Variational Inference on Infinite Mixtures of Inverse Gaussian, Multinomial Probit and Exponential Regression

Minhazul Islam Sk and Arunava Banerjee
School of Computer and Information Science and Engineering, University of Florida
Email: {smislam, arunava}@cise.ufl.edu

***Abstract*—We introduce a new class of methods and inference techniques for infinite mixtures of Inverse Gaussian, Multinomial Probit and Exponential Regression, models that belong to the widely applicable framework of Generalized Linear Model (GLM). We characterize the joint distribution of the response and covariates via a Stick-Breaking Prior. This leads to, in the various cases, nonparametric models for an infinite mixture of Inverse Gaussian, Multinomial Probit and Exponential Regression. Estimates of the localized mean function which maps the covariates to the response are presented. We prove the weak consistency for the posterior distribution of the Exponential model (SB-EX) and then propose mean field variational inference algorithms for the Inverse Gaussian, Multinomial Probit and Exponential Regression. Finally, we demonstrate their superior accuracy in comparison to several other regression models such as, Gaussian Process Regression, Dirichlet Process Regression, etc.**

***Keywords*—Probit Regression, Exponential Regression, Inverse Gaussian Regression, Dirichlet Process and Variational Inference.**

I. Introduction

Inverse Gaussian, Exponential, and Multinomial Probit Regression belong to a unified framework called the Generalized Linear Model (GLM) [1]. Regression, in its canonical form, assumes that the response variable follows a given probability distribution with its support determined by a linear combination of the covariates. Formally stated, $Y|\mathbf{X} \sim f\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta}\right)$. There are two aspects to this equation that GLM generalizes. Firstly, f is generalized to the exponential family. f, in the case of Inverse Gaussian regression, is the Inverse-Gaussian distribution, and in the case of Exponential and Multinomial Probit Regression, is the Exponential and Multinomial distributions respectively. Secondly, the function that maps the response mean (μ) to $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta}$, which in the case of Multinomial Probit regression is the probit function $\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta} = g\left(\mu\right) = \Phi^{-1}\left(\mu\right)\right)$, ($\Phi$ is the Normal CDF) is generalized to one of any member of a set of link functions. Link functions for Inverse Gaussian regression is $g\left(\mu\right) = -\mu^{-2}$ and for Exponential regression is $g\left(\mu\right) = -\mu^{-1}$.

Notwithstanding its generality, all three regression models suffer from two intrinsic weaknesses. Firstly, the covariates are associated with the model via only a linear function. Secondly, the variance of the responses are not associated with the individual covariates. This is not desirable because it is conceivable that the response depend non-linearly on the covariates, and furthermore, the variance vary with the values of the covariates [3].

We overcome these limitations by introducing a mixture of regression model where each mixture component is capable of localizing itself to covariate ranges that shows similar responses. Although each mixture component is a localized regression model (therefore a linear combination of covariates), marginalizing out the variance of the local densities creates a non-linear regression model. It also models heterosecdasticity, where the response variance changes across local densities and therefore also varies across the covariates. Furthermore, in order to allow the data to choose the number of clusters non-parametrically [5] we impose a Stick-Breaking prior [4] in the form of a Dirichlet Process [2]. We refer to the infinite mixture of Inverse Gaussian as SB-IG, the infinite mixture of Multinomial Probit as SB-PBT and the infinite mixture of Exponential as SB-EX.

We prove the weak consistency for the SB-EX model which acts as a frequentist justification for Bayesian Methods: as more observations arrive, the posterior distribution converges to the true density of the response-covariate pair. Weak consistency of the other two models are almost the same, therefore we have not presented it due to lack of space.

For inference with SB-EX, SB-PBT and SB-IG, a widely used MCMC algorithm, namely Gibbs sampling [7]-[8] is an immediate choice. However, the inherent deficiencies of Gibbs sampling significantly reduces its practical utility. As is well known, Gibbs sampling approximates the original posterior distribution by sampling using a Markov Chain. However, Gibbs sampling is prohibitively slow and moreover, its convergence is very difficult to diagnose. In high dimensional regression problems, Gibbs Sampling seldom converges to the target posterior distribution in suitable time, leading to significantly poor density estimation and prediction [9]. To alleviate these problems, we introduce a fast and deterministic mean field variational inference algorithm [10], [11], [12], [13] for superior prediction and density estimation. Variational inference is deterministic and possesses an optimization criterion which can be used to assess convergence.

We derive the variational inference separately for the SB-EX, SB-IG and SB-PBT models. These models differ significantly in terms of the type of covariate and response data, which results in markedly different variational distributions, parameter estimations and predictive distributions. In each case, we formulate a class of decoupled and factorized variational distributions as surrogates for the true posterior distribution. We then maximize the lower bound (resulting

from imposing Jensen's inequality on the log likelihood) to obtain the optimal variational parameters. Finally, we derive the predictive distribution from the posterior approximation to predict the response variable conditioned on a new covariate and the past response-covariate pairs.

We demonstrate the accuracy of the variational approach for SB-EX and SB-IG models across different metrics such as relative mean square and absolute error, in high dimensional problems against Linear Regression, Bayesian and variational Linear Regression, Gaussian Process Regression, and ordinary Dirichlet Process Regression. We test the SB-PBT model against Multiclass SVM, Multinomial Logistic Regression and Naive Bayes Model.

