Automatic Detection of Latent Common Clusters of Groups in MultiGroup Regression

Minhazul Islam Sk(✉) and Arunava Banerjee

University of Florida, Gainesville, Florida, USA
smislam@cise.ufl.edu

Abstract. We present a flexible non-parametric generative model for multigroup regression that detects latent common clusters of groups. The model is founded on techniques that are now considered standard in the statistical parameter estimation literature, namely, Dirichlet process(DP) and Generalized Linear Model (GLM), and therefore, we name it "Infinite MultiGroup Generalized Linear Models" (iMG-GLM). We present two versions of the core model. First, in iMG-GLM-1, we demonstrate how the use of a DP prior on the groups while modeling the response-covariate densities via GLM, allows the model to capture latent clusters of groups by noting similar densities. The model ensures different densities for different clusters of groups in the multigroup setting. Secondly, in iMG-GLM-2, we model the posterior density of a new group using the latent densities of the clusters inferred from previous groups as prior. This spares the model from needing to memorize the entire data of previous groups. The posterior inference for iMG-GLM-1 is done using Variational Inference and that for iMG-GLM-2 using a simple Metropolis Hastings Algorithm. We demonstrate iMG-GLM's superior accuracy in comparison to well known competing methods like Generalized Linear Mixed Model (GLMM), Random Forest, Linear Regression etc. on two real world problems.

1 Introduction

Multigroup Regression is the method of choice for research design whenever response-covariate data is collected across multiple groups. When a common regressor is learned on the amalgamated data, the resultant model fails to identify effects for the responses specific to individual groups because the underlying assumption is that the response-covariate pairs are drawn from a single global distribution, when the reality might be that the groups are not statistically identical, making the joining of them inappropriate. Modeling separate groups via separate regressors results in a model that is devoid of common latent effects across the groups. Such a model does not exploit the patterns common among the groups ensuring in turn the transferability of information among groups in the regression setting. This is of particular importance when the training set is very small for many of the groups. Joint learning, by sharing knowledge between the statistically similar groups, strengthens the model for each group, and the resulting generalization in the regression setting is vastly improved.

© Springer International Publishing Switzerland 2016
P. Perner (Ed.): MLDM 2016, LNAI 9729, pp. 251–266, 2016.
DOI: 10.1007/978-3-319-41920-6_19

The complexities that underlie the utilization of the information transfer between the groups are best motivated through examples. In Clinical Trials, for example, a group of people are prescribed either a new drug or a placebo to estimate the efficacy of the drug for the treatment of a certain disease. At a population level, this efficacy may be modeled using a single Normal or Poisson mixed model distribution with mean set as a (linear or otherwise) function of the covariates of the individuals in the population. A closer inspection might however disclose potential factors that explain the efficacy results better. For example, there might be regularities at the group level—Caucasians as a whole might react differently to the drug than, say, Asians, who might, furthermore, comprise many groups. Identifying this across group information would therefore improve the accuracy of the regressor. Similarly in the Stock Market, future values and trends for a group of stocks are predicted for various sectors such as Energy, Materials, Consumer discretionary, Financials, Telecomm., Technology, etc. Within each sector, various stocks share trends and therefore predicting them together (modeling them with the same time series via autoregressive density) is usually much more accurate than predicting and capturing individual trends. Modeling the latent common clustering effects of cross-cutting subgroups is therefore an important problem to solve. We present a framework here that accomplishes this.

We begin with a brief description of the weaknesses of the most popular multilevel regression techniques, namely, Generalized Linear Models [19] and Mixed model [7]. In regression theory, Generalized Linear Model (GLM), proposed in [19], brings erstwhile disparate techniques such as, Linear regression, Logistic regression, and Poisson regression, under a unified framework. GLM is formally defined as:

$$f\left(y;\theta,\psi\right)=exp\left\{\frac{y\theta-b\left(\theta\right)}{a\left(\psi\right)}+c\left(y;\psi\right)\right\} \tag{1}$$

Here, ψ is a dispersion parameter. exp denotes the exponential family density. The mean response is $E\left[Y|\mathbf{X}\right]=b\left(\theta\right)=\mu=g^{-1}\left(X^{T}\beta\right)$, where g is the link function, $X^{T}\beta$ is the linear predictor. For multigroup regression, Generalized Linear Mixed Model (GLMM) [7] and Hierarchical Generalized Linear Mixed Model [14] have been developed where similarities between groups is captured though a Fixed effect and variation across groups is captured through random effects. Statistically, these models are very rigid since every group is forced to manifest the same fixed effect, while the random effect only represents the intercept parameter of the linear predictors. Cluster of groups may have significantly different properties from other clusters of groups, a feature that is not captured in these traditional GLM based models. Furthermore, various clusters of groups may have different uncertainties with respect to the covariates which we denote as heteroscedasticity. In recent progress, [3] has proposed a Bayesian Hierarchical model, where a prior is used for the mixture of groups. Nevertheless, individual groups are given weights as opposed to jointly learning various groups. Also, the number of mixtures are fixed in advance.

Before, presenting our algorithm, we describe our basis for identifying group-correlation. First, two groups are correlated if their responses follow the same distribution. Second, two groups that have the same response variance with respect to the covariates are deemed to be correlated. This is achieved via a Dirichlet Process prior on the groups and the covariate co-efficients (β). The posterior is obtained by appropriately combining the prior and the data likelihood from the given groups. The prior helps cluster the groups and the likelihood from the individual groups help in the sharing of trends between groups to create the single posterior density between the many potential groups, thereby leading to group-correlation.

We now present an overview of our iMG-GLM framework. Our objective is to achieve (a) shared learning of various groups in a regression setting, where data may vary in terms of temporal, geographical or other modalities and (b) automatic clustering of groups which display correlation. iMG-GLM-1 solves this task. In iMG-GLM-2, we model a completely new group after modeling previous groups through parameters learned in iMG-GLM-1. In the first part, the regression parameters are given a Dirichlet Process prior, that is, they are drawn from a DP with the base distributions set as the density of the regression parameters. Since a draw from a DP is an atomic density, to begin, one group will be assigned one density of the regression parameters which signifies the response density with respect to its covariates. As the drawn probability weight from the DP increases, the cluster starts to consume more and more groups in this mutigroup setting. We employ a variational Bayes algorithm for the inference procedure in iMG-GLM-1 for computational efficiency. iMG-GLM-1 is then extended to iMG-GLM-2 for modeling a completely new group. Here we transfer the information (covariate coefficients) obtained in the first part, to learning a new group. In essence, the cluster parameters (covariate coefficients for the whole group) are used as a prior distribution for the model parameters of the new group's response density. This therefore leads to a mixture model where the weights are given by the number of groups that one cluster consumed in the first part and the mixture components are the regression parameters obtained for that specific cluster. The likelihood comes from the data of the new group. We use a simple accept-reject based Metropolis Hastings algorithm to generate samples from the posterior for the new group regression parameter density. For both iMG-GLM-1 and iMG-GLM-2, we use Monte Carlo integration for evaluating the predictive density of the new test samples.

