A Stick-Breaking Likelihood for Categorical Data Analysis with Latent Gaussian Models

Mohammad Emtiyaz Khan[1], **Shakir Mohamed**[1], **Benjamin M. Marlin**[2] **and Kevin P. Murphy**[1]
[1]Department of Computer Science, University of British Columbia, Vancouver, Canada
[2]Department of Computer Science, University of Massachusetts, Amherst, USA

Abstract

The development of accurate models and efficient algorithms for the analysis of multivariate categorical data are important and long-standing problems in machine learning and computational statistics. In this paper, we focus on modeling categorical data using Latent Gaussian Models (LGMs). We propose a novel stick-breaking likelihood function for categorical LGMs that exploits accurate linear and quadratic bounds on the logistic log-partition function, leading to an effective variational inference and learning framework. We thoroughly compare our approach to existing algorithms for multinomial logit/probit likelihoods on several problems, including inference in multinomial Gaussian process classification and learning in latent factor models. Our extensive comparisons demonstrate that our stick-breaking model effectively captures correlation in discrete data and is well suited for the analysis of categorical data.

1 Introduction

The development of accurate models and efficient learning and inference algorithms for high-dimensional, correlated, multivariate categorical data are important and long-standing problems in machine learning and computational statistics. They have applications for data analysis in a wide variety of areas such as discrete choice modeling in econometrics, analysis of survey responses in social science, medical diagnostics, and recommender systems.

In this paper, we focus on the class of Latent Gaussian Models (LGMs), which model data distributions using Gaussian latent variables. LGMs include popular models such as factor analysis and probabilistic principal components analysis for continuous data (Knott and Bartholomew, 1999; Tipping and Bishop, 1999), binary and multinomial factor analysis for discrete data (Wedel and Kamakura, 2001; Collins et al., 2002; Mohamed et al., 2008; Khan et al., 2010), and Gaussian process regression and classification (Nickisch and Rasmussen, 2008). LGMs allow for a principled handling of missing data and can be used for dimensionality reduction, data prediction and visualization.

In the case of LGMs for categorical data, the two most widely used likelihoods are the multinomial-probit, and the multinomial-logit or softmax. The key difficulty with the LGM model class is that the latent variables must be integrated away in order to obtain the marginal likelihood needed to learn the model parameters. This integration can be performed analytically in Gaussian-likelihood LGMs such as factor analysis because the model is jointly Gaussian in the latent factors and the observed variables (Bishop, 2006). LGMs with logit and probit-based likelihoods lack this property, resulting in intractable integrals for the marginal likelihood.

The main contribution of this paper is the development of a novel stick-breaking likelihood function for categorical data. The stick-breaking likelihood function is an alternative generalization of the binary logit likelihood to the case of categorical data. It is more amenable to the application of variational bounds than the traditional multinomial-logit construction, and is specifically designed to exploit recently proposed linear and quadratic bounds on the logistic-log-partition function developed by Marlin et al. (2011). These bounds are much more accurate than variational quadratic bounds used in previous work. We thoroughly compare our proposed framework to existing algorithms for both multinomial-probit and multinomial-logit likelihoods on several problems in-

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

cluding inference in multinomial Gaussian process classification and learning in categorical factor models. Our results demonstrate that the proposed stick-breaking model effectively captures correlation in discrete data and is well suited for the analysis of categorical data.

2 Categorical Latent Gaussian Models

For a generic latent Gaussian model, we consider N data instances, with a visible data vector $\mathbf{y}_n$ and corresponding latent vector by $\mathbf{z}_n$, for the n'th observation. In general, $\mathbf{y}_n$ and $\mathbf{z}_n$ will have dimensions D and L, respectively. Each element of $\mathbf{y}_n$, denoted by y_{dn}, can take values from a finite discrete set $S_d = \{C_0, C_1, C_2, \ldots, C_{K_d}\}$ where C_k is the k'th category. For simplicity, we assume that $K_d = K, \forall d$.

In LGMs, the latent variables $\mathbf{z}_n$ follow a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ as shown in Eq. 1. The probability of each categorical variable y_{dn} is parameterized in terms of the linear projection $\boldsymbol{\eta}_{dn}$ as seen in Eqs. 2 and 3. Here, $\mathbf{W}_d \in \mathbb{R}^{(K+1)\times L}$ is the factor loading matrix and $\mathbf{w}_{0d} \in \mathbb{R}^{K+1}$ is the offset vector, implying that $\boldsymbol{\eta}_{dn} \in \mathbb{R}^{K+1}$. The likelihood in Eq. 3 factorizes over data dimensions d. We consider the problem of choosing a parameterization for each likelihood term $p(y_{dn}|\boldsymbol{\eta}_{dn})$ in the next section.

$$p(\mathbf{z}_n|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

$$\boldsymbol{\eta}_{dn} = \mathbf{W}_d \mathbf{z}_n + \mathbf{w}_{0d} \tag{2}$$

$$p(\mathbf{y}_n|\mathbf{z}_n) = \prod_{d=1}^{D} p(y_{dn}|\boldsymbol{\eta}_{dn}). \tag{3}$$

We denote the set of parameters by $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{W}, \mathbf{w}_0\}$, where $\mathbf{W}$ and $\mathbf{w}_0$ are the sets containing $\mathbf{W}_d$ and $\mathbf{w}_{0d}$ for all dimensions. To make the model identifiable, we set the last row of $\mathbf{W}_d$ and last element of $\mathbf{w}_{0d}$ to zero. Also, the prior mean $\boldsymbol{\mu}$ and the offset $\mathbf{w}_0$ are interchangeable in all the models so we use the mean only.

Different models for categorical data can be obtained by restricting the above generic model in different ways. We obtain the categorical factor analysis (cFA) model by assuming that $L \leq DK$ and $\boldsymbol{\Sigma}$ is the identity matrix, while $\mathbf{W}$ and $\boldsymbol{\mu}$ are unrestricted. Conversely, we obtain the categorical latent Gaussian graphical model (cLGGM) by assuming that $L = DK$ and $\mathbf{W}$ is the identity matrix, while $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unrestricted. We obtain a multi-class Gaussian process classification (mGPC) model by restricting $N = 1$ and $\mathbf{W}$ to be the identity matrix, and specifying $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ using a kernel function that depends on features. In the mGPC case, D is the number of data points and the set of parameters consists of the hyperparameters of the mean and covariance function.

3 Categorical Parameterizations

There are many choices available for the categorical distribution $p(y|\boldsymbol{\eta})$ in Eq. 3. We review two popular parameterizations: the multinomial-probit and multinomial-logit, and then introduce a new *stick-breaking* parameterization and discuss its properties.

The form for the multinomial-probit function is given in Eq. 8 and makes use of auxiliary variables $u_j \sim \mathcal{N}(u_j|\eta_j, 1)$. The probability of each category is defined through an integral over the region R_k where auxiliary variable $u_k > u_j$ for all $j \neq k$.

$$p(y = C_k|\boldsymbol{\eta}) = \int_{R_k} \prod_{j=0}^{K} p(u_j|\eta_j) d\mathbf{u} \tag{8}$$

$$p(y = C_k|\boldsymbol{\eta}) = \frac{e^{\eta_k}}{\sum_{j=0}^{K} e^{\eta_j}} = \exp\left[\eta_k - \mathrm{lse}(\boldsymbol{\eta})\right] \tag{9}$$

The form of the multinomial-logit function is given in Eq. 9. It is defined using the log-sum-exp (LSE) function $\mathrm{lse}(\boldsymbol{\eta}) = \log \sum_j \exp(\eta_j)$, and is the natural generalization of the binary logit function to three or more categories.

The standard multinomial-logit construction is not the only way to extend the logit function to the case of multiple categories. We propose an alternative generalization of the logit function, which we refer to as the *stick-breaking* parameterization. The stick-breaking process is part of the more general framework of random allocation processes, and is very closely associated with Bayesian non-parametric methods, where it is used in constructive definitions of the Dirichlet process, for example Sethuraman (1994). In our stick-breaking parameterization, we use a logit function to model the probability of the first category as $\sigma(\eta_0)$ where $\sigma(x) = 1/(1 + \exp(-x))$. This is the first piece of the stick. The length of the remainder of the stick is $(1 - \sigma(\eta_0))$. We can model the probability of the second category as a fraction $\sigma(\eta_1)$ of the remainder of the stick left after removing $\sigma(\eta_0)$. We can continue in this way until we have defined all the stick lengths up to K. The last category then receives the remaining stick length, as seen below.

$$\begin{aligned}
p(y = C_0|\boldsymbol{\eta}) &= \sigma(\eta_0) \\
p(y = C_k|\boldsymbol{\eta}) &= \prod_{j \leq k-1} (1 - \sigma(\eta_j))\sigma(\eta_k), 0 < k < K \\
p(y = C_K|\boldsymbol{\eta}) &= \prod_{j=1}^{K-1} (1 - \sigma(\eta_j))
\end{aligned} \tag{10}$$

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \int_{\mathbf{z}} p(\mathbf{z}|\boldsymbol{\theta}) p(\mathbf{y}_n|\mathbf{z}) d\mathbf{z} = \sum_{n=1}^{N} \log \int_{\mathbf{z}} q(\mathbf{z}|\boldsymbol{\gamma}_n) \frac{p(\mathbf{z}|\boldsymbol{\theta}) p(\mathbf{y}_n|\mathbf{z})}{q(\mathbf{z}|\boldsymbol{\gamma}_n)} d\mathbf{z} \tag{4}$$

$$\mathcal{L}(\boldsymbol{\theta}) \geq \mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma}) := \sum_{n=1}^{N} - \int_{\mathbf{z}} q(\mathbf{z}|\boldsymbol{\gamma}_n) \log \frac{q(\mathbf{z}|\boldsymbol{\gamma}_n)}{p(\mathbf{z}|\boldsymbol{\theta})} d\mathbf{z} + \int_{\mathbf{z}} q(\mathbf{z}|\boldsymbol{\gamma}_n) \log p(\mathbf{y}_n|\mathbf{z}) d\mathbf{z} \tag{5}$$

$$= \sum_{n=1}^{N} -D_{KL}\left[q(\mathbf{z}|\boldsymbol{\gamma}_n)||p(\mathbf{z}|\boldsymbol{\theta})\right] + \sum_{d=1}^{D} \mathbb{E}_{q(\boldsymbol{\eta}|\tilde{\boldsymbol{\gamma}}_{dn})}[\log p(y_{dn}|\boldsymbol{\eta})] \tag{6}$$

$$D_{KL}(q_n(\mathbf{z}|\boldsymbol{\gamma}_n)||p(\mathbf{z}|\boldsymbol{\theta})) = \frac{1}{2}\left[-\log|\mathbf{V}_n\boldsymbol{\Sigma}^{-1}| + \mathrm{tr}(\mathbf{V}_n\boldsymbol{\Sigma}^{-1}) + (\mathbf{m}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{m}_n - \boldsymbol{\mu}) - L\right] \tag{7}$$

The probabilities (stick lengths) are all positive and sum to one; they thus define a valid probability distribution. We can also use a different function for $\sigma(x)$ such as the probit function, but we use the logit function since it allows us to use efficient variational bounds. The stick-breaking parameterization can be written more compactly as shown in Eq. 11.

$$p(y = C_k|\boldsymbol{\eta}) = \exp[\eta_k - \sum_{j \leq k} \log(1 + e^{\eta_j})] \tag{11}$$

Models for multinomial *regression* have wide coverage in the statistics and psychology literature. Both the multinomial-probit and multinomial-logit links are used extensively (Albert and Chib, 1993; Holmes and Held, 2006; Vijverberg, 2000). These link functions do not assume any ordering of categories and it is understood that these parameterizations give similar performance and qualitative conclusions. On the other hand, inference with these link functions is difficult.

Our stick-breaking construction simplifies the inference by constructing a categorical likelihood using simpler binary likelihood functions as shown in Eq. 10. Each η_k can be interpreted as the log-ratio: $\eta_k = \log[p(y = C_k|\boldsymbol{\eta})/p(y > C_k|\boldsymbol{\eta})]$. This implies that, given a particular ordering of categories, each η_k defines a decision boundary in the latent space $\mathbf{z}$, that separates the k'th category from all categories $j > k$. If such a separation is difficult to attain given an ordering of categories, the stick-breaking likelihood may not give good predictions. In practice, such separability is easier to achieve in latent variable models such as ours. Our results on real-world datasets confirm this.

The stick-breