yes” or “no” responses by subjects change along with time or other covariates (Smith et al., 2009). The binary regression model is a useful technique for analyzing binary data. In a binary regression model, the response probability function is specified as $p_i = P(Y_i = 1 | \mathbf{x}_i) = H(w(\mathbf{x}_i))$, $i = 1, \dots, n$, where y_i is the binary response and $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,k})'$ is a k -dimensional vector of covariates. The covariate structure $w(x)$ is called the latent regression function. The common latent regression function is $w(\mathbf{x}_i) = \mathbf{x}_i' \boldsymbol{\beta}$ with a vector of unknown parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)'$. The inverse of H serves as the link function. The logistic, probit and complementary log-log (cloglog) are commonly used.

Figure 1(a) shows binary time series responses from a monkey performing the attention paradigm described in Smith et al. (2009). The task consists of making a saccade to a visual target followed by a variable period of fixation on the target and detection of a change in target color followed by a bar release. The monkey must release the bar within a brief time cued by the change in target color in order to get a reward. The data include a temporal sequence of binary observations with 1's corresponding to reward being delivered and 0's corresponding to reward not being delivered at each trial (circles at 0 and 1 in Figure 1(a)). To ease the visualization on how the reward evolves over time, Figure 1(a) also includes the percentages of 1's in every 40 consecutive trials (solid dots connected by the solid curve). Two complications arise in employing the “standard” binary regression model. First, it is clear that the proportions of the rewarded trials do not evolve with time by a simple, monotonic trend. A parametric model such as $w(t_i) = \beta_0 + \beta_1 \times t_i$, where t_i 's are time points, is apparently not appropriate for the data. Second, the rate of the response probability curve does not necessarily approach zero at the same rate as that approaches one in each local up and down trends in the curve, especially for trials after 1000.

The second motivating example provides another common scenario in many fields such as clinical trials, occupational epidemiology and public health where the latent regression function includes more covariates other than time. As part of the National Coal Board's Pneumoconiosis Field Research (Fay, 1957), the binary observations were collected on 916 men along with the length of their employment as coal-getters on the coalface or as haulage-workers in the underground roadways. Figure 1(b) shows the data on the proportions of men having pneumoconiosis with different length of employment as coal-getters or haulage-workers. Similarly, the usual binary regression model may not be appropriate because of the potential non-linearity in the latent regression function and the assumed skewness in a given link function, such as the symmetric probit and the asymmetric cloglog links.

Limiting the latent regression function to a simple linear or parametric form is clearly restrictive in modeling the binary data. Parametric modeling often leads to inconsistent estimates when the model is mis-specified (Frölich, 2006). Moreover, mis-specification in the link function leads to an increase in mean square error of the estimated probability and substantial bias in estimating the regression parameters and the mean response probability

(Czado and Santner, 1992). In an effort to create a more flexible binary regression model, we adopt a semiparametric Bayesian approach by imposing a nonparametric Gaussian process (GP) prior on the latent regression function of covariates and proposing a parametric class of the generalized extreme value (GEV) link functions with an unconstrained shape parameter. Our methodology allows flexible inference in binary regression models without placing structural restrictions on $w(\cdot)$ or assuming a fixed cumulative function H .

There are a large collection of research work on flexible binary regression modeling by using Bayesian semi-parametric or nonparametric approaches. A key difference among these models is about where the nonparametric prior is imposed. It can be applied on the inverse link function H , or the latent regression function $w(\cdot)$, or the joint response-covariate distribution $f(y, x)$. Below we briefly review these three lines of research to place our proposed model in the line of ongoing Bayesian semi-parametric and nonparametric binary regression modeling. It also helps show our model's connection with and difference from the existing literature.

Earlier work investigated construction of flexible nonparametric link functions by putting a Dirichlet process prior on H . The work of Newton et al. (1996) and Basu and Mukhopadhyay (2000) fall in this category. While these models provide flexible symmetric and asymmetric links, they usually assume linearity in $w(x)$, which restricts the response curve to be monotonic. As the motivating data example 1 shows, monotonicity may not be a reasonable assumption. In order to relax this restriction, polynomial regression or a linear combination of other basis functions need to be included to accommodate the nonlinearity and potential interaction effects. This adds to complexity and subjectiveness of the model.

Another line of research models the latent regression function $w(x)$ by placing a GP prior on it. The GP provides a natural framework for modeling nonlinear relationship. It has recently become widely used in nonlinear regression, classification, and clustering (Neal, 1999; Rasmussen and Williams, 2006; Choudhuri et al., 2007; Iwata et al., 2013). GP models have superior predictive performance to many other nonparametric regression models in various tasks with additional advantages of conceptual simplicity, interpretability, and flexibility (Rasmussen, 1997). A much less investigated aspect is the choice of link function in the model. Most of the GP binary regressions are only limited to symmetric links, probit and logit links, for computational convenience. There exists a large body of literature showing that symmetric link functions are not always appropriate or adequate for fitting binary data, especially when the probability of the binary response approaches 0 at a different rate from that of approaching 1 (Chen et al., 1999; Kim et al., 2008; Wang and Dey, 2010).

A more general nonparametric approach is to model the joint distribution of the response and the covariates with Dirichlet process priors (Shahbaba and Neal, 2009; Dunson and Bhattacharya, 2011; Hannah et al., 2011; DeYoreo and Kottas, 2015). The previously discussed two semi-parametric methods are both restrictive in the sense that they allow only one structure component of the regression model to be estimated from data. Instead, the “fully” nonparametric Bayesian approach models the link function and the latent regression function simultaneously. It is a very flexible approach and to some degree it is a Bayesian analogy of a locally linear kernel estimator of the response curve (Hannah et al., 2011).

We would distinguish our work from the previous semi-parametric and nonparametric approaches. Our contribution is to investigate the importance of an appropriate link function in a GP binary regression model. Choudhuri et al. (2007) studied GP binary regression with a probit or an arbitrary symmetric link function. They argued that once the functional form of the latent regression function allows to be arbitrary, there is no loss of generality in shape by keeping the link function fixed, such as a probit link. As the GP with a given covariance kernel structure does not allow totally arbitrary skewness, our results support a flexible parametric link function which can further improve the modeling fitting. By conjoining a nonparametric GP prior and a parametric GEV link function, the proposed semi-parametric model (GP-GEV model) provides flexibility in detecting the skewness in the response curve while allowing nonparametric feature learning. Our results show that the GP prior component allows global nonlinear latent regression function, and the GEV link function component accommodates local skewness in the response curve. The GP-GEV model outperforms several alternatives in the simulated datasets under various scenarios. Its flexibility is also illustrated in the two above-mentioned motivating examples. Besides proposing the new GP-GEV model, we develop an efficient sampling algorithm for the model with a free-shaped GEV link function in our Bayesian computation. To our best knowledge, it is the first attempt at sampling a GP binary model with a link function that falls out of the normal or logistic family, thus provides a valuable addition to the sampling literature.

The paper is organized as follows. Section 2 presents the GP binary regression model. The GEV link is introduced and compared with the probit link and the cloglog link by studying the response probability functions. Additionally, we prove the posterior consistency of our model. In Section 3, we specify the priors, develop an efficient sampling algorithm for posterior inference, and describe model selection criteria. Section 4 discusses findings from simulated data sets by comparing the proposed model with a set of alternatives. We analyze the two motivating data examples in Section 5 and conclude in Section 6.