We evaluate both iMG-GLM-1 and iMG-GLM-2 Normal models in two real world problems. The first is the prediction and finding of trends in the Stock Market. We show how information transfer between groups help our model to effectively predict future stock values by varying the number of training samples in both previous and new groups. In the second, we show the efficacy of i-MG-GLM-1 and 2 Poisson model against its competitors in a very important Clinical Trial Problem Setting.

2 Mathematical Background

2.1 Models Related to iMG-GLM

After its introduction, Generalized Linear Model was extended to Hierarchical Generalized Linear Model (HGLM) [14]. Then it included structured dispersion in [15] and models for spatio-temporal co-relation in [16]. Generalized Linear Mixed Models (GLMMs) were proposed in [7]. The random effects in HGLM were specified by both mean and dispersion in [17]. Mixture of Linear Regression was proposed in [22]. Hierarchical Mixture of Regression was done in [13]. Varying co-efficient models were proposed in [11]. Multi-tasking Model for classification in Non-parametric Bayesian scenario was introduced in [23]. Sharing Hidden Nodes in Neural Networks was introduced in [4,5]. General Multi-Task learning was described first in [8]. Common prior in hierarchical Bayesian model was used in [24,25]. Common structure sharing in the predictor space was presented in [1]. All of these models suffer the shortcomings of not identifying the latent clustering effect across groups as well as varying uncertainty with respect to covariates across groups, which the iMG-GLM inherently models.

2.2 Dirichlet Process and its Stick-Breaking Representation

A Dirichlet Process [10], $D\left(\alpha, G_0\right)$ is defined as a probability distribution over a sample space of probability distributions, $G \sim DP\left(\alpha, G_0\right)$ and $\eta_j|G \sim G$. Here, α is the concentration parameter and G_0 is the base distribution.

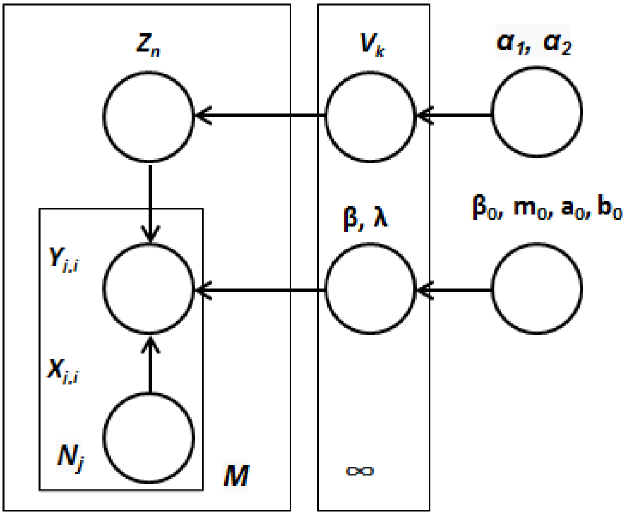

Fig. 1. Graphical Representation of iMG-GLM-1 Model.

When we integrate over G, the conditional density of η_j, given previous $\eta_{1:j-1}$ is given by the *Chinese Restaurant process* [2]. $\eta_j|\theta_{1:j-1}, \alpha, G_0 \sim \frac{\alpha}{\alpha+j-1}G_0 + \frac{1}{\alpha+j-1}\sum_{k=1}^{j-1} n_{-j,k}\delta_{\eta_k^*}$. Here, $n_{-m,k}$ denotes the number of η's equal to η_k^* (From K distinct values) excluding η_j.

According to the stick-breaking construction [21] of DP, G, which is a sample from DP, is an atomic distribution with countably infinite atoms drawn from G_0.

$$\begin{aligned} &v_k|\alpha, G_0 \sim Beta(1,\alpha), \quad \theta_k|\alpha, G_0 \sim G_0, \\ &\pi_i = v_k \prod_{p=1}^{k-1}(1-v_p), \quad G = \sum_{k=1}^{\infty} \pi_k.\delta_{\theta_k} \end{aligned} \tag{2}$$

In the DP mixture model [9], DP is used as a non-parametric prior over parameters of an Infinite Mixture model.

$$\begin{aligned} &z_n|\{v_1, v_2, ...\} \sim Categorical\{\mathit{\Pi}_1, \mathit{\Pi}_2, \mathit{\Pi}_3....\}, \\ &X_n|z_n, (\theta_k)_{i=1}^{\infty} \sim F(\theta_{z_n}) \end{aligned} \tag{3}$$

Here, F is a distribution parametrized by θ_{z_n}. $\{\pi_1, \pi_2, \pi_3, ...\}$ is defined by Eq. 2.2.

3 iMG-GLM Model Formulation

We consider M groups indexed by $j = 1, ..., M$ and the complete data as $\mathcal{D} = \{x_{j,i}, y_{j,i}\}$ s.t. $i = 1, ...N_j$. $\{x_{j,i}, y_{j,i}\}$ are covariate-response pairs and are drawn i.i.d. from an underlying density which differs along with the nature of $\{x_{j,i}, y_{j,i}\}$ among various models.

3.1 Normal iMG-GLM-1 Model

In the Normal iMG-GLM-1 model, the generative model of the covariate-response pair is given by the following set of equations. Here, X_{ji} and Y_{ji} represent the i^{th} continuous covariate-response pairs of the j^{th} group. The distribution of $Y_{j,i}|X_{j,i}$ is normal parametrized by $\beta_{0:D}$ and λ. The distribution, $\{\beta_{kd}, \lambda_k\}$ (Normal-Gamma) is the prior distribution on the covariate coefficient β. This distribution is the base distribution (G) of the Dirichlet Process. The set $\{m_0, \beta_0, a_0, b_0\}$ constitute the hyper-parameters for the covariate coefficients (β) distribution. The graphical representation of the normal model is given in Figure 1.

$$\begin{aligned} &v_k \sim Beta(\alpha_1, \alpha_2), \quad \pi_k = v_k \mathit{\Pi}_{n=1}^{k-1}(1-v_n) \\ &\quad \mathrm{N}\left(\beta_{kd}|m_0, (\beta_0, \lambda_k)^{-1}\right) \mathrm{Gamma}\left(\lambda_k|a_0, b_0\right) \\ &\quad Z_j|v_k \sim Categorical\left(\pi_1, ...\pi_\infty\right) \\ &\quad Y_{ji}|X_{ji} \sim \mathcal{N}\left(Y_{ji}|\sum_{d=0}^{D}\beta_{Z_j d}X_{jid}, \lambda_{Z_j}^{-1}\right) \end{aligned} \tag{4}$$

3.2 Logistic Multinomial iMG-GLM-1 Model

In the Logistic Multinomial iMG-GLM-1 model, a Multinomial Logistic Framework is used for a Categorical response, Y_{ji}, for a continuous covariate, X_{ji}, in the case of i^{th} data point of the j^{th} group. t is the index of the category. The distribution of $Y_{j,i}|X_{j,i}$ is Categorical parametrized by $\beta_{0:D,0:T}$. The distribution, $\{\beta_{ktd}\}$ (Normal) is the prior distribution on the covariate coefficient β which is the base distribution (G) of the Dirichlet Process. The set $\{m_0, s_0\}$ constitute the hyper-parameters for the covariate coefficients (β) distribution.

$$\begin{aligned}
&v_k \sim Beta(\alpha_1, \alpha_2), \quad \pi_k = v_k \Pi_{n=1}^{k-1}(1 - v_n) \\
&\beta_{ktd} \sim \mathcal{N}\left(\beta_{ktd}|m_0, s_0^2\right), \quad Z_j|v_k \sim Categorical\,(\pi_1, ...\pi_\infty) \\
&Y_{ji} = t|X_{ji}, Z_j \sim \frac{\exp\left(\sum_{d=0}^{D} \beta_{Z_j td} X_{jid}\right)}{\sum_{t=1}^{T} \exp\left(\sum_{d=0}^{D} \beta_{Z_j td} X_{jid}\right)}
\end{aligned} \tag{5}$$

3.3 Poisson iMG-GLM-1 Model

In the Poisson iMG-GLM model, a Poisson distribution is used for the count response. Here, X_{ji} and Y_{ji} represent the i^{th} continuous/ordinal covariate and categorical response pair of the j^{th} group. The distribution of $Y_{j,i}|X_{j,i}$ is Poisson parametrized by $\beta_{0:D,0:T}$. The distribution, $\{\beta_{kd}\}$ (Normal) is the prior distribution on the covariate coefficient β which is the base distribution (G) of the Dirichlet Process. The set $\{m_0, s_0\}$ constitute the hyper-parameters for the covariate coefficients (β) distribution.

$$\begin{aligned}
&v_k \sim Beta(\alpha_1, \alpha_2), \quad \pi_k = v_k \Pi_{n=1}^{k-1}(1 - v_n), \\
&\{\beta_{k,d}\} \sim \mathcal{N}\left(\beta_{kd}|m_0, s_0^2\right) \\
&Y_{ji}|X_{ji}, Z_j \sim Poisson\left(y_{ji}|\exp\left(\sum_{d=0}^{D} \beta_{Z_j d} X_{jid}\right)\right)
\end{aligned} \tag{6}$$

4 Variational Inference

The inter-coupling between Y_{ji}, X_{ji} and z_j in all three models described above makes computing the posterior of the latent parameters analytically intractable. We therefore introduce the following fully factorized and decoupled variational distributions as surrogates.

4.1 Normal iMG-GLM-1 Model

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

$$\begin{aligned}
&q\,(z, v, \beta_{kd}, \lambda_k) = \prod_{k=1}^{K} Beta\left(v_k|\gamma_k^1, \gamma_k^2\right) \prod_{j=1}^{M} Multinomial\,(z_j|\phi_j) \\
&\prod_{k=1}^{K} \prod_{d=0}^{D} \mathcal{N}\left(\beta_{kd}|m_{kd}, (\beta_k, \lambda_k)^{-1}\right) \text{Gamma}\,(\lambda_k|a_k, b_k)
\end{aligned} \tag{7}$$

Firstly, each v_k follows a Beta distribution. As in [6], we have truncated the infinite series of $v_k^{'}$s into a finite one by making the assumption $p\,(v_K = 1) = 1$ and $\pi_k = 0 \forall k > K$. Note that this truncation applies to the variational surrogate distribution and *not* the actual posterior distribution that we approximate. Secondly, z_j follows a variational multinomial distribution. Thirdly, $\{\beta_{kd}, \lambda_k\}$ follows a Normal-Gamma distribution.

4.2 Logistic Multinomial iMG-GLM-1 Model

The variational distribution for the Logistic Multinomial model is given by:

$$\begin{aligned}&q\left(z,v,\beta_{kd},\lambda_k\right)=\prod_{k=1}^{K} Beta\left(v_k|\gamma_k^1,\gamma_k^2\right)\prod_{j=1}^{M} Multinomial\left(z_j|\phi_j\right)\\&\prod_{k=1}^{K}\prod_{t=1}^{T}\prod_{d=0}^{D}\left\{\mathcal{N}\left(\beta_{ktd}|m_{ktd},s_{ktd}^2\right)\right\}\end{aligned}\tag{8}$$

Here, v_k and z_j represent the same distributions as described in the Normal iMG-GLM-1 model above. $\{\beta_{ktd}\}$ follows a variational Normal Model.

4.3 Poisson iMG-GLM-1 Model

The variational distribution for the Poisson iMG-GLM-1 model is given by:

$$\begin{aligned}&q\left(z,v,\beta_{kd},\lambda_k\right)=\prod_{k=1}^{K} Beta\left(v_k|\gamma_k^1,\gamma_k^2\right)\\&\prod_{j=1}^{M} Multinomial\left(z_j|\phi_j\right)\prod_{k=1}^{K}\prod_{d=0}^{D}\left\{\mathcal{N}\left(\beta_{ktd}|m_{ktd},s_{ktd}^2\right)\right\}\end{aligned}\tag{9}$$

Here, v_k and z_j represent the same distributions as described in the Normal iMG-GLM-1 model above. $\{\beta_{kd}\}$ follows a variational Normal Model.

5 Parameter Estimation for Variational Distribution

We bound the log likelihood of the observations in the generalized form of iMG-GLM-1 (same for all the models) using Jensen's inequality, $\phi\left(E\left[X\right]\right)\geq$E$[\phi\left(X\right)]$, where, ϕ is a concave function and X is a random variable. In this section, we differentiate the individually derived bounds with respect to the variational parameters of the specific models to obtain their respective estimates.

5.1 Parameter Estimation of iMG-GLM-1 Normal Model

The parameter estimation of the Normal Model is as follows:

$$\begin{aligned}&\gamma_k^1=1+\sum_{i=1}^{M}\phi_{ik},\quad\gamma_k^2=\alpha+\sum_{i=1}^{M}\sum_{p=k+1}^{K}\phi_{n,p}\\&\phi_{jk}=\frac{exp\left(S_{jk}\right)}{\sum_{k=1}^{K}exp\left(S_{jk}\right)}\quad s.t.\\&S_{jk}=\sum_{j=1}^{k}\left\{\Psi\left(\gamma_j^1\right)-\Psi\left(\gamma_j^1+\gamma_j^2\right)\right\}+P_{jk}\quad s.t.\\&P_{jk}=\frac{1}{2}\sum_{j=1}^{M}\sum_{i=1}^{N_j}\phi_{jk}\{log\left(\frac{1}{2\pi}\right)+\Psi\left(a_k\right)-log\left(b_k\right)\\&-\beta_k\left(1+\sum_{d=1}^{D}X_{jid}^2\right)-\frac{a_k}{b_k}\left(Y_{ji}-m_{k0}-\sum_{d=1}^{D}m_{kd}X_{jid}\right)^2\}\\&\beta_k=\frac{(D+1)\beta_0+\sum_{j=1}^{M}\sum_{i=1}^{N_j}\phi_{jk}\left(1+\sum_{d=1}^{D}X_{jid}^2\right)}{D+1}\\&a_k=\sum_{d=0}^{D}a_0+\frac{1}{2}\sum_{j=1}^{M}\sum_{i=1}^{N_j}\phi_{jk}\\&b_k=\frac{1}{2}\{\sum_{d=0}^{D}\beta_0\left(m_{kd}-m_0\right)^2+2b_0\\&+\sum_{j=1}^{M}\sum_{i=1}^{N_j}\phi_{jk}\left(Y_{ji}-m_{k0}-\sum_{d=1}^{D}m_{kd}X_{jid}\right)^2\}\\&m_{k0}=\frac{m_0\beta_0+\sum_{j=1}^{M}\sum_{i=1}^{N_j}\phi_{ji}\left(Y_{ji}-\sum_{d=1}^{D}m_{kd}X_{jid}\right)}{\beta_0+\sum_{j=1}^{M}\sum_{i=1}^{N_j}\phi_{jk}}\\&m_{kd}=\frac{m_0\beta_0+\sum_{j=1}^{M}\sum_{i=1}^{N_j}\phi_{ji}\left(Y_{ji}-m_{k0}-\sum_{d=1}^{D-(d)}m_{kd}X_{jid}\right)X_{jid}}{\beta_0+\sum_{j=1}^{M}\sum_{i=1}^{N_j}\phi_{jk}X_{jid}^2}\end{aligned}\tag{10}$$

5.2 Parameter Estimation of iMG-GLM-1 Multinomial Model

For the Logistic Multinomial Model, the estimation of $\gamma_i^1, \gamma_i^2, \phi_{jk}$ and are identical to the Normal model with the only difference being that P_{jk} is given as,

$$\begin{aligned}&P_{jk} = \tfrac{1}{2}\sum_{j=1}^{M}\sum_{i=1}^{N_j}\phi_{jk}\{log\left(\tfrac{1}{2\pi}\right) + \\ &\sum_{t=1}^{T} Y_{jit}\left(m_{k0t} + \sum_{d=1}^{D} X_{jid}m_{kdt}\right)\\ &m_{kdt} = m_0 s_0^2 + s_{kdt}^2 \sum_{j=1}^{M}\phi_{jk}\sum_{j=1}^{N_j} Y_{jit}X_{jid}, \quad s_{kdt}^2 = s_0^2 + \\ &\sum_{j=1}^{M}\phi_{jk}\sum_{j=1}^{N_j}\left(\sum_{d=0}^{D} X_{jid}^2 \exp\left(\sum_{d=0}^{D} X_{jid}m_{kdt}\right)\right)\end{aligned} \tag{11}$$

5.3 Parameter Estimation of Poisson iMG-GLM-1 Model

Again, in the Poisson Model, estimation of $\gamma_i^1, \gamma_i^2, \phi_{jk}$, are similar to the Normal model with the only difference being that the term P_{jk} is given as,

$$\begin{aligned}&P_{jk} = \tfrac{1}{2}\sum_{j=1}^{M}\sum_{i=1}^{N_j}\phi_{jk}\{-\sum_{d=0}^{D}\exp\left(\tfrac{s^{kd}}{2} + \tfrac{m_{kd}X_{jid}}{s^{kd}}\right) + \\ &Y_{ji}\left(\sum_{d=0}^{D} X_{jid}m_{kd}\right) - \log\left(Y_{ji}\right)\\ &\tfrac{m_{kd}}{s_{kd}^2} + \exp\left(m_{kd}\right) + \sum_{j=1}^{M}\phi_{jk}\sum_{i=1}^{N_j}\tfrac{X_{jid}}{s_{kd}^2}\\ &= \sum_{j=1}^{M}\sum_{i=1} N_j\phi_{jk}Y_{ji}X_{jid}\end{aligned} \tag{12}$$

For, m_{kd} and s_{kd}, does not have a close form solution. However, it can be solved quickly via any iterative root-finding method.

5.4 Predictive Distribution

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

$$\begin{aligned}&p\left(Y_{j,new}|X_{j,new}, Z_j, \beta_{k=1:K,d=0:D}\right) = \\ &\sum_{k=1}^{K}\int Z_{jk}p\left(Y_{j,new}|X_{j,new}, \beta_{k,d=0}^{D}\right) q\left(z, v, \beta_{kd}, \lambda_k\right)\end{aligned} \tag{13}$$

Table 1. Algorithm: Variational Inference Algorithm for iMG-GLM-1 Normal Model.

```
1. Initialize Generative Model Latent Parameters $q(z, v, \beta_{kd}, \lambda_k)$ Randomly in its State Space.
Repeat
2. Estimate $\gamma_k^1$ and $\gamma_k^2$ according to Eq.5.10. for $k = 1$ to $K$.
3. Estimate $\phi_{jk}$ according to Eq.5.10. for $j = 1$ to $M$ and for $k = 1$ to $K$.
4. Estimate the model density parameters, $\{m_{kd}, \beta_k, a_k, b_k\}$ according to Eq.5.10. for $k = 1$ to $K$ and $d = 0$ to $D$.
until converged
6. Evaluate $E[Y_{j,new}]$ for a new covariate, $X_{j,new}$, according to Eq.5.14 and Eq.5.15.
```

Integrating out the $q\,(z, v, \beta_{kd}, \lambda_k)$, we get the following equation for the Normal model.

$$\begin{aligned} &p\,(Y_{j,new}|X_{j,new}) = \\ &\textstyle\sum_{k=1}^{K} \phi_{jk} St\left(Y_{j,new}|\left(\sum_{d=0}^{D} m_{kd} X_{j,new,,d}, L_k, B_k\right)\right) \end{aligned} \tag{14}$$

Here, $L_k = \frac{(2a_k - D)\beta_k}{2(1+\beta_k)b_k}$, which is the precision parameter of the Student's t-distribution and $B_i = 2a_{y,i} - D$ is the degrees of freedom. For the Poisson and Multinomial Models, the integration of the densities is not analytically tractable. Therefore, we use Monte Carlo integration to obtain,

$$\begin{aligned} &E\,[Y_{j,new}|\mathbf{X_{j,new}}, \mathbf{X}, \mathbf{Y}] = E\,[E\,[Y_{j,new}|\mathbf{X_{j,new}}, \mathbf{q}\,(\beta_{\mathbf{kd}})]\,|\mathbf{X}, \mathbf{Y}] \\ &= \tfrac{1}{S} \textstyle\sum_{s=1}^{S} E\,[Y_{j,new}|\mathbf{X_{j,new}}, \mathbf{q}\,(\beta_{\mathbf{kd}})] \end{aligned} \tag{15}$$

In all experiments presented in this paper, we collected 100 i.i.d. samples (S=100) from the density of β to evaluate the expected value of $Y_{j,new}$. The complete Variational Inference Algorithm for iMG-GLM-1 Normal Model is given Table 1.

6 iMG-GLM-2 Model

We can now learn a new group $M + 1$, after all of the first M groups have been trained. For this process, we memorize the learned latent parameters from the previously learned data.

6.1 Information Transfer From Prior Groups

First, we write down the latent parameter conditional distribution given all the parameters in the previous groups. We define the set of latent parameters (Z, v, β, λ) as η. From the description of Dirichlet Process we write down the probability for the latent parameters for the $(M + 1)^{th}$ group given previous ones,

$$p\,(\eta_{M+1}|\eta_{1:M}, \alpha, G_0) = \tfrac{\alpha}{M+\alpha} G_0 + \tfrac{1}{M+\alpha} \textstyle\sum_{k=1}^{K} n_k \delta_{\eta_k^*} \tag{16}$$

Where, $n_k = \sum_{j=1}^{M} Z_{jk}$, represents count where $\eta_j = \eta_k^*$. If we substitute $\eta_k^* = E\,[\eta_k^*]$, which we define by $\Omega = \{\phi_{jk}, \gamma_k, m_{dk}, \lambda_k, s_{dk}\}$, we get,

$$p\,(\eta_{M+1}|\eta_k^*, \alpha, G_0) = \tfrac{\alpha}{M+\alpha} G_0 + \tfrac{1}{M+\alpha} \textstyle\sum_{k=1}^{K} n_k \delta_{\eta_k^*} \tag{17}$$

Where, $n_k = \sum_{j=1}^{M} index_{jk}$ and $index_{jk} = \delta_{argmax(\phi_{jk})}$. This distribution represents the prior belief about the new group latent parameters in the Bayesian setting. Now our goal is to compute the posterior distribution of the new group latent parameters after we view the likelihood with the data in $(M+1)^{th}$ group.

$$p\,(\eta_{M+1}|\Omega, \alpha, D_{M+1}) = \frac{p\left(D_{M+1}|\eta_{M+1}\right) p\left(\eta_{M+1}|\Omega, G_0\right)}{p\left(D_{M+1}|\Omega, G_0\right)} \tag{18}$$

Here, $p\,(D_{M+1}|\eta_{M+1}) = \Pi_{i=1}^{N_{M+1}} p\,(Y_{M+1,i}|\eta_{M+1}, X_{M+1,i})$.

6.2 Posterior Sampling

The posterior of Eq. 6.18 does not have a closed form solution apart from the Normal Model. So, we apply a Metropolis Hastings Algorithm [18,20] for the Logistic Multinomial and Poisson Model. For the Normal model, $p\left(\eta_{M+1}|\Omega, \alpha, D_{M+1}\right)$ turns out to be a mixture of Normal-Gamma density, $Normal - Gamma\left(\eta_{M+1}|m_k^{'}, \beta_k^{'}, a_k^{'}, b_k^{'}\right)$ with following parameters,

$$\begin{array}{l} m_k^{'} = \left\{X_{M+1}^T X_{M+1} + (\beta_k)\, I\right\}^{-1} \left\{X_{M+1}^T Y_{M+1} + \beta_k I m_k\right\} \\ \beta_k^{'} = \left(X_{M+1}^T X_{M+1} + \beta_k I\right), \quad a_k^{'} = a_k + N_{M+1}/2 \\ b_k^{'} = b_k + \frac{1}{2}\left\{Y_{M+1}^T Y_{M+1} + m_k^T \beta_k m_k - m_k^{'T} \beta_k^{'} m_k^{'}\right\} \end{array} \tag{19}$$

For the Poisson and Logistic Multinomial Model, The Metropolis Hastings Algorithm has the following steps. First, we draw a sample $\acute{\eta}$ from Eq. 6.17. Then we draw a candidate sample η, Next, we compute the acceptance probability, $\left[min\left[1, \frac{p(D_{M+1}|\eta)}{p(D_{M+1}|\acute{\eta})}\right]\right]$. We set the new $\acute{\eta}$ to η with this acceptance probability. Otherwise, it remains the old value. We repeat the above 4 steps until enough samples has been collected. This yields the approximation of the posterior.

6.3 Prediction for New Group Test Samples

We seek to predict the future $Y_{M+1,new}|X_{M+1,new}, \Omega$, by the following equation with the previous collection of posterior samples $\eta_{t=1:T}$. T is the number of samples.

$$\begin{array}{l} p\left(Y_{M+1,new}|X_{M+1,new}, \Omega\right) \\ = \frac{1}{T}\sum_{t=1}^{T} p\left(Y_{M+1,new}|X_{M+1,new}, \eta_t\right) \end{array} \tag{20}$$

7 Experimental Results

We present empirical studies on two realworld applications: (a) a Stock Market Accuracy and Trend Detection problem and (b) a Clinical Trial problem on the efficacy of a new drug.

7.1 Trends in Stock Market

We propose iMG-GLM-1 and iMG-GLM-2 as a trend spotter in Financial Markets where we have chosen daily close out stock prices over 51 stocks from NYSE and Nasdaq in various sectors, such as, Financials (BAC, WFC, JPM, GS, MS, Citi, BRK-B, AXP), Technology (AAPl, MSFT, FB, GOOG, CSCO, IBM, VZ), Consumer Discretionary (AMZN, DIS, HD, MCD, SBUX, NKE, LOW), Energy (XOM, CVX, SLB, KMI, EOG), Health Care (JNJ, PFE, GILD, MRK, UNH, AMGN, AGN), Industrials (GE, MMM, BA, UNP, HON, UTX, UPS), Materials (DOW, DD, MON, LYB) and Consumer Staples (PG, KO, PEP, PM, CVS, WMT). The task is to predict future stock prices given past stock value for all these stocks and spot general trends in the cluster of the stocks which might be

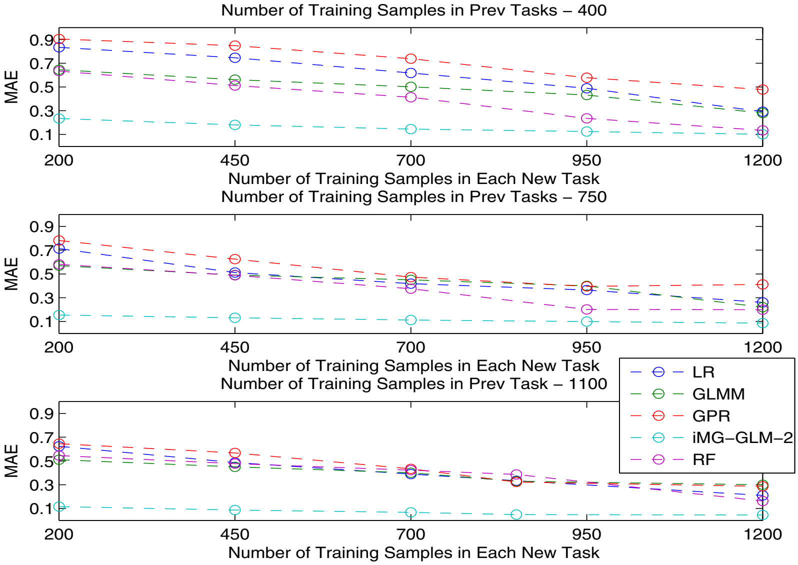

Fig. 2. The Average Mean Absolute Error for 10 New Stocks for 50 random runs for iMG-GLM-2 Model with varying number of training samples in both previous and New Groups

helpful in finding a far more powerful model for prediction. The general setting is a auto-regressive process via the Normal iMG-GLM-1 model with lags representing the predictor variables and response being the current stock price. The lag-length was determined to be 3 by trial and error with 50-50 training-testing split. Data was collected from September 13th, 2010 to September 13th, 2015 with 1250 data points, from Google Finance.

Some very interesting trends were noteworthy. After the clustering was accomplished for the Normal model, the stocks became grouped almost entirely by the sectors they came from. Specifically, we witnessed a total of 9 clusters of stocks, close in makeup to the 8 sectors chosen originally consolidating all the stocks sectors such as, financial, healthcare etc. For example, Apple, Microsoft Verizon, Google, Cisco and AMZN were clubbed together in one cluster. This signifies that all of these stocks share the same auto-regressive density with the same variance. In comparison, single and separate modeling of the stocks resulted in a much inferior model. Joint modeling was particularly useful because we had only 625 data points per stocks for training purposes over the past 5 years. As a result, transfer of stock data points from one stock to another helped mitigate the problem of over-fitting the individual stocks while ensuring a much improved model for density estimation for a cluster of stocks. We report the clustering of the stocks in Table 2. We also show the accuracy of the prediction for the iMG-GLM-1 model in terms of the Mean Absolute error (MAE) in Table 3. Note that MAE for the Normal model significantly outperformed the GLMM normal model, stock specific Random Forest, Linear Regression and Gaussian Process Regression.

We now highlight the utilization of information transfer in the iMG-GLM-1 model. We trained the first 51 stocks where we varied the number of training samples in each group/stock from 200 to 1200 in steps of 250. For each group we chose the training samples randomly from the datasets and the remaining were used for testing. The hyper-parameters were set as, $\{m_0, \beta_0, a_0, b_0\} = 0, 1, 2, 2$. We also ran our inference with different settings of the hyper-parameters but found the results not to be particularly sensitive to the hyper-parameters settings. We plot the average MAE of 50 random runs in Figure 3. The iMG-GLM-1 Normal Model generally outperformed the other competitors. Few interesting results were found in this experiment. When very few training samples were used for training, virtually all the algorithms performed poorly. In particular, iMG-GLM-1 clubbed all stocks into one cluster as sufficient data was not present to identify the statistical similarities between stocks. As the number of training samples increased iMG-GLM-1 started to pick out cluster of groups/stocks as it was able find latent common densities among different groups. As, the training samples got closer to the number of data points (1200), all other models started to perform close to the iMG-GLM-1 model, because they managed to learn each stock well in isolation, indicating that further data from other groups became less useful.

We now proceed to iMG-GLM-2, where we trained 10 new stocks from different sectors (CMCSA, PCLN, WBA, COST, KMI, AIG, GS, HON, LMT, T). Two features which influenced the learning were considered. First, we varied the number of training samples from 400 to 750 to 1100 for each previous groups that were used to further train β_{M+1}. Then, we changed the number of training samples for the new groups from 200 to 1200 in steps of 250. We plot the MAE results for 50 random runs in Figure 2. The prior belief is that the new groups are similar in response density to the previous groups. iMG-GLM-2 efficiently transfers this information from a previous groups to new groups. The iMG-GLM-1 model learns an informative prior for new groups when the number of training samples for each previous group is very small (as seen in the first part in Figure 2). The accuracy increases very slightly as the number of training samples increases in each group. But, with the number of training samples for the new groups increasing, iMG-GLM-2 does not improve at all. This is due to the flexible information transfer from the previous groups. The model does not require more training samples for its own group to model its density, because it has already obtained sufficient information as prior from the previous groups.

7.2 Clinical Trial Problem Modeled by Poisson iMG-GLM Model

Finally, we explored a Clinical Trial problem [12] for testing whether a new anticonvulsant drug reduces a patient's rate of epileptic seizures. Patients were assigned the new drug or the placebo and the number of seizures were recorded over a six

Table 2. Clusters of Stocks from Various Sectors. We note 9 clusters of stocks consolidating all the pre-chosen sectors such as, financials, materials etc.

Group No.	1	2	3	4	5	6	7	8	9
	AAPL, MSFT, VZ, GOOG, CSCO, AMZN	BAC, WFC, JPM, AXP, PG, CITI, GS,MS	DIS, HD, LOW, SBUX, MCD	XOM, CVX, SLB, EOG, KMI	GILD, MRK, UNH, AMGN, AGN	GE, MMM, BA, UNP, HON	DOW, DD, MON, LYB, JNJ, PFE	KO, PEP, PM, CVS, WMT	BRK-B, IBM, FB, NKE, UTX, UPS

Table 3. Mean Absolute Error (MAE) for All Stocks. iMG-GLM has Much Higher Accuracy than Other Competitors.

	AAPL	MSFT	VZ	GOOG	CSCO	AMZN	BAC	WFC	JPM	AXP	PG	CITI	GS	MS	DIS	HD	LOW
GPR	.023	.004	.087	.078	.093	.189	.452	.265	.176	.190	.378	.018	.037	.098	.278	.038	.011
RF	.278	.903	.370	.256	.290	.570	.159	.262	.329	.592	.746	.894	.956	.239	.934	.189	.045
LR	.381	.865	.280	.038	.801	.706	.589	.491	.391	.467	.135	.728	.578	.891	.389	.790	.624
GLMM	.378	.489	.389	.208	.972	.786	.289	.768	.189	.389	.590	.673	.901	.490	.209	.391	.991
iMG-GLM	.012	.002	.009	.011	.018	.028	.047	.038	.035	.079	.069	.087	.019	.030	.139	.189	.213
	SBUX	MCD	XOM	CVX	SLB	EOG	KMI	GILD	MRK	UNH	AMGN	AGN	GE	MMM	BA	UNP	HON
GPR	.837	.289	.849	.583	.185	.810	.473	.362	.539	.289	.306	.438	.769	.848	.940	.829	.691
RF	.884	.321	.895	.843	.774	.863	.973	.729	.894	.794	.695	.549	.603	.738	.481	.482	.482
LR	.380	.391	.940	.995	.175	.398	.539	.786	.591	.320	.793	.839	.991	.839	.698	.389	.298
GLMM	.649	.720	.364	.920	.529	.369	.837	.630	.729	.481	.289	.970	.740	.649	.375	.439	.539
iMG-GLM	.003	.018	.128	.291	.005	.060	.052	.017	.014	.078	.009	.067	.191	.034	.098	.145	.238
	DOW	DD	LYB	JNJ	PFE	KO	PEP	PM	CVS	WMT	BRK-B	IBM	FB	NKE	UTX	UPS	MON
GPR	.689	.890	.745	.907	.678	.378	.867	.945	.361	.934	.589	.845	.901	.310	.483	.828	.748
RF	.181	.098	.489	.237	.692	.827	.490	.295	.749	.692	.957	.295	.478	.694	.747	.806	.945
LR	.67	.386	.984	.982	.749	.294	.256	.567	.345	.767	.893	.956	.294	.389	.694	.921	.702
GLMM	.727	.389	.288	.592	.402	.734	.923	.900	.571	.312	.839	.956	.638	.490	.390	.372	.512
iMG-GLM	.038	.078	.063	.019	.024	.007	.089	.192	.138	.111	.289	.390	.289	.218	.200	.149	.087

week period. A measurement was made before the trial as a baseline. The objective was to model the number of seizures, which being a count datum, is modeled using a Poisson distribution with a Log link. The covariates are: Treatment Center size (ordinal), number of weeks of treatment (ordinal), type of treatment–new drug or placebo (nominal) and gender (nominal). A Poisson distribution with log link was used for the count of seizures. Here, X_{ji} and Y_{ji} represent the i^{th} covariate and count response pair of the j^{th} group. The distribution, $\{\beta_{kd}\}$ (Normal) is the prior distribution on the covariate coefficient β.

We found that a patient's number of seizures are clustered (they form the groups) in multiple collections. This signifies that a majority of the patients across groups show the same response to the treatment. We obtained 8 clusters from 300 out of 565 patients for the iMG-GLM-1 model (the remaining 265 were set aside for modeling through the iMG-GLM-2 model). Among them 5 clusters showed that the new drug reduces the number of epileptic seizures with increasing number of weeks of treatment while the remaining 3 clusters did not show any improvement. We also report the forecast error of the number of epileptic seizures of the remaining 265 patients in Table 4. Our recommendation for the usage of the new drug would be a cluster based solution. For a specific patient, if she falls in one of those clusters with decreasing trend in the number of seizures

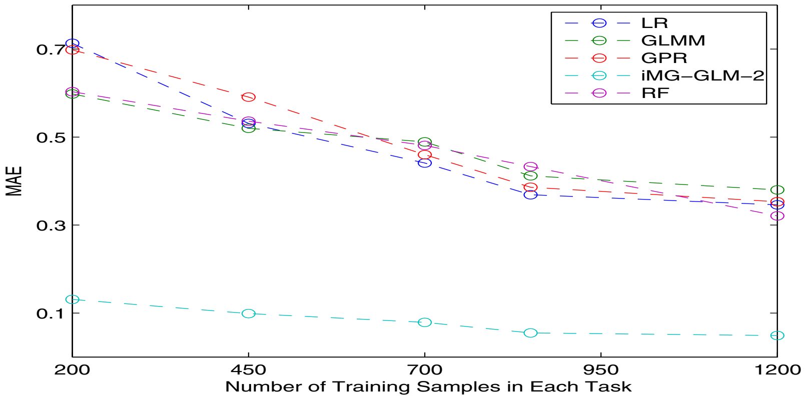

Fig. 3. The Average Mean Absolute Error for 51 Stocks for 50 random runs for iMG-GLM-1 Model with varying number of training samples.

Table 4. MSE and MAE of the Algorithms for the Clinical Trial Dataset and Number of Patients in Clusters for iMG-GLM-1 and iMG-GLM-2 Model.

<table>
<tr><td colspan="8">Patient Number in Clusters for iMG-GLM-1 Model</td></tr>
<tr><td colspan="5">Positive</td><td colspan="3">Negative</td></tr>
<tr><td>46</td><td>30</td><td>40</td><td>27</td><td>33</td><td>24</td><td>37</td><td>24</td></tr>
<tr><td colspan="8">Patient Number in Clusters for iMG-GLM-2 Model</td></tr>
<tr><td colspan="5">Positive</td><td colspan="3">Negative</td></tr>
<tr><td>33</td><td>24</td><td>41</td><td>29</td><td>30</td><td>31</td><td>34</td><td>43</td></tr>
<tr><td colspan="2">iMG-GLM</td><td colspan="2">Poisson GLMM</td><td colspan="2">Poisson Regression</td><td colspan="2">RForest</td></tr>
<tr><td colspan="8">Mean Square Root Error(L2 Error) fpr iMG-GLM-2 Model</td></tr>
<tr><td colspan="2">1.53</td><td colspan="2">1.58</td><td colspan="2">1.92</td><td colspan="2">1.75</td></tr>
<tr><td colspan="8">Mean Absolute Error Root Error(L1 Error) for iMG-GLM-2 Model</td></tr>
<tr><td colspan="2">1.14</td><td colspan="2">1.34</td><td colspan="2">1.51</td><td colspan="2">1.62</td></tr>
</table>

with time, we would recommend the new drug, and otherwise not. Out of 265 test case patients modeled through iMG-GLM-2, 180 showed signs of improvements while 85 did not. We kept all the weeks as training for the iMG-GLM-1 model and the first five weeks as training and the last week as testing data for the iMG-GLM-2 model. Traditional Poisson GLMM cannot infer these findings since the densities are not shared at the patient group level. Moreover, only the Poisson iMG-GLM-1/2 based prediction is formally equipped to recommend a patient cluster based solution for the new drug, whereas all traditional mixed models predict a global recommendation for all patients.

8 Conclusion

In this paper, we have formulated an infinite multigroup Generalized Linear Model (iMG-GLM), a flexible model for shared learning among groups in grouped regression. The model clusters groups by identifying identical response-covariate densities for different groups. It also models heteroscedasticity among groups by modeling different uncertainty among groups. We experimentally evaluated the model on a wide range of problems where traditional mixed effect models and group specific regression models fail to capture structure in the grouped data. Although the Metropolis Hastings algorithm turned out to be fairly accurate for the iMG-GLM-2 model, developing a variational inference alternative would be an interesting topic for future research. Finally, the number of groups in each cluster depends on the scale factors α_1 and α_2 (scale parameters of the DP) of the model, and at times grows large in specific cluster. This occurs mostly when any cluster has a large number of groups which becomes representative of the whole data. In most cases, beyond a few primary clusters, the remaining clusters represent outliers. Although, careful tuning of scale parameters can mitigate these problems, a theoretical understanding of the dependence of the model on scale parameters could lead to better modeling and application.

References

1. Ando, R.K., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research **6**, 1817–1853 (2005)
2. Antoniak, C.: Mixtures of dirichlet processes with applications to bayesian nonparametric problems. Annals of Statistics **2**(6), 1152–1174 (1974)
3. Bakker, B., Heskes, T.: Task clustering and gating for bayesian multitask learning. Journal of Machine Learning Research **4**, 83–99 (2003)
4. Baxter, J.: Learning internal representations. In: International Conference on Computational Learning Theory, pp. 311–320 (1995)
5. Baxter, J.: A model of inductive bias learning. Journal of Artificial Intelligence Research **12**, 149–198 (2000)
6. Blei, D.: Variational inference for dirichlet process mixtures. Bayesian Analysis **1**, 121–144 (2006)
7. Breslow, N.E., Clayton, D.G.: Approximate inference in generalized linear mixed models. Journal of the American Statistical Association **88**(421), 9–25 (1993)
8. Caruana, R.: Multitask learning. Machine Learning **28**(1), 41–75 (1997)
9. Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association **90**, 577–588 (1994)
10. Ferguson, T.: A bayesian analysis of some nonparametric problems. Annals of Statistics **1**, 209–230 (1973)
11. Hastie, T., Tibshirani, R.: Varying-coefficient models. Journal of the Royal Statistical Society. Series B (Methodological) **55**(4), 757–796 (1993)
12. IBM: Ibm spss version 20. IBM SPSS SOFTWARE (2011)
13. Jordan, M., Jacobs, R.: Hierarchical mixtures of experts and the EM algorithm. In: International Joint Conference on Neural Networks (1993)
14. Lee, Y., Nelder, J.A.: Hierarchical generalized linear models. Journal of the Royal Statistical Society. Series B (Methodological) **58**(4), 619–678 (1996)
15. Lee, Y., Nelder, J.A.: Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. Biometrika **88**(4), 987–1006 (2001)

16. Lee, Y., Nelder, J.A.: Modelling and analysing correlated non-normal data. Statistical Modelling **1**(1), 3–16 (2001)
17. Lee, Y., Nelder, J.A.: Double hierarchical generalized linear models (with discussion). Journal of the Royal Statistical Society: Series C (Applied Statistics) **55**(2), 139–185 (2006)
18. Neal, R.M.: Markov chain sampling methods for dirichlet process mixture models. Journal of Computational and Graphical Statistics **9**(2), 249–265 (2000)
19. Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. Journal of the Royal Statistical Society, Series A (General) **135**(3), 370–384 (1972)
20. Robert, C., Casella, G.: Monte Carlo Statistical Methods (Springer Texts in Statistics). Springer-Verlag New York, Inc. (2005)
21. Sethuraman, J.: A constructive definition of dirichlet priors. Statistica Sinica **4**, 639–650 (1994)
22. Viele, K., Tong, B.: Modeling with mixtures of linear regressions. Statistics and Computing **12**(4), 315–330 (2002)
23. Xue, Y., Liao, X., Carin, L.: Multi-task learning for classification with dirichlet process priors. Journal of Machine Learning Research **8**, 35–63 (2007)
24. Yu, K., Tresp, V., Schwaighofer, A.: Learning gaussian processes from multiple tasks. International Conference on Machine Learning, pp. 1012–1019 (2005)
25. Zhang, J., Ghahramani, Z., Yang, Y.: Learning multiple related tasks using latent independent component analysis. Advances in Neural Information Processing Systems, pp. 1585–1592 (2005)