The remainder of the paper is organized as follows. The next section describes the Dirichlet Process and its Stick-breaking Representation. Section 3 presents the SB-EX, SB-IG and SB-PBT models as probabilistic graphical models. In section 4, we prove the weak consistency of the SB-EX model. In section 5, we formulate the variational distributions of the models. Section 6 is devoted to the estimation of the parameters of the variational distributions. Experimental results are reported in section 7. Finally, section 8 presents concluding remarks.

II. Dirichlet Process and its Stick-breaking Representation

A Dirichlet Process [2], $DP(\alpha_0, G_0)$ is defined as a probability distribution over a sample space of probability distributions,

$$G \sim DP(\alpha_0, G_0) \tag{1}$$

Here, α_0 is a concentration parameter and G_0 is the base distribution. According to the stick-breaking construction [4] of DP, G, which is a sample from DP, is an atomic distribution with countably infinite atoms drawn from G_0.

$$\begin{aligned} v_i|\alpha_0, G_0 \sim Beta(1, \alpha_0), \quad \theta_i|\alpha_0, G_0 \sim G_0 \\ M_i = v_i \prod_{l=1}^{i-1} (1 - v_l), \quad G = \sum_{i=1}^{\infty} M_i.\delta_{\theta_i} \end{aligned} \tag{2}$$

In the DP mixture model, DP [5], [6] is used as a non-parametric prior over parameters of an infinite mixture model [7].

$$\begin{aligned} z_n|\{v_1, v_2, ...\} \sim Categorical\{M_1, M_2, M_3....\} \\ X_n|z_n, (\theta_i)_{i=1}^{\infty} \sim F(\theta_{z_n}) \end{aligned} \tag{3}$$

Here, F is a distribution parametrized by θ_{z_n}.

III. SB-IG, SB-EX and SB-PBT Models as Probabilistic Graphical Models

We begin by viewing the continuous covariate-response pairs in the SB-EX, SB-PBT, and SB-IG models through the lens of probabilistic graphical models according to their stick breaking representation.

A. Exponential Model (SB-EX)

In SB-EX model, the generative model of the covariate-response pair is given by the following set of equations.

$$\begin{aligned} v_i|\alpha_1, \alpha_2 \sim Beta(\alpha_1, \alpha_2) \\ \{\lambda_{x,i,d}\} \sim Gamma(\lambda_{x,i,d}|a_x, b_x) \\ \{\beta_{i,d}\} \sim Gamma(\beta_{i,d}|c_{y,d}, b_{y,d}) \\ z_n|\{v_1, v_2, ...\} \sim Categorical\{M_1, M_2, M_3....\} \\ X_{n,d}|z_n \sim Exp(X_{n,d}|\lambda_{x,z_n,d}) \\ Y_n|X_n, z_n \sim Exp\left(Y_n|\beta_{z_n,0} + \sum_{d=1}^{D} \beta_{z_n,d} X_{n,d}\right) \end{aligned} \tag{4}$$

Here, X_n and Y_n represents the continuous response-covariate pairs. $\{z, v, \lambda_{x,i,d}, \beta_{i,d}\}$ is the set of latent variables and the distributions, $\{\lambda_{x,i,d}\}$ and $\{\beta_{i,d}\}$ are the base distributions of the DP.

B. Inverse Gaussian Model (SB-IG)

In the SB-IG model, the covariate and the response is modeled by Inverse Gaussian distribution. Here, too v_i and z_n follow the same distributions as before. The remainder of the generative model is given by,

$$\begin{aligned} \{\mu_{i,d}, \lambda_{x,i,d}\} \sim \mathcal{N}\left(\mu_{i,d}|a_{x,d}, (b_{x,d}, \lambda_{x,i,d})^{-1}\right) \\ Gamma(\lambda_{x,i,d}|c_{x,d}, d_{x,d}) \\ \{\beta_{i,d}, \lambda_{y,i}\} \sim \mathcal{N}\left(\beta_{i,d}|a_{y,d}, (b_y, \lambda_{y,i})^{-1}\right) \\ Gamma(\lambda_{y,i}|c_y, d_y) \\ X_{n,d}|z_n \sim IG(X_{n,d}|\mu_{z_n,d}, \lambda_{x,z_n,d}) \\ Y_n|X_n, z_n \sim IG\left(Y_n|\beta_{z_n,0} + \sum_{d=1}^{D} \beta_{z_n,d} X_{n,d}, \lambda_{y,z_n}\right) \end{aligned} \tag{5}$$

Here, X_n and Y_n represents the continuous response-covariate pairs. $\{z, v, \mu_{i,d}, \lambda_{x,i,d}, \beta_{i,d}, \lambda_{y,i}\}$ is the set of latent variables and the distributions, $\{\mu_{i,d}, \lambda_{x,i,d}\}$ and $\{\beta_{i,d}, \lambda_{y,i}\}$ are the base distributions of the DP.

C. Multinomial Probit Model (SB-PBT)

In the Multinomial Probit model, the continuous covariates are modeled by a Gaussian mixture and a Multinomial Probit framework is used for the categorical response. Here, too v_i and z_n follow the same distributions as before. The remainder of the generative model of the covariate-response pair is given by the following set of equations.

$$\begin{aligned} \{\mu_{i,d}, \lambda_{x,i,d}\} \sim \mathcal{N}\left(\mu_{i,d}|a_{x,d}, (b_{x,d}, \lambda_{x,i,d})^{-1}\right) \\ Gamma(\lambda_{x,i,d}|c_{x,d}, d_{x,d}) \\ X_{n,d}|z_n \sim \mathcal{N}\left(X_{n,d}|\mu_{z_n,d}, \lambda_{x,z_n,d}^{-1}\right) \\ \beta_{i,d,k} \sim \mathcal{N}\left(\beta_{i,d,k}|m_{y,d,k}, s_{y,d,k}^2\right) \\ \lambda_{y,i,k} \sim Gamma(\lambda_{y,i,k}|a_{y,k}, b_{y,k}) \\ Y_{n,k,i}^*|X_n, z_n \sim \mathcal{N}\left(Y_n|\beta_{i,0,k} + \sum_{d=1}^{D} \beta_{i,d,k} X_{n,d}, \lambda_{y,i,k}^{-1}\right) \\ Y_n|Y_{n,k,z_n}^* \sim \frac{Y_{n,k,z_n}^*}{\sum_{k=1}^{K} Y_{n,k,z_n}^*} \end{aligned} \tag{6}$$

Here, $\left\{z, v, \mu_{i,d}, \lambda_{x,i,d}, \beta_{i,d,k}, \lambda_{y,i,k}, Y_{n,k,i}^*\right\}$ are the latent variables and the distributions, $\{\mu_{i,d}, \lambda_{x,i,d}\}$, $\{\beta_{i,d,k}\}$, $\{\lambda_{y,i,k}\}$ and $\left\{Y_{n,k,i}^*\right\}$ are the DP base distributions.

IV. Weak Consistency of the SB-EX Model

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

$$\mathcal{W}_\epsilon\left(f_0\right)=\{f:|\int f_0\left(x,y\right)g\left(x,y\right)\,dx\,dy-\int f\left(x,y\right)g\left(x,y\right)\,dx\,dy<\epsilon\} \tag{7}$$

for every bounded, continuous function g. Now, the proof of the weak consistency of SB-EX model depends on a theorem by Schwartz [14], which says, if Π^f is a prior on $\mathcal{F}$ and if Π^f places a positive probability on all neighborhoods,

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

for every $\delta>0$, then Π^f is weak consistent at f_0. The proof is similar to Ghosal et al. [15] and Hannah et al. [3]. f_0 for the Exponential Model (SB-EX) has a compact support. So, there exists x_0 and y_0, such that $f_0(x,y)=0$ for $|x|>x_0$ or $|y|>y_0$, fixing $\epsilon>0$, we have,

$$\int\int f_0\left(x,y\right)\log\frac{f_0\left(x,y\right)}{\int\int\theta_x exp(-\theta_x x)\theta_y exp(-\theta_y y)f_0\left(x,y\right)\,d\theta_x\,d\theta_y}<\epsilon/2 \tag{8}$$

Let P_0 is a measure on $\{\lambda_x,\beta_0,\beta_0\}$. We define $dP_0=f_0\times\delta_0$. Fixing $k>0$, we choose a set K such that support of $P_0\subset K$. Let $\mathcal{B}=\{P:|P(K)/P_0(K)-1|<k\}$, therefore, $\Pi(\mathcal{B})>0$. From Ghosal et al. [15] and Hannah et al. [3], there exists a set $\mathcal{C}$ such that $\Pi(\mathcal{B}\cap\mathcal{C})>0$ and for every $P\in\mathcal{B}\cap\mathcal{C}$, for some k,

$$\int_0^{x_0}\int_0^{y_0}f_0\left(x,y\right)\log\frac{\int_K\theta_x exp(-\theta_x x)\theta_y exp(-\theta_y y)\,dP_0}{\int_K\theta_x exp(-\theta_x x)\theta_y exp(-\theta_y y)\,dP}<k/(1-k)+2k<\epsilon/2 \tag{9}$$

Therefore, from previous two equations, for every $P\in\mathcal{B}\cap\mathcal{C}$, for $f=\phi*P$,

$$\int f_0\left(x,y\right)\log\frac{f_0\left(x,y\right)}{f\left(x,y\right)}\,dx\,dy<\epsilon \tag{10}$$

So, the positive measure by Π^f on weak neighborhoods of f_0 ensures that the SB-EX model is weak consistent.

V. Variational Distribution of the Models

The inter-coupling between Y_n, X_n and z_n in all three models described above makes computing the posterior of Y_n analytically intractable. We therefore introduce the following fully factorized and decoupled variational distributions as surrogates.

A. Exponential Model (SB-EX)

The variational distribution for the Exponential model is defined formally as:

$$q\left(\mathbf{z,v,}\lambda_{x,i,d}\mathbf{,}\beta_{i,d}\right)=\prod_{i=1}^{T-1}q\left(v_i|\gamma_i\right)\prod_{n=1}^{N}q\left(z_n|\phi_n\right)$$
$$\prod_{i=1}^{T}\prod_{d=1}^{D}q\left(\lambda_{x,i,d}|a_{x,i,d},b_{x,i,d}\right)\prod_{i=1}^{T}\prod_{d=0}^{D}q\left(\beta_{i,d}|c_{y,i,d},d_{y,i,d}\right) \tag{11}$$

Firstly, each v_i follows a Beta distribution. As in [13], we have truncated the infinite series of $v_i^{'}$s into a finite one by making the assumption $q\left(v_T=1\right)=1$ and $M_i=0\forall i>T$. Note that this truncation applies to the variational surrogate distribution and *not* the actual posterior distribution that we approximate. Secondly, z_n follows a variational multinomial distribution. Thirdly, $\{\lambda_{x,i,d}\}$ and $\{\beta_{i,0}:\beta_{i,D}\}$, both follow a variational Gamma distribution.

B. Inverse Gaussian Model (SB-IG)

The variational distribution for the Inverse Gaussian Model model is given by:

$$q\left(\mathbf{z,v,}\mu_{i,d},\lambda_{x,i,d}\mathbf{,}\beta_{i,d},\lambda_{y,i}\right)=\prod_{i=1}^{T-1}q\left(v_i|\gamma_i\right)\prod_{n=1}^{N}q\left(z_n|\phi_n\right)$$
$$\prod_{i=1}^{T}\prod_{d=1}^{D}q\left(\mu_{i,d}|a_{x,i,d},\left(b_{x,i,d},\lambda_{x,i,d}\right)^{-1}\right)q\left(\lambda_{x,i,d}|c_{x,i,d},d_{x,i,d}\right)$$
$$\prod_{i=1}^{T}\prod_{d=0}^{D}q\left(\beta_{i,d}|a_{y,i,d},\left(b_{y,i},\lambda_{y,i}\right)^{-1}\right)q\left(\lambda_{y,i}|c_{y,i},d_{d,i}\right) \tag{12}$$

$\{\mu_{i,d},\lambda_{x,i,d}\}$ and $\{\beta_{i,0}:\beta_{i,D},\lambda_{y,i}\}$ both follows a variational Normal-Gamma distribution.

C. Multinomial Probit Model(SB-PBT)

The variational distribution for the Multinomial probit Model is

$$q\left(\mathbf{z,v,}\eta_x\mathbf{,}\eta_y\right)=\prod_{i=1}^{T-1}q\left(v_i|\gamma_i\right)\prod_{n=1}^{N}q\left(z_n|\phi_n\right)$$
$$\prod_{i=1}^{T}\prod_{d=1}^{D}q\left(\mu_{i,d}|a_{x,i,d},\left(b_{x,i,d},\lambda_{x,i,d}\right)^{-1}\right)q\left(\lambda_{x,i,d}|c_{x,i,d},d_{x,i,d}\right)$$
$$\prod_{i=1}^{T}\prod_{d=1}^{D}\prod_{k=1}^{K}q\left(\beta_{i,d,k}|m_{y,i,d,k},s_{y,i,d,k}^2\right)\prod_{k=1}^{K}\prod_{i=1}^{T}q\left(\lambda_{y,i,k}|a_{y,i,k},b_{y,i,k}\right)$$
$$\prod_{n=1}^{N}\prod_{k=1}^{K}\prod_{i=1}^{T}q\left(Y_{n,k,i}^*|\beta_{i,0,k}+\prod_{d=1}^{D}\beta_{i,d,k}X_{n,d},\lambda_{y,i,k}^{-1}\right) \tag{13}$$

Here, $\beta_{i,d,k}$ follows a Normal distribution. $\{\mu_{i,d},\lambda_{x,i,d}\}$ and $\left\{Y_{n,k,i}^*,\lambda_{y,i,k}\right\}$ follows a variational Normal-Gamma distribution. $\beta_{i,d,k}$ follows a normal distribution.

VI. Parameter Estimation for the Variational Distributions

We bound the log likelihood of the observations (same for all the models) using Jensen's inequality, $\phi\left(E\left[X\right]\right)\geq\mathrm{E}[\phi\left(X\right)]$, where, ϕ is a concave function and X is a random variable. This generalized ELBO is the same for all the three models under investigation and it is a function of the variational parameters as well as the hyper-parameters. We differentiate the individuall ELBOs with respect to the variational parameters of the specific models to obtain their respective estimates.

A. Parameter Estimation for the SB-EX Model

We differentiate the ELBO w.r.t. γ_i^1 and γ_i^2 and set them to zero to obtain estimates of γ_i^1 and γ_i^2,

$$\gamma_i^1=\alpha_1+\sum_{n=1}^{N}\phi_{n,i},\quad\gamma_i^2=\alpha_2+\sum_{n=1}^{N}\sum_{j=i+1}^{T}\phi_{n,j} \tag{14}$$

Estimating $\phi_{n,i}$ is a constrained optimization with $\sum \phi_{n,i} = 1$. We differentiate the Lagrangian w.r.t. $\phi_{n,i}$ to obtain,

$$\phi_{n,i} = \frac{exp\left(M_{n,i}\right)}{\sum_{i=1}^{T} exp\left(M_{n,i}\right)} \tag{15}$$

The term $M_{n,i}$ is represented as,

$$M_{n,i} = \sum_{j=1}^{i} \left\{\Psi\left(\gamma_j^2\right) - \Psi\left(\gamma_j^1 + \gamma_j^2\right)\right\} + P_{n,i} \tag{16}$$

where,

$$\begin{aligned} P_{n,i} = \sum_{n=1}^{N}\sum_{i=1}^{T}\sum_{d=1}^{D}\left\{\Psi\left(a_{x,i,d}\right) - ln\left(b_{x,i,d}\right) - X_{n,d}\frac{a_{x,i,d}}{b_{x,i,d}}\right\} + \\ \sum_{n=1}^{N}\sum_{i=1}^{T}\{-\frac{c_{y,i,0}}{d_{y,i,0}} - \sum_{d=1}^{D} X_{n,d}\frac{c_{y,i,d}}{d_{y,i,d}} - Y_n\frac{\Gamma\left(c_{y,i,0}\right)}{\left(d_{y,i,0}+1\right)c_{y,i,0}} \\ + Y_n\sum_{d=1}^{D}\frac{\Gamma\left(c_{y,i,d}\right)}{\left(d_{y,i,d}+X_{n,d}\right)c_{y,i,d}}\} \end{aligned} \tag{17}$$

The variational parameters for the covariates and responses are found by maximizing the ELBO w.r.t. them.

$$a_{x,i,d} = a_{x,d} + \sum_{n=1}^{N}\phi_{n,i},\ b_{x,i,d} = b_{x,d} + \sum_{n=1}^{N}\phi_{n,i}X_{n,d} \tag{18}$$

$$c_{y,i,d} = c_{y,d} + \sum_{n=1}^{N}\left(\phi_{n,i} + Y_n\right),\ d_{y,i,d} = d_{y,d} + \sum_{n=1}^{N}\phi_{n,i}\left(X_{n,d} + Y_n\right) \tag{19}$$

B. Parameter Estimation for the SB-IG Model

For the Inverse-Gaussian Model, the estimation of $\gamma_i^1, \gamma_i^2, \phi_{n,i}$ are identical to the Exponential model with the only difference being that $P_{n,i}$ is given as,

$$\begin{aligned} P_{n,i} = \frac{1}{2}\sum_{d=1}^{D}\{log\left(\frac{1}{2\pi}\right) + \Psi\left(c_{x,i,d}\right) - log\left(d_{x,i,d}\right) \\ -b_{x,i,d}^{-1} - \frac{c_{x,i,d}}{d_{x,i,d}}\left(X_{n,d} - a_{x,i,d}\right)^2\} + \frac{1}{2}\{log\left(\frac{1}{2\pi}\right) \\ +\Psi\left(c_{y,i}\right) - log\left(d_{y,i}\right) - b_{y,i}^{-1}\left(1 + \sum_{d=1}^{D}X_{n,d}^2\right) \\ -\frac{c_{y,i}}{d_{y,i}}\left(Y_n - a_{y,i,0} - \sum_{d=1}^{D}a_{y,i,d}X_{n,d}\right)^2\} \end{aligned} \tag{20}$$

The variational parameters for the covariates and responses are found by maximizing the ELBO w.r.t. them.

$$b_{x,i,d} = b_{x,d} + \sum_{n=1}^{N}\phi_{n,i},\ c_{x,i,d} = c_{x,d} + \sum_{n=1}^{N}\phi_{n,i} \tag{21}$$

$$\begin{aligned} d_{x,i,d} = \frac{1}{2}\{b_{x,d}\left(a_{x,i,d} - a_{x,d}\right)^2 + 2d_{x,d} \\ + \sum_{n=1}^{N}\frac{\phi_{n,i}\left(X_{n,d} - a_{x,i,d}\right)^2}{a_{x,i,d}^2 X_{n,d}}\} \end{aligned} \tag{22}$$

$$a_{x,i,d} = \frac{\sum_{n=1}^{N}\phi_{n,i}X_{n,d} + b_{x,d}m_{x,d}}{\sum_{n=1}^{N}\phi_{n,i} + b_{x,d}} \tag{23}$$

$$b_{y,i} = \frac{(D+1)b_y + \sum_{n=1}^{N}\phi_{n,i}\left(1 + \sum_{d=1}^{D}X_{n,d}^2\right)}{D+1} \tag{24}$$

$$c_{y,i} = \sum_{d=0}^{D}c_y + \frac{1}{2}\sum_{n=1}^{N}\phi_{n,i} \tag{25}$$

$$\begin{aligned} d_{y,i} = \frac{1}{2}\{\sum_{d=0}^{D}b_y\left(a_{y,i,d} - a_{y,d}\right)^2 + 2d_y \\ + \sum_{n=1}^{N}\frac{\phi_{n,i}\left(Y_n - a_{y,i,0} - \sum_{d=1}^{D}a_{y,i,d}X_{n,d}\right)^2}{\left(a_{y,i,0} - \sum_{d=1}^{D}a_{y,i,d}\right)^2 X_{n,d}}\} \end{aligned} \tag{26}$$

$$a_{y,i,0} = \frac{a_{y,d}b_y + \sum_{n=1}^{N}\phi_{n,i}\left(Y_n - \sum_{d=1}^{D}a_{y,i,d}X_{n,d}\right)}{b_y + \sum_{n=1}^{N}\phi_{n,i}} \tag{27}$$

$$\begin{aligned} a_{y,i,d} = \frac{a_{y,d}b_y}{b_y + \sum_{n=1}^{N}\phi_{n,i}X_{n,d}^2} \\ + \frac{\sum_{n=1}^{N}\phi_{n,i}\left(Y_n - a_{y,i,0} + a_{y,i,d}X_{n,d}\right)}{b_y + \sum_{n=1}^{N}\phi_{n,i}X_{n,d}^2} \\ - \frac{\sum_{n=1}^{N}\phi_{n,i}\sum_{d=1}^{D}a_{y,i,d}X_{n,d}}{b_y + \sum_{n=1}^{N}\phi_{n,i}X_{n,d}^2} \end{aligned} \tag{28}$$

C. Parameter Estimation for the SB-PBT Model

Again, in the Poisson Model, estimation of $\gamma_i^1, \gamma_i^2, a_{x,i,d}, b_{x,i,d}, c_{x,i,d}, d_{x,i,d}$, are similar to the Exponential model. The variational parameters are given by,

$$a_{y,i,k} = a_{y,k} + \sum_{n=1}^{N}\phi_{n,i},\ b_{y,i,k} = b_{y,k} \tag{29}$$

And, $m_{y,i,0,k} = m_{y,d,k} + s_{y,d,k}^2\sum_{n=1}^{N}\phi_{n,i}Y_{n,k}$

$$m_{y,i,d,k} = m_{y,d,k} + s_{y,d,k}^2\sum_{n=1}^{N}\phi_{n,i}Y_{n,k}X_{n,d} \tag{30}$$

D. Predictive Distribution

Finally, we derive the predictive distribution for a new response given a new covariate and the set of previous covariate-response pairs.

$$\begin{aligned} p\left(Y_{N+1}|\mathbf{X_{N+1}}, \mathbf{X}, \mathbf{Y}\right) = \\ \sum_{z}\int\int p\left(Y_{N+1}|\mathbf{X_{N+1}}, \eta_{\mathbf{y}}, \mathbf{z}\right)p\left(\mathbf{v}, \eta_{\mathbf{y}}|\mathbf{Y}, \mathbf{X}\right)p\left(\mathbf{z}|\mathbf{v}\right)d\mathbf{v}d\eta_y \end{aligned} \tag{31}$$

Since the inner integrals are analytically intractable, we approximate the predictive distribution by replacing the true posterior, $p\left(\mathbf{v}, \eta_{\mathbf{y}}|\mathbf{Y}, \mathbf{X}\right)$, with its variational surrogate, $q\left(\mathbf{v}\right)q\left(\eta_{\mathbf{y,i}}\right)$. The density, $q(v)$, with the density, $p\left(\mathbf{z}|\mathbf{v}\right)$ is

Initialize Hyper-parameters of the Generative Model.
Repeat
Evaluate γ_i^1 and γ_i^2 According to Eq. 14 (SB-EX model).
Evaluate $\phi_{n,i}$ of the respective Model According to Eq. 15 (SB-EX model).
Evaluate Variational Parameters of the Covariate Distribution Accroding to Eq. 18. (SB-EX model)
Evaluate Variational Parameters of the Response Distribution Accroding to Eq. 19. (SB-EX model)
until converged

TABLE I. THE COMPLETE VARIATIONAL INFERENCE ALGORITHM

integrated out to give the weight factor w_i for each mixture. Here, w_i is given by,

$$w_i = \frac{\gamma_i^1 \gamma_i^2 \left(\gamma_i^2 + 1\right) \ldots\ldots\ldots\ldots \left(\gamma_i^2 + T - 1 - i\right)}{\left(\gamma_i^1 + \gamma_i^2\right)\left(\gamma_i^1 + \gamma_i^2 + 1\right) \ldots\ldots\ldots \left(\gamma_i^1 + \gamma_i^2 + T - i\right)} \tag{32}$$

The integration of the densities $q\left(\eta_{y,i}\right)$ and $p\left(Y_{N+1}\right)$ is not analytically tractable. We have therefore used basic Monte Carlo Integration by sampling from the distribution of the latent variables, just like computing erf(z). This does not at all suffer the difficulties faced in a MCMC sampler as there is no Markov chain used for sampling. The integrals were computed within milliseconds.

$$\begin{aligned} E\left[Y_{N+1}|\mathbf{X_{N+1}}, \mathbf{X}, \mathbf{Y}\right] &= E\left[E\left[Y_{N+1}|\mathbf{X_{N+1}}, \eta_{\mathbf{y},\mathbf{i}(\mathbf{1}:\mathbf{T})}\right]|\mathbf{X}, \mathbf{Y}\right] \\ &= \frac{1}{M}\sum_{m=1}^{M} E\left[Y_{N+1}|\mathbf{X_{N+1}}, \eta^{\mathbf{m}}_{\mathbf{y},\mathbf{i}(\mathbf{1}:\mathbf{T})}\right] \end{aligned} \tag{33}$$

In all experiments presented in this paper, we collected 100 i.i.d. samples from the density of $\eta_{y,i}$ to evaluate the expected value of Y_{N+1} from the density of $p\left(Y_{N+1}\right)$. The complete variational inference algorithm is given in Table 1.

VII. EXPERIMENTAL RESULTS

A broad set of experiments were conducted to evaluate the SB-EX, SB-IG and SB-PBT models. Samples from the predictive posterior were used to evaluate the accuracy of the SB-EX and SB-IG models against its competitor algorithms, such as, Linear regression with no feature selection (OLS), Bayesian Linear regression, Variational Linear regression [17], Gaussian Process regression [16], and ordinary DP regression [3]. The accuracy of the SB-PBT model was evaluated against Multiclass Support Vector Machine [19], Naive Bayes Model [18] and Multinomial Logistic Model [17].

Next, to highlight SB-EX and SB-IG as a practical tool, it was employed as a new GLM-based technique to model the volatility dynamics of the stock market. Specifically, it was used to determine how individual stocks tract predetermined baskets of stocks over time.

A. Datasets

One artificial group of datasets and three real world datasets were used. In the artificial set, we generated several 50 to 100 dimensional regression datasets with 10 clusters each in the covariate-response space (Y, X). The covariates were generated from independent Inverse Gaussians with means varying from 1 to 27 in steps of 3 for the 10 clusters. The shape parameter was drawn independently from the range $[.1, 1]$ for the 10 clusters. For a fixed cluster, the shapes were set to be the same for each dimension. The second dataset was a compilation of daily stock price data (retrieved from Google Finance) for the "Dow 30" companies from Nov 29, 2000 to Dec 29, 2013. It had 3268 instances and was viewed as 30 different 29-1 covariate-response datasets. The goal was to model the stock price of an individual Dow-30 company as a function of the remaining 29 companies, over time. Accuracy results were averaged over all 30 regressions. The third dataset was the Parkinson's telemonitoring dataset [20] from the UCI Machine Learning Repository that has 5875 instances over 16 covariates. The final dataset was the Breast Cancer Wisconsin (Original) dataset [21] from the UCI Repository that has 699 instances over 10 covariates. This dataset was used to evaluate SB-PBT against competitors like Multiclass SVM [19], Multinomial Logistic regression [17] and Naive Bayes model [18].

Time-Period	Cisco	Goldman Sachs	Chevron	McDonald	Boeing
2000-07	Verizon	JPM	XOM	J and J	DD
	IBM	VISA	Boeing	Coca-Cola	GE
	GE	AXP	MMM	NKE	GS
2007-09	AXP	XOM	AT-T	MMM	MCD
	INTEL	NKE	PG	IBM	VISA
	DIS	DD	Coca-Cola	TRX	MMM
2009-13	INTEL	AXP	XOM	Coca-cola	CAT
	MSFT	PG	CAT	Merck	DD
	DD	JPM	GE	J and J	JPM

TABLE II. LIST OF FIVE DIFFERENT STOCKS WITH TOP 3 MOST SIGNIFICANT STOCKS THAT INFLUENCE EACH STOCK. HERE, INTEL, VERIZON, CISCO, IBM, AT-T ARE TECH. STOCKS, MMM, CAT, DD, BOEING, GE ARE MACHINERY/CHEMICAL STOCKS, XOM, CHEVRON ARE ENERGY STOCKS, AXP, GS, PG, TRX, JPM, VISA ARE FINANCE/RETAIL STOCKS AND MCD, J-J, COCA-COLA ARE FOOD STOCKS.

B. Accuracy

We report the mean absolute error (MAE) and Mean Square Error (MSE) for all the algorithms in Table 3 for the first 3 datasets. Note that SB-EX and SB-IG yield the least error values among its competitors. For the classification dataset, we have reported the class accuracy percentage where SB-PBT has obtained the highest class accuracy percentage.

C. SB-IG and SB-EX as a Tool to Understand Stock Market Dynamics

SB-EX and SB-IG is presented as a new tool to analyze the dynamics of stocks from the "Dow 30" companies. "Dow 30" stocks belong to disparate market sectors such as, technology (Microsoft, Intel etc.), finance (Goldman Sachs, American Express etc.), food/pharmaceuticals (Coca-cola, McDonald,Johnson and Johnson), Energy and Machinery (Chevron, GE, Boeing, Exxon Mobil). We divided the dataset into 3 time segments on the two sides of the financial crisis of 2008. The first comprised of the stock values from Nov-00 to Nov-07 and the third of the stock values from Dec-08-Dec13. The middle, set as the remainder, was representative of the financial crisis.

Using SB-EX and SB-IG, we modeled each company's stock value as a function of the values of the others in DOW 30. We recorded the stocks having the most impact on the determination of the value of each stock. The impacts are necessarily the magnitude of the weighted coefficients of

Synthetic Data	MAE			MSE		
Training Percent	30	60	90	30	60	90
SB-IG	1.21	.89	.79	1.78	1.55	1.39
SB-EX	1.32	1.26	1.16	1.85	1.78	1.44
ODP	1.47	1.37	1.29	1.95	1.82	1.52
GPR	1.56	1.42	1.63	2.34	2.17	1.79
VLR	1.71	1.53	1.29	2.49	2.28	2.82
BLR	1.92	1.59	1.41	2.71	2.44	1.92
LR	1.55	1.47	1.36	2.78	2.57	2.12
Stock Market Data	MAE			MSE		
Training Percent	30	60	90	30	60	90
SB-IG	.74	.63	.56	1.39	1.28	1.13
SB-EX	1.01	.92	.79	1.62	1.51	1.40
ODP	.99	.88	.73	1.74	1.57	1.38
GPR	.83	.76	.68	1.53	1.44	1.29
VLR	1.07	.99	.90	1.82	1.71	1.50
BLR	1.16	1.05	.92	1.89	1.76	1.56
LR	1.25	1.13	1.01	1.94	1.83	1.64
Telemonitoring Data	MAE			MSE		
Training Percent	30	60	90	30	60	90
LR	1.86	1.55	1.36	2.09	1.66	1.36
BLR	1.91	1.60	1.32	2.13	1.63	1.30
VLR	1.88	1.52	1.28	2.07	1.70	1.33
ODP	1.85	1.59	1.33	2.10	1.64	1.29
GPR	1.80	1.56	1.27	2.04	1.57	1.26
SB-IG	1.79	1.54	1.25	2.01	1.59	1.25
SB-EX	1.77	1.48	1.23	1.99	1.53	1.20
Breast Cancer Data	Class Percentage Accuracy					
Training Percent	30		60		90	
SB-PBT	86.4		92.1		98.3	
Naive Bayes	69.7		76.9		82.8	
SVM	74.4		78.7		86.9	
Logistic	75.3		81.2		89.5	

TABLE III. FIRST 3 PORTION DEPICTS MSE AND MAE OF THE ALGORITHMS FOR THE SYNTHETIC DATASET(50,75,100 DIMENSIONS), STOCK MARKET DATASET AND TELEMONITORING DATA SET WITH 30, 60 AND 90 % OF DATA SET AS TRAINING. THE LAST PORTION REPRESENTS CLASS PERCENTAGE ACCURACY OF THE ALGORITHMS FOR THE BREAST CANCER DATA.

the covariates (the stock values) in SB-EX and SB-IG. Two significant trends were noteworthy.

Firstly, when the market was stable (the first and third segments), stocks from any given sector had impact largely on the same sector, with few stocks being influential overall. Secondly, the sectors having the most impact on a specific stock were the same on both sides of the crisis. For example, Microsoft (tech. sector), is largely modeled by Intel, IBM (tech), GE (machinery) and JPM (finance) previous to the crisis and modeled by Cisco, Intel (tech), Boeing (machinery) and GS (finance) (in descending order of weights) post crisis. However, during the crisis, the stocks showed no such trends. For example, Microsoft is impacted by GS, MMM, TRX and Cisco showing no sector wise trend. We report 5 additional such results in Table 2.

VIII. CONCLUSION

In this paper, we have formulated infinite mixtures of Inverse Gaussian, Multinomial Probit and Exponential Regression via a Stick Breaking Prior as Hierarchical Bayesian graphical models. We have derived fast mean field variational inference algorithms for each of the models. The algorithm is particularly useful for high dimensional datasets where Gibbs sampling fails to scale and is slow to converge. The algorithm has been tested successfully on four datasets against its well known competitor algorithms across many settings of training/testing splits. While SB-IG, SB-EX and SB-PBT has been developed in a simple setting, developing its counterparts for the Hierarchical Generalized Linear Models remain topics for future research. Furthermore, we have considered here mean field variational methods; it would be worth exploring other variational methods in the non-parametric Bayesian context to the Generalized Linear Model.

REFERENCES

[1] J.A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society, Series A (General)*, vol. 135, pp. 370-384, 1972.

[2] T.S. Ferguson, "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, vol. 1, pp. 209-230, 1973.

[3] L. Hannah, D. Blei and W. Powell, "Dirichlet Process Mixtures of Generalized Linear Models," *Journal of Machine Learning Research*, vol. 12, pp. 1923-1953, 2011.

[4] J. Sethuraman, "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, vol. 4, pp. 639-650, 1994.

[5] C.E. Antoniak, "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *Annals of Statistics*, vol. 2, no. 6, pp. 1152-1174, 1973.

[6] M. Escobar and M. West, "Bayesian Density Estimation and Inference Using Mixtures," *Journal Of American Statistical Association*, vol. 90, pp.577-588, 1995.

[7] J. Ishwaran and L. James, "Gibbs Sampling Methods for Stick Breaking Priors," *Journal Of American Statistical Association*, vol. 96, pp. 161-174, 2001.

[8] R. Neal, "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal Of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249-265, 2001.

[9] C. Robert and G. Casella, "Monte Carlo Statistical Methods," *Springer-Verlag*, 2001.

[10] D. Blei, A. Ng and M. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[11] M. Jordan, Z. Ghahramani, T. Jaakkola and L. Saul, "Introduction to Variational Methods for Graphical Models," *Machine Learning* vol. 37, pp. 183-233, 2001.

[12] Z. Ghahramani and M. Beal, "Propagation Algorithms for Variational Bayesian Learning," *Proceedings of 13th Advances in Neural Information Processing Systems*, pp. 507-513, 2000.

[13] D. Blei and M. Jordan, "Variational Inference for Dirichlet Process Mixtures," *Bayesian Analysis*, vol. 1, pp. 121-144, 2006.

[14] L. Schwartz, "On Bayes procedures," *Z. Wahrsch. Verw. Gebiete*, vol. 4, no. 1, pp. 1026, 1965.

[15] S. Ghosal, JK Ghosh, and RV Ramamoorthi, "Posterior consistency of Dirichlet mixtures in density estimation," *The Annals of Statistics*, vol. 27, no. 1, pp. 143158, 1999.

[16] C.E. Rasmussen and C. K. I. Williams, "Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)," *MIT Press*, 2005.

[17] C. M. Bishop, "Pattern Recognition and Machine Learning," *Springer-Verlag*, 2006.

[18] D. Lowd and P. Domingos, "Naive Bayes models for probability estimation," *Proceedings of the 22nd international conference on Machine learning*, pp. 529-536, 2005.

[19] C. Cortes and V. Vapnik, "Support Vector Networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.

[20] A. Tsanas, M. A. Little, P. E. McSharry and L. O. Ramig, "Accurate Telemonitoring of Parkinsons Disease Progression by Non-invasive Speech Tests," *IEEE transactions on Biomedical Engineering*, vol. 57, pp. 884-893, 2009.

[21] W.H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proceedings of the National Academy of Sciences*, vol. 87, pp. 9193-9196, 1990.