2. GP-GEV Binary Regression Model

The GP binary regression model is introduced in Section 2.1. In the literature, a binary regression model is called nonparametric when a GP prior is used in modeling the potential nonlinear latent structure. The proposed GP-GEV model is constructed by employing a GEV link on the GP binary regression model in Section 2.2. Section 2.3 presents the results on the posterior consistency of the GP-GEV model.

2.1 Gaussian Process Binary Regression Model

A random, real-valued function $w(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^k$, is said to follow a GP denoted as $\text{GP}(\mu(\cdot), R(\cdot, \cdot))$ with mean function $\mu(\mathbf{x})$ and covariance kernel $R(\mathbf{x}_i, \mathbf{x}_j)$ if given any finite n distinct vectors, $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^k$, $(w(\mathbf{x}_1), \dots, w(\mathbf{x}_n))'$ follows a multivariate Gaussian distribution with mean vector $(\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n))'$ and covariance matrix Σ . That is, $(w(\mathbf{x}_1), \dots, w(\mathbf{x}_n))' \sim N((\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n))', \Sigma)$, where $\Sigma_{i,j} = R(\mathbf{x}_i, \mathbf{x}_j)$. A common choice of the covariance kernel R is the squared exponential kernel,

$$(\sum_{\theta})_{ij} = R(\mathbf{x}_i, \mathbf{x}_j) = \alpha \exp \left\{ - \sum_{d=1}^k (x_{i,d} - x_{j,d})^2 / \beta_d \right\} \quad (1)$$

with a set of hyperparameters $\theta = \{\alpha, \{\beta_d\}_{d=1}^k\}$. The scaling parameter α controls the vertical scale of variation of the response function and the length-scale parameters $\beta_1, \dots, \beta_k$ control the smoothness of sample paths (Rasmussen and Williams, 2006). A GP with covariance structure (1) supports a large class of functions with various shapes.

We are interested in using GP as a prior distribution over the latent regression function in a binary regression model. Denote the observed data as $\mathcal{D} = \{\mathbf{y}, \mathbf{X}\} = \{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$, where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ is an $n \times k$ matrix with covariates $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,k})'$ and $\mathbf{y} = (y_1, \dots, y_n)'$ are 0-1 valued responses. Let $\mathbf{w} = (w_1, w_2, \dots, w_n)'$ be latent variables with $w_i = w(\mathbf{x}_i)$, $i = 1, \dots, n$. Define $p_i = P(y_i = 1) = H(w_i | \phi)$, $i = 1, 2, \dots, n$, where $H(\cdot | \phi)$ is a cumulative distribution function depending on the potentially vector valued indexing parameter ϕ , and H^{-1} is called the link function. The GP binary regression model, also called a nonparametric binary regression model with a GP prior, is then given by

$$y_i | p_i \stackrel{\text{ind}}{\sim} \text{Bin}(1, p_i), i = 1, \dots, n, \quad (2)$$

$$p_i = E(Y_i = 1 | \mathbf{x}_i) = H(w_i | \mathbf{x}_i, \theta, \phi), \quad (3)$$

$$\mathbf{w} | \mathbf{X}, \theta = (w_1, w_2, \dots, w_n)' \sim N(0, \sum_{\theta}). \quad (4)$$

2.2 Response Probability Function with Different Link Functions

A key component of the model given in (2)-(4) is the specification of the link function. The commonly used probit link is specified as $w_i = H^{-1}(p_i) = \Phi^{-1}(p_i)$, where Φ^{-1} is the inverse cumulative distribution function of $N(0, 1)$ distribution. The symmetry in the normal distribution leads to the symmetry in the probit link. Wang and Dey (2010) showed that the symmetric link has an inferior performance when the data structure requires a skewed response probability function. They proposed a link function based on the GEV distribution. The distribution function of GEV (μ, σ, ξ) is given by:

$$F(x | \mu, \sigma, \xi) = \begin{cases} \exp \left[- \left\{ 1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right\}_+^{-\frac{1}{\xi}} \right], & \xi \neq 0 \\ \exp \left\{ - \exp \left(- \frac{x - \mu}{\sigma} \right) \right\}, & \xi = 0, \end{cases} \quad (5)$$

where $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$ and $\xi \in \mathbb{R}$ are respectively the location, scale and shape parameters, and $x_+ = \max(0, x)$. The shape of this distribution function is very flexible with the tail behavior controlled by the shape parameter ξ . When $\xi = 0$, it is the Gumbel distribution and decays exponentially. When $\xi < 0$, it reduces to the negative Weibull distribution with a finite short upper endpoint. When $\xi > 0$, it becomes the Fréchet distribution with a heavy tail behavior. The GEV link is the inverse of H which is assumed as

$$p_i = H(w_i | \xi) = 1 - \text{GEV}(-w_i; \xi) = \begin{cases} 1 - \exp\left\{-(1 - \xi w_i)_+^{-\frac{1}{\xi}}\right\}, & \xi \neq 0 \\ 1 - \exp\{-\exp(w_i)\}, & \xi = 0, \end{cases} \quad (6)$$

where $\text{GEV}(x; \xi)$ represents the cumulative probability at x for the GEV distribution with parameters $\phi = (\mu = 0, \sigma = 1, \xi)'$. Note μ and σ are set to fixed constants for model identifiability. Wang and Dey (2010) showed that the GEV link model specified in (6) is negatively skewed for $\xi < \log 2 - 1$ and positively skewed for $\xi > \log 2 - 1$. The link function is approximately symmetric at $\xi = \log 2 - 1$. The cloglog link, specified as $w_i = H^{-1}(p_i) = \log\{-\log(1 - p_i)\}$, is a special case of the GEV link with $\xi = 0$.

The GP binary regression model under the GEV link is then given by (2) and (4) along with $p_i = \mathbb{E} Y_i = 1 | \mathbf{x}_i = H(w_i | \mathbf{x}_i, \boldsymbol{\theta}, \xi)$, where the parameters $\boldsymbol{\theta}$ controls the GP and ξ controls skewness in the GEV link model. Figure 2 shows the response curves under different link functions with the GP prior on the latent regression function. The x -axis represents x_i generated from a univariate standard normal distribution, $i = 1, \dots, n$. The scaling parameter α is fixed at 1 while the length-scale parameter β varies with values 0.25, 4, and 16. Given $\mathbf{X}$ and the hyperparameters, the latent variables $\mathbf{w}$ are generated from Equation (4). The response curves $E(Y_i = 1 | x_i)$ are then obtained by Equation (3) under the probit, cloglog, and GEV ($\xi = -1$ and 1) links, respectively. Comparison across the three panels ($\beta = 0.25, 4$, and 16) shows that the GP structure controls the global structural trend in the response curves. The difference among the four response curves within each panel indicates that the link function influences the local structure of the response curves. For example, when $\beta = 4$, the four response curves have similar global shapes. However, the probit link (dot-dashed line) exhibits local symmetric feature, while the GEV links show local positive skewness ($\xi = 1$, dotted line) or negative skewness ($\xi = -1$, solid line). The positively skewed cloglog link response curve (dashed line) overlaps with the GEV link response curve at $\xi = 0$. Similar features can be observed when $\beta = 0.25$ and $\beta = 16$.

We employ Bayesian computation to obtain inference of the GP-GEV model. Sampling positive variables can be realized by using Gamma types of proposals (Robert and Casella, 2005). Each positive GP hyperparameter is given a diffuse inverse-gamma prior with a large variance. A priori all GP hyperparameters are independent. For the shape parameter ξ of the GEV distribution, either a diffuse prior or a weakly informative normal prior is possible (Northrop and Attalides, 2015). We will discuss the posterior consistency of our model under these priors in next section. Details of the priors, including the values of hyperparameters, will be discussed in Section 3.1.

Let $\Theta = \{\xi, w\}$, which consists of the shape parameter ξ and the latent function w . Strictly speaking, the parameter vector Θ is *unidentifiable* (see a proof in Web Appendix A). That is, one can find two distinct values Θ_1 and Θ_2 ($\Theta_1 \neq \Theta_2$) such that $p(\text{data} | \Theta_1) = p(\text{data} | \Theta_2)$ for all the data. However, model identifiability is a concept of likelihood, regardless of the estimation paradigms (whether it is frequentist or Bayesian) (see Kadane (1974); Poirier (1998)). Model unidentifiability does not imply that there is no Bayesian learning of the parameter (see e.g., Xie and Carlin (2006)) as one can resolve the issue in a Bayesian model by incorporating prior information. Lindley (1971) first argued that unidentifiability causes no real difficulty in the Bayesian approach as long as one specifies proper priors on the parameters, which is clearly the case in our model. Although there is a concern about “No Bayesian free lunch” by Poirier (1998) in the sense that the prior and posterior distribution might be identical given the prior and the data, our extensive simulation study clearly indicates the contrary. In our simulation study, a vague prior is specified on the ξ and the posterior distribution still clearly concentrates around the true parameter η_0 . Therefore, we argue that the parameters in our model can be adequately learnt.

2.3 Posterior Consistency

A binary GP model admits a much richer latent structure than that of a generalized linear model which only allows linear specifications, thus potentially better approximates the true response function. In this subsection, we will show in theory that under mild assumptions the posterior distribution under our model is consistent with respect to any true response function. More precisely, the posterior mass on any neighborhood of the true response function goes to one as the sample size goes to infinity. Let $p_0(x)$ ($x \in \mathbb{R}^k$) be the true response function and P_0^n be the true distribution of the data. If the covariates x come from a random design, that is, $x \sim Q$ for some distribution Q , we define an ε neighborhood of $p_0(x)$ as $U_\varepsilon(p_0) = \{p: \int |p(x) - p_0(x)| Q(dx) < \varepsilon\}$. If x is from a fixed design, an ε neighborhood of p_0 is given as $U_{\varepsilon,n}(p_0) = \{p: \int |p(x) - p_0(x)| Q_n(dx) < \varepsilon\}$, where Q_n is the empirical distribution on $x_1, \dots, x_n$. Assume the covariates lie in some bounded subset χ of $\mathbb{R}^k$. The following theorem shows that the posterior distribution of $p(x)$ is consistent with respect to the topology defined above. The proof is presented in Web Appendix B.

Theorem 1—Assume the true response function $p_0(x)$ is continuous and bounded away from zero and one. Let Π be the prior distribution of p induced by a Gaussian process prior on w with covariance kernel given in (1) and hyperpriors Π_α and Π_β given as the Inverse-gamma distributions. Π_ξ is a prior on the shape parameter of the link function which is assumed to have a continuous density with respect to the Lebesgue measure on the real line $\mathbb{R}$. Then for x from a random design, one has

$$\Pi(U_\varepsilon(p_0) | ((x_1, y_1), \dots, (x_n, y_n))) \rightarrow 1 \text{ almost surely } P_0^n \text{ as } n \rightarrow \infty.$$

If the covariates x are from a fixed design, one has

$$\prod (U_{\epsilon,n}(p_0)|((x_1, y_1), \dots, (x_n, y_n))) \rightarrow 1 \text{ almost surely } P_0^n \text{ as } n \rightarrow \infty.$$

Remark 1—Note that the assumption on the true response function being bounded away from zero or one is imposed in Theorem 1 although in real applications GEV link allows flatness in the tail with appropriate ξ values. But such assumptions which are used in proving the inequality between the Kullback-Leibler divergence and the super distance as well as in constructing the testing conditions are only sufficient conditions.

3. Priors Specification, Posterior Inference, and Model Comparison Criteria

In this section, we specify priors for parameters θ and ξ , carry out posterior inference, and introduce two model comparison criteria. For posterior inference, details on the sampling algorithm are provided. The prediction of a new binary observation is briefly described.

3.1 Priors Specification

As we mentioned in Section 2.2, the GP hyperparameters are given diffuse inverse-gamma prior distributions. An inverse-gamma prior on a scalar hyperparameter τ has density as $P(\tau, a, b) \propto \tau^{-(a+1)} \exp(-b/\tau) \mathbb{1}_{(0,+\infty)}(\tau)$, where a and b are shape and rate parameters. The prior can be fairly vague by setting values of a and b , for example $a = 2$ and $b = 1$. The GP hyperparameters $\theta = \{\alpha, \{\beta_d\}_{d=1}^k\}$ are assumed to be independent, i.e.,

$P(\theta) = P(\alpha) \prod_{k=1}^d P(\beta_k)$, where each hyperparameter is given an inverse gamma prior with $a = 2$ and $b = 1$. We place a weakly informative normal prior $N(0, 25)$ on the shape parameter ξ .

3.2 Posterior Inference

In a GP binary regression model, the latent variables $\mathbf{w}$ are assumed to be $\mathbf{w}|\theta \sim N(\mathbf{0}, \Sigma_\theta)$. We have joint posterior distribution $P(\mathbf{w}, \theta, \xi|\text{data}) \propto L(\text{data}|\mathbf{w}, \xi) \cdot N(\mathbf{w}; \mathbf{0}, \Sigma_\theta) \cdot P(\theta) \cdot P(\xi)$, where $L(\cdot)$ is the likelihood function, $N(\cdot)$ is the Gaussian process prior on the latent variables $\mathbf{w}$, and $P(\cdot)$ are the priors on θ and ξ .

Using the Metropolis-Hastings (MH) algorithm, samples can be drawn from the joint posterior distribution by alternately fixing the unknowns $\mathbf{w}$, θ , and ξ . However, chains from MCMC sampling by the “standard” MH algorithm are not only computationally expensive but also converge slowly and mix poorly. One reason for such poor performance is that the samples of $\mathbf{w}$ are highly informative about the hyperparameters, especially for short lengthscales. It significantly limits the ability to update θ and ξ with fixed $\mathbf{w}$ for any Markov chain. Another reason is that it is difficult to design proposals that lead to reasonable acceptance rate, especially with high dimensional latent variables. Efficient sampling algorithm for a GP binary regression model is only available for probit links. To improve sampling efficiency, computational simplicity and speed of convergence, we construct an efficient sampling algorithm called the GEV link slice sampling algorithm (SSLS-GEV) extended from the work of Murray and Adams (2010). To our best understanding, our

sampling method is the first successful attempt at sampling a GP binary regression model with a flexible skewed link such as GEV.

Auxiliary model and surrogate data—As defined in Murray and Adams (2010), the auxiliary variables $\mathbf{g} = (g_1, \dots, g_n)'$ is $g_i = w_i + \text{noise}$, where the noise is typically Gaussian. It yields conditional density as $\mathbf{g}|\mathbf{w}, \boldsymbol{\theta} \sim N(\mathbf{w}, S_{\boldsymbol{\theta}})$, where $S_{\boldsymbol{\theta}}$ is the diagonal noise covariance matrix which is often chosen to be $S_{\boldsymbol{\theta}} = \gamma \mathbf{I}$ and γ can be set by hand to a fixed value. The diagonal elements of $S_{\boldsymbol{\theta}}$ can also be obtained by individually matching variance from a Gaussian fit to the posterior of each latent variable. Given $\boldsymbol{\theta}$, the auxiliary values are drawn from the marginal distribution $\mathbf{g}|\boldsymbol{\theta} \sim N(0, \Sigma_{\boldsymbol{\theta}} + S_{\boldsymbol{\theta}})$. The posterior distribution for the latent variables is then $\mathbf{w}|\mathbf{g}, \boldsymbol{\theta} \sim N(\mathbf{m}_{\boldsymbol{\theta}, \mathbf{g}}, \Psi_{\boldsymbol{\theta}})$, where $\Psi_{\boldsymbol{\theta}} = \Sigma_{\boldsymbol{\theta}} - \Sigma_{\boldsymbol{\theta}}(\Sigma_{\boldsymbol{\theta}} + S_{\boldsymbol{\theta}})^{-1} \Sigma_{\boldsymbol{\theta}}$ and $\mathbf{m}_{\boldsymbol{\theta}, \mathbf{g}} = \Psi_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}^{-1} \mathbf{g}$. Using the surrogate data $\mathbf{g}$ allows the proposed draw of the latent variables to be plausible for the proposed hyperparameters while being close to the current state.

Reparameterization—Murray and Adams (2010) also proposed reparameterization in the auxiliary model to further deal with sampling from strongly coupled variables. A Cholesky decomposition gives $\Psi_{\boldsymbol{\theta}} = L_{\Psi_{\boldsymbol{\theta}}} L_{\Psi_{\boldsymbol{\theta}}}'$, where $L_{\Psi_{\boldsymbol{\theta}}}$ is the lower triangular matrix. During sampling, a draw is first obtained from a standard multivariate Gaussian $\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{I})$ and then the latent variables are calculated as $\mathbf{w} = L_{\Psi_{\boldsymbol{\theta}}} \boldsymbol{\eta} + \mathbf{m}_{\boldsymbol{\theta}, \mathbf{g}}$. This reparameterization allows updating $\boldsymbol{\theta}$ for fixed $\boldsymbol{\eta} = L_{\Psi_{\boldsymbol{\theta}}}^{-1}(\mathbf{w} - \mathbf{m}_{\boldsymbol{\theta}, \mathbf{g}})$ instead of fixed $\mathbf{w}$, where $\boldsymbol{\eta}$ is regarded as a random reparameterization and independent of $\boldsymbol{\theta}$. The posterior of hyperparameters $\boldsymbol{\theta}$ and ξ is

$$P(\boldsymbol{\theta}, \xi | \boldsymbol{\eta}, \mathbf{g}, \text{data}) \propto L(\text{data} | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{g}, \xi) \cdot N(\mathbf{g}; \mathbf{0}, \Sigma_{\boldsymbol{\theta}} + S_{\boldsymbol{\theta}}) \cdot P(\boldsymbol{\theta}) \cdot P(\xi). \quad (7)$$

SSLS-GEV sampling algorithm—The efficiency of MH algorithm for the posterior distribution (7) depends crucially on a careful choice of the parameters in the proposal distribution. An improved version of the MH algorithm is slice sampling (Neal, 2003) which is a robust simulation method that works well on a variety of problems and has little need for tuning. Murray and Adams (2010) developed surrogate data slice sampling (SSLS) algorithm. SSLS is an adaptive and robust procedure that has a free parameter σ which is not required to be carefully tuned. Based on surrogate data model and the use of the free-shaped GEV link function, we propose a slice sampling algorithm, named SSLS-GEV, to jointly update strongly-coupled latent variables and hyperparameters $\boldsymbol{\theta}$, as well as the shape parameter ξ . This sampling method is more efficient than the previous MH method used in Wang and Dey (2010). The SSLS-GEV algorithm is described in Web Appendix C.

Prediction—For prediction of y^* at a new point with $\mathbf{x} = \mathbf{x}^*$, the estimated latent variable $w(\mathbf{x}^*)$ need to be estimated. The value of the predictive probability, $\hat{p}(y^* = 1 | \mathbf{x} = \mathbf{x}^*)$, can be calculated based on the estimated link function. A rule of thumb can be applied to obtain the prediction of y^* based on the estimated predictive probability. For example, $y^* = 1$ if $\hat{p}(y^* = 1 | \mathbf{x} = \mathbf{x}^*) \geq 0.5$ or 0 otherwise. Shi and Choi (2011) provided details regarding the prediction

in a GP binary regression model. In Web Appendix F, we show how the prediction can be obtained using MCMC sampling for motivating data example 2.

3.3 Model Comparison Criteria

We use two formal model comparison criteria, namely, the deviance information criterion (DIC) (Spiegelhalter et al., 2002) and log-pseudo marginal likelihood (LPML) based on the conditional predictive ordinates (CPO) (Gelfand et al., 1992). The minimum DIC and maximum LPML indicate the best fit. CSDPbinary function of DPpackage in R computes LPML but not DIC. We also include the predicted error (PE) as an exploratory measures for model comparison. In a sample with size n , if $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_n)'$ is a vector of n predictive probabilities with each element $\hat{p}_i = P(y_i = 1 | \mathbf{x} = \mathbf{x}_i)$ and $\mathbf{y} = (y_1, \dots, y_n)'$ is the vector of observed binary responses, then PE is calculated as $PE = \sum_{i=1}^n (y_i - \hat{p}_i)^2 / n$.

4. Simulation Study

Five candidate models are considered in the simulation study, including the GP binary regression models with the probit link (GP-probit), the GP binary regression models with the cloglog link functions (GP-Cloglog), the GP binary regression models with the GEV link (GP-GEV), the semi-parametric binary regression model adopting Newton et al. (1996)'s centrally standardized Dirichlet process (CSDP) prior for the inverse link function with linear (CSDP-linear) and polynomial (CSDP-polynomial) latent regression functions. The results from these simulated datasets demonstrate how appropriate choice of link affects the fitting performance, as well as how the GEV link provides better flexibility in fitting skewness in the response curve than other links. Moreover, we also examine how the fitting performance scales with the sample size and the number of covariates.

Simulated datasets are generated from each of the five candidate models, respectively. We consider both one and four covariates cases. The true parameters in the models with flexible links, including GP-GEV, CSDP-linear, and CSDP-polynomial, are set such that the proportions of 1's (denoted as % 1) in the simulated data sets are around 30%, 50% and 70%. The covariate $\mathbf{x}_1$ for one covariate case is an equally space sequence from -5 to 5. The data set with four covariates includes $\mathbf{x}_1$, one continuous covariate $\mathbf{x}_2$ generated from a normal distribution with a mean of 0 and a variance of 3, a binary categorical data $\mathbf{x}_3$ generated from a Bernoulli distribution with a probability of success $p = 0.6$, and another continuous covariate $\mathbf{x}_4$ generated from a uniform distribution on the interval [0, 10]. Under each scenario, two sizes of datasets are generated with $n = 200$ or $n = 1000$. Thus, we generate a total of 44 simulated datasets, including two levels of the number of covariates, two levels of sample sizes, and 11 model setups (1 GP-probit, 1 GP-cloglog, 3 GP-GEV, 3 CSDP-linear, and 3 CSDP-polynomial).

The five candidate models are then fitted to each of the 44 simulated datasets. The computations of GP models are implemented in MATLAB software. Computational implementation of the CSDP binary regression model is realized using DPpackage (version 1.1-6) (Jara et al., 2011) in R. This package contains functions to fit the binary regression with nonparametric link in a Bayesian context, e.g., CSDPbinary function, using CSDP prior

and allowing centering the DP around a logit distribution. For each case, we conduct the MCMC simulation for 100000 iterations. The first 50000 draws are discarded as a burn-in phase. Every 10th values are stored to reduce level of correlation between successive values of the simulated chain. The posterior mean and 95% highest posterior density (HPD) interval are calculated with the resulting 5000 values. All results pass the convergence test by the CODA package in R.

Table 1 shows the model comparison results of the 22 simulated data sets with one covariate. When $n = 200$, the GEV model performs better than or as well as other models in all 11 data sets. Even when the true model is the CSDP model (with either linear or polynomial latent regression function), the GEV model has almost the same LPML and PE as the true model. In some cases, the criteria of the cloglog model are very close to the GEV model when the estimation of ξ is very close to 0, which is consistent with the fact that the cloglog link is a special case of the GEV link with $\xi = 0$. When $n = 1000$, the GEV model consistently outperforms the others in all 11 data sets. Plots of the posterior mean of the fitted $\hat{p}_i$ for three simulated examples: GP-probit data ($n = 1000$, % 1 = 54.8%), GP-GEV data ($n = 1000$, % 1 = 76%) and CSDP-polynomial data ($n = 1000$, % 1 = 74.2%) are provided in Web Figures 1-3. As shown in these figures, the curve of the posterior mean of predictive probabilities under the GEV link fits well to the data (closer to the true response curve). In the cases of four covariates, although the GEV model still performs the best when the data is simulated from the GP binary model with different links, the CSDP model often performs better than the GEV model when the data are simulated from the CSDP model. The model comparison results for simulated datasets with four covariates are provided in Web Tables 1 and 2.

The GEV model can approximate a symmetric or asymmetric link model and provides great flexibility in fitting skewness in the response curve. In the simulated example of the GP-cloglog data ($n = 200$, % 1 = 67%), we obtain $\hat{\xi} = 0.46$ with 95% HPD interval for ξ as $[-0.20, 1.24]$ for the GEV model, which indicates that ξ is significantly larger than $\log 2 - 1 \approx -0.3069$ (positively skewed) and zero is contained in the interval. The values of DIC and LPML of the GEV model and the cloglog model are quite similar. Both positively skewed link models are better than the symmetric probit link model. When the data has higher skewness in the response curve, e.g., the GP-GEV data ($n = 200$, % 1 = 75.5%, $\xi = 2$), the GEV model exhibits benefits with the lowest DIC and largest LPML. The data with significantly different numbers of 0's and 1's is not uncommon in many practical problems, since the measurement may be a "rare event" that happens with a small probability. Such an unbalance is an indicator of the possible skewness in the response curve. Wang and Dey (2010) pointed out the choice on the direction of the skewness is more important on the choice between symmetric or asymmetric link. The wrong choice of direction gives an even worse fit than a wrong choice between symmetric and asymmetric links. The flexibility of the GEV link models is apparent because it allows the model to detect the skewness in the data instead of presuming any direction. This is important in practice. Reversing the coding of "0" and "1" does not affect the fitness of the symmetric links, but it changes the asymmetry in the data. Description and results of comparison for "flipped" data are presented in Web Appendix D, Web Table 3 and Web Figure 4. The performance of the positively skewed cloglog link depends on how the data is coded and deteriorates simply by flipping "0" and "1" in a dataset previously requiring a positively skewed link. The

performance of the GP-GEV model is not affected by this arbitrary change in the data and still outperforms the other models.

Choice of a diffuse prior on ξ tends to increase the chance that ξ is poorly learnt especially in complex models with small sample size. Elicitation of more informative prior may assist Bayesian learning, see Dey and Liu (2007). In our simulation studies, ξ can be adequately learnt by giving it a weakly informative normal prior ($N(0, 25)$). The true value of ξ for each GEV model is contained in the 95% HPD interval. With the same number of covariates and increasing the sample size, the variance of ξ decreases, which implies that more information provided by the data helps obtain a more precise estimator of ξ . With the same sample size, the variance of ξ increases when more covariates are in the model. The impacts of sample size and number of covariates on fitting the GEV model are reflected by model selection criteria. It can be better learnt with larger sample sizes or smaller number of covariates.

5. Data Examples

We now apply our model to the two motivating data examples presented in Introduction.

5.1 Example 1: Experiment on Attention Paradigm

This example studies the attention paradigm described in Smith et al. (2009) to which we refer the readers for more details of the experiment. The goal is to determine whether deep brain stimulation allows the monkey to recover its pre-fatigue level of performance once performance has diminished as a result of spontaneous fatigue. The monkey performed 1250 trials. Stimulation was applied during four periods across trials 300-364, 498-598, 700-799 and 1000-1099, indicated by shaded gray regions in Figure 3. A total of 741 (59.28%) out of the 1250 observations are the correct responses. This is an example of modeling binary time series responses. To improve the sampling performance and convergence, we center and scale the covariate $\mathbf{x}$ to an equally spaced sequence from -6.24 to 6.25 with an increment of 0.01.

Table 2 (Example 1: Experiment on Attention Paradigm) shows the posterior estimation and model comparison results of the five models. Note that the parameters of the CSDP models include precision parameter a and regression parameters $\mathbf{b}$ (Newton et al., 1996). For the GEV model, the posterior mean and 95% HPD interval of ξ are -0.8748 and [-0.9995, -0.6889] respectively, indicating that the response function is negatively skewed. The model comparison criteria select the GEV model as the best model. The cloglog model fits the data poorly as expected, since the cloglog link is positively skewed.

In Figure 3(a), we plot the posterior mean of the latent variables $\mathbf{w}$ for the three fitted GP models. In all the GP models, the estimates follow a similar pattern, but expressive differences between the estimates are found in the last OFF-stimulation period. Figure 3(b) plots the posterior mean of the probability of a correct response for the GP models and the CSDP-polynomial model. The GP model response curves all lie around 0.7 and decrease significantly in the middle of the fourth OFF-stimulation period around trial 900, and at the end of the final ON-stimulation period around trial 1100, especially with the GP-GEV model. The results show a significant increase in performance during the final ON-

stimulation period from trial 1000 with the proportion of 1's increased from 30% to 57.5%. Then there is a dramatic drop in the performance with 0% of 1's in the final OFF-stimulation period (from around trial 1200 to the end). For all the GP models, the results indicate the stimulation has a positive effect on performance, except during the first stimulation period. These results are consistent with Smith et al. (2009). The best CSDP-polynomial model we selected to fit the data is with $w(x) = \beta_0 + \sum_{j=1}^3 \beta_j x^j$. However, its corresponding response curve does not seem to catch the upward and downward trends as the GP models, in particular fails to exhibit the dramatic fluctuates from around trial 900 to the end of the experiment. Abanto-Valle et al. (2015) analyzed this data using a Bayesian state-space mixed model (BSSM) under different link functions, including the probit, the logistic, and the GEV links. The latent state is assumed to follow a AR(1) process. The BSSM-GEV provides the best fit with DIC = 1427, which is much larger than DIC = 1415.3 of our GP-GEV model. It implies that along with a flexible parametric link function, the nonparametric latent regression function based on a GP prior can provide better fit to the data. For the binary time series, it is often of interested to compute the joint distribution of $Y_{t-p}, \dots, Y_{t-1}, Y_t$, given the parameters in the model, for some integer p (Di Lucca et al., 2013). A discussion is included in Web Appendix E.

5.2 Data Example 2: Pneumoconiosis Field Research

Here we apply the GP-GEV model to data from the field of research on the causes of pneumoconiosis among coal miners (Ashford and Smith, 1964) as presented in Introduction. Out of 916 observations, 303 (or 33.08%) were observed to have pneumoconiosis.

We make the bubble plot as shown in Figure 1(b) after classifying the workers into 40 groups based on durations of two types of work. For each individual, the x- and y-axis respectively represent the duration of working as a coal-getter on the coal surface and as a haulage-worker in the underground roadway. Each proportionally sized bubble indicates the proportion of observed pneumoconiosis in that group. We can see the proportions vary with different combinations of durations.

Posterior estimation and model comparison results are shown in Table 2 (Example 2: Pneumoconiosis Field Research). The GEV model performs the best with the largest LPML and the lowest DIC and PE. The posterior mean of the shape parameter $\hat{\xi} = -0.3622$ with 95% HPD interval $[-0.6033, -0.1474]$.

Prediction in a GP regression model is conceptually straightforward (Rasmussen and Williams, 2006; Shi and Choi, 2011). We show the prediction of predictive probabilities on a 20 by 20 testing data set (see Web Appendix F for the method of prediction). Figure 4 are the image and contour plots of the posterior mean of predictive probabilities, as well as the image plots of the posterior standard errors of predictive probabilities.

From the image and contour plots, we can visualize that the overall trend of the estimated predictive probabilities increases with the incremental number of years the worker spent on the coalface. The patterns of the subregions can also be locally captured in details. Workers who worked both underground for less than 15 years and on the coalface for less than 10

years only have the value of predictive probability less than around 0.3. And the probability is less than 0.35 if workers worked as coal-getters for no more than 5 years, no matter how long they worked as haulage-workers. If they worked on the coalface for 5 to 20 years as well as underground for more than 20 years, their probability of having pneumoconiosis is larger than 0.5. The probability is also likely to be greater than 0.5 if workers worked on coalface for more than 20 years, no matter for how many years they worked as haulage-workers. The plots of the posterior standard errors of predictive probabilities show the northeast has large standard errors of predictions, since the northeast area lacks observations. Standard errors are quite small in the southwest area, where the training data are mainly located.

6. Conclusion and Discussion

We introduced a nonparametric binary regression model that incorporates a GP prior through a flexible skewed link function in modeling binary response data. Comparing with standard link functions, the GEV link model is attractive because of its flexibility in modeling positively skewed, symmetric or negatively skewed scenarios. Efficient sampling algorithms were developed which are valuable additions to current sampling algorithms that only apply to symmetric links. Empirical results from simulation studies and real examples show that our proposed model outperforms other models in terms of model fitting, predictive power and flexibility. It is possible that other flexible link functions, such as CSDP in Newton et al. (1996), can outperform the proposed GP-GEV when they are coupled with a richer class of basis functions as the latent regression function. However, it lacks the simplicity in the model structure, parameter interpretation, and computational tractability as in the GP-GEV model. Our results from the simulated datasets show that either the GP-GEV model or the CDSP model can have superior performance depending on the actual structure in a dataset.

In this work, stepwise slice samplers are used, which are known to be more efficient than standard MCMC algorithms such as Metropolis-Hastings with a random walk proposal. We employed methods that based on sparse matrices to reduce the computational burden of evaluating the operations of matrices in the algorithm, such as inverse of large covariance matrix or multiplication of matrices. One possible extension to our sampling method might be exploring parallel multivariate slice sampling which can offer potential gains in efficiency over stepwise slice sampler. Furthermore, it is well-known that the GP regression model faces the computation challenges as it typically scales as $O(n^3)$. Chapter 8 of Rasmussen and Williams (2006) provides a comprehensive review on approximation methods for large datasets. It is possible to implement approximation methods in the GP-GEV model when dealing with “big” data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by an allocation of computing time from the Ohio Supercomputer Center. Lizhen Lin was partially supported by R01-ES017240, a grant from the National Institute of Health. The authors are

grateful to the Editor, the Associate Editor and the reviewers for their helpful and constructive comments that have led to substantial improvement on this manuscript.

References

- Abanto-Valle CA, Dey DK, Jiang X. Binary state space mixed models with flexible link functions: a case study on deep brain stimulation on attention reaction time. *Statistics and Its Interface*. 2015; 8:187–194.
- Ashford JR, Smith CS. General models for quantal response to the joint action of a mixture of drugs. *Biometrika*. 1964; 51:413–428.
- Basu S, Mukhopadhyay S. Bayesian analysis of binary regression using symmetric and asymmetric links. *Sankhyā : The Indian Journal of Statistics, Series B*. 2000; 62:372–387.
- Chen MH, Dey DK, Shao QM. A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*. 1999; 94:1172–1186.
- Choudhuri N, Ghosal S, Roy A. Nonparametric binary regression using a Gaussian process prior. *Statistical Methodology*. 2007; 4:227–243.
- Czado C, Santner TJ. The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*. 1992; 33:213–231.
- Dey DK, Liu J. A quantitative study of quantile based direct prior elicitation from expert opinion. *Bayesian Analysis*. 2007; 2:137–166.
- DeYoreo M, Kottas A. A fully nonparametric modeling approach to binary regression. *Bayesian Analysis*. 2015; 10:821–847.
- Di Lucca MA, Guglielmi A, Müller P, Quintana FA. A simple class of Bayesian nonparametric autoregression models. *Bayesian Analysis*. 2013; 8:63–88. [PubMed: 26052373]
- Dunson, DB.; Bhattacharya, A. Nonparametric Bayes regression and classification through mixtures of product kernels. In: Bernardo, JM.; Bayarri, MJ.; Berger, JO.; Dawid, AP.; Heckerman, D.; Smith, AFM.; West, M., editors. *Bayesian Statistics. Vol. 9*. Oxford University Press; 2011. p. 145–164.
- Fay JWJ. The National Coal Board's Pneumoconiosis Field Research. *Nature*. 1957; 180:309–311. [PubMed: 13464825]
- Frölich M. Non-parametric regression for binary dependent variables. *Econometrics Journal*. 2006; 9:511–540.
- Gelfand, AE.; Dey, DK.; Chang, H. Model determination using predictive distributions with implementation via sampling-based methods. In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. *Bayesian Statistics. Vol. 4*. Oxford University Press; 1992. p. 147–167.
- Hannah LA, Blei DM, Powell WB. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*. 2011; 12:1923–1953.
- Iwata T, Duvenaud D, Ghahramani Z. Warped mixtures for nonparametric cluster shapes. *arXiv: 1206.1846v2 [stat ML]*. 2013
- Jara A, Hanson TE, Quintana FA, Müller P, Rosner GL. DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*. 2011; 40:1–30. [PubMed: 21796263]
- Kadane, JB. The role of identification in Bayesian theory. In: Fienberg, SE.; Zellner, A., editors. *Studies in Bayesian Econometrics and Statistics*. North-Holland Amsterdam; 1974. p. 175–191.
- Kim S, Chen MH, Dey DK. Flexible generalized t-link models for binary response data. *Biometrika*. 2008; 95:93–106.
- Lindley, DV. *Bayesian Statistics: A Review*. SIAM; Philadelphia: 1971.
- Murray I, Adams RP. Slice sampling covariance hyperparameters of latent Gaussian models. *Advances in Neural Information Processing Systems*. 2010; 23:1723–1731.
- Neal R. Regression and classification using Gaussian process priors. *Bayesian Statistics*. 1999; 6:475–501.
- Neal RM. Slice sampling. *The Annals of Statistics*. 2003; 31:705–767.
- Newton MA, Czado C, Chappell R. Bayesian inference for semiparametric binary regression. *Journal of the American Statistical Association*. 1996; 91:142–153.

- Northrop PJ, Attalides N. Posterior propriety in Bayesian extreme value analyses using reference priors. arXiv:1505.04983v3 [stat ME]. 2015
- Poirier DJ. Revising beliefs in nonidentified models. *Econometric Theory*. 1998; 14:483–509.
- Rasmussen, CE. Technical report. University of Toronto; 1997. Evaluation of Gaussian processes and other methods for non-linear regression.
- Rasmussen, CE.; Williams, CKI. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press; 2006.
- Robert, CP.; Casella, G. *Monte Carlo Statistical Methods*. Springer-Verlag New York, Inc; 2005.
- Shahbaba B, Neal RM. Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*. 2009; 10:1829–1850.
- Shi, JQ.; Choi, T. *Gaussian Process Regression Analysis for Functional Data*. CRC Press; Boca Raton, FL; 2011.
- Smith AC, Shah SA, Hudson AE, Purpura KP, Victor JD, Brown EN, Schiff ND. A Bayesian statistical analysis of behavioral facilitation associated with deep brain stimulation. *Journal of Neuroscience Methods*. 2009; 183:267–276. [PubMed: 19576932]
- Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2002; 64:583–639.
- Wang X, Dey DK. Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *The Annals of Applied Statistics*. 2010; 4:2000–2023.
- Xie Y, Carlin BP. Measures of Bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*. 2006; 136:3458–3477.

(a)

(b)

Figure 1.

Two motivating data examples. Panel (a) Experiment on attention paradigm. Circles at 0 and 1 represent the observed binary responses. The solid dots connected by the solid line are proportions of 1's in every 40 consecutive trials. Panel(b) Pneumoconiosis field research. The size of the circles is scaled according to the proportions of men having pneumoconiosis at different length of employment as coal-getters or haulage-workers. The larger the circle, the higher the proportion. This figure appears in color in the electronic version of this article.

Figure 2.

The response probability functions under the probit, cloglog and GEV ($\xi = -1$ and 1) links in the GP binary regression model with $\alpha = 1$ and $\beta = 0.25, 4$ and 16 , respectively. This figure appears in color in the electronic version of this article.

Figure 3.

Estimation results for the experiment on attention paradigm example. Panel (a): Posterior mean of latent variable $\mathbf{w}$ of GP models under probit, cloglog, and GEV links; Panel (b): Posterior mean of predictive probabilities of GP models and CSDP-polynomial model. This figure appears in color in the electronic version of this article.

Figure 4.

Pneumoconiosis Field Research. Image plot of the posterior mean of predictive probabilities [left], image plot of the posterior standard errors of predictive probabilities [middle] and contour plot of the posterior mean of predictive probabilities [right] for the GP models under (a) probit link; (b) cloglog link; (c) GEV link. This figure appears in color in the electronic version of this article.

Table 1

Comparison of GP-probit, GP-cloglog, GP-GEV, CSDP-linear and CSDP-polynomial model for different simulated data sets with 1 covariate. The column % 1 represents the proportion of 1's in the simulated data set.

Model		Sample size $n = 200$						Sample size $n = 1000$					
		%1	probit	cloglog	GEV	CSDP-linear	CSDP-poly	%1	probit	cloglog	GEV	CSDP-linear	CSDP-poly
Fitted True													
True													
probit	DIC		140.83	142.57	138.01	—	—		135.95	139.91	131.70	—	—
	LPML		-69.64	-70.16	-68.42	-126.55	-84.35		-67.22	-69.38	-65.33	-408.25	-125.43
	PE	50%	0.102	0.103	0.102	0.207	0.134	54.8%	0.019	0.020	0.019	0.125	0.026
	ξ				-0.36 [-0.81, 0.19]						-0.56 [-0.86, -0.23]		
cloglog	DIC		120.11	116.62	118.01	—	—		173.14	161.01	163.74	—	—
	LPML		-58.72	-57.33	-58.24	-110.87	-73.52		-87.18	-80.36	-81.38	-361.31	-158.53
	PE	67%	0.077	0.075	0.078	0.167	0.091	56.6%	0.023	0.023	0.023	0.120	0.029
	ξ				0.46 [-0.20, 1.24]						-0.04 [-0.18, 0.10]		
GEV													
$\xi = -2$	DIC		89.71	111.71	77.50	—	—		482.47	742.86	366.29	—	—
	LPML		-43.36	-55.20	-37.46	-103.89	-78.33		-237.12	-371.77	-181.73	-407.02	-315.30
	PE	32%	0.055	0.064	0.051	0.157	0.124	22.5%	0.065	0.094	0.059	0.120	0.088
	ξ				-2.10 [-3.65, -0.97]						-2.00 [-2.51, -1.51]		
$\xi = 0.5$	DIC		173.55	162.75	164.72	—	—		1009.40	969.32	967.13	—	—
	LPML		-87.16	-80.83	-82.04	-139.11	-106.36		-504.28	-484.23	-483.50	-687.01	-572.80
	PE	54.5%	0.111	0.111	0.115	0.229	0.171	54%	0.162	0.159	0.159	0.238	0.180
	ξ				0.50 [-0.09, 1.16]						0.33 [0.03, 0.61]		
$\xi = 2$	DIC		134.39	126.50	117.76	—	—		575.24	559.48	547.88	—	—
	LPML		-65.72	-61.98	-58.02	-104.11	-62.57		-286.20	-279.31	-273.56	-482.34	-307.72
	PE	75.5%	0.094	0.092	0.096	0.154	0.092	76%	0.096	0.095	0.095	0.157	0.100
	ξ				2.86 [0.74, 5.43]						1.65 [0.81, 2.53]		
CSDP-linear													
DIC		28%	128.28	129.13	126.35	—	—	27.5%	1143.30	1162.50	1136.50	—	—

Model	Sample size $n = 200$										Sample size $n = 1000$									
	Fitted	True	%1	probit	cloglog	GEV	CSDP-linear	CSDP-poly	%1	probit	cloglog	GEV	CSDP-linear	CSDP-poly						
CSDP-polynomial	LPML			-63.42	-64.54	-62.47	-61.15	-60.99		-571.39	-580.76	-568.04	-568.13	-568.89						
	PE			0.092	0.093	0.092	0.090	0.090		0.188	0.192	0.188	0.188	0.188						
	ξ					-0.64 [-1.08, -0.28]						-1.11 [-1.44, -0.80]								
	DIC			122.17	122.24	122.00	—	—		491.06	503.80	488.00	—	—						
	LPML	56%		-60.65	-60.47	-60.41	-59.74	-59.57	48.3%	-245.31	-253.07	-247.02	-250.93	-251.64						
	PE			0.087	0.087	0.087	0.084	0.084		0.075	0.075	0.074	0.073	0.073						
	ξ					-0.12 [-0.66, 0.36]						-0.54 [-0.93, -0.04]								
	DIC			176.83	177.50	174.51	—	—		1203.60	1205.40	1201.20	—	—						
	LPML	77%		-88.48	-89.01	-84.92	-93.85	-93.28	76.4%	-601.88	-602.73	-600.72	-603.71	-603.46						
	PE			0.130	0.128	0.126	0.133	0.131		0.202	0.202	0.202	0.202	0.203						
	ξ					1.80 [-0.47, 5.26]						0.30 [-0.61, 1.18]								
CSDP-polynomial	DIC			50.63	54.81	43.65	—	—		377.89	396.18	330.30	—	—						
	LPML	18.5%		-23.72	-26.84	-22.61	-59.81	-22.15	27.8%	-188.16	-200.92	-164.39	-521.12	-170.36						
	PE			0.028	0.028	0.029	0.063	0.026		0.051	0.058	0.050	0.189	0.049						
	ξ					-1.61 [-3.03, -0.61]						-1.33 [-1.70, -1.02]								
	DIC			103.49	104.20	100.06	—	—		347.81	398.08	339.52	—	—						
	LPML	56%		-52.01	-52.55	-50.41	-59.74	-53.31	47.8%	-173.17	-199.92	-169.28	-418.77	-177.73						
	PE			0.068	0.068	0.072	0.084	0.069		0.051	0.053	0.051	0.140	0.051						
	ξ					-1.96 [-3.25, -0.88]						-0.59 [-0.75, -0.45]								
	DIC			197.28	198.14	198.50	—	—		827.29	813.75	790.65	—	—						
	LPML	78.5%		-98.58	-99.07	-99.08	-104.11	-100.64	74.2%	-412.44	-405.86	-395.64	-527.59	-404.31						
	PE			0.152	0.151	0.151	0.161	0.153		0.135	0.132	0.132	0.182	0.129						
ξ					0.02 [-1.76, 2.23]						1.78 [1.20, 2.60]									

Table 2
Comparison of the GP-probit, GP-cloglog, GP-GEV, CSDP-linear and CSDP-polynomial model on two motivating data examples

Example 1: Experiment on Attention Paradigm						
	GP-Probit model		GP-Cloglog model		GP-GEV model	
	mean	95% HPD	mean	95% HPD	mean	95% HPD
α	0.6838	[0.2873, 1.2631]	0.8524	[0.3087, 1.5578]	0.4009	[0.1473, 0.7315]
β	0.7947	[0.3789, 1.2660]	0.8135	[0.3976, 1.2956]	0.7504	[0.3059, 1.2554]
ξ	—	—	—	—	-0.8748	[-0.9995, -0.6889]
DIC		1424.2		1434.8		1415.3
LPML		-711.7211		-717.2409		-707.2327
PE		0.1905		0.1910		0.1903

CSDP-linear model				CSDP-polynomial model	
	95% HPD		mean	95% HPD	
	mean	95% HPD		mean	95% HPD
a	1.4722	[0.7766, 2.1291]	1.3393	[0.5608, 2.0028]	
$b_0: 1$	2.3382	[1.2546, 3.3586]	1.2304	[0.5151, 1.8701]	
$b_1: x$	-0.6940	[-0.9987, -0.3937]	0.0992	[-0.1094, 0.2557]	
$b_2: x^2$	—	—	-0.0487	[-0.0803, -0.0181]	
$b_3: x^3$	—	—	-0.0132	[-0.0219, -0.0054]	
DIC		—		—	
LPML		-731.0377		-720.8887	
PE		0.1981		0.1945	

Example 2: Pneumoconiosis Field Research						
	GP-Probit model		GP-Cloglog model		GP-GEV model	
	mean	95% HPD	mean	95% HPD	mean	95% HPD
α	0.5834	[0.1958, 1.1596]	1.3275	[0.4298, 2.6957]	0.7085	[0.2053, 1.4493]
β_1	1.5281	[0.1951, 3.6382]	1.6403	[0.2726, 3.6309]	1.8688	[0.2370, 4.3983]
β_2	6.7744	[0.3878, 19.1160]	6.2836	[0.3869, 15.9535]	8.3987	[0.4668, 22.1415]
ξ	—	—	—	—	-0.3622	[-0.6033, -0.1474]

Example 1: Experiment on Attention Paradigm						
	GP-Probit model		GP-Cloglog model		GP-GEV model	
	mean	95% HPD	mean	95% HPD	mean	95% HPD
DIC		958.7876		962.6674		954.7931
LPML		-479.5721		-481.6969		-477.6529
PE		0.1708		0.1717		0.1695

CSDP-linear model				CSDP-polynomial model			
	mean		95% HPD		mean		95% HPD
a	1.3367		[0.7098, 1.9808]	1.1044		[0.3731, 1.7873]	
$b_0: 1$	-1.9940		[-2.7110, -1.2950]	-3.9280		[-5.8000, -2.3240]	
$b_1: x_1$	0.0951		[0.0640, 0.1304]	0.2812		[0.1428, 0.4501]	
$b_2: x_2$	0.0276		[0.0142, 0.0440]	0.1385		[0.0302, 0.2560]	
$b_3: x_1^2$	—		—	-0.0041		[-0.0076, -0.0010]	
$b_4: x_2^2$	—		—	-0.0017		[-0.0035, 0.0005]	
$b_5: x_1 x_2$	—		—	-0.0041		[-0.0087, 0.0004]	
DIC			—			—	
LPML			-494.0234			-494.4930	
PE			0.1750			0.1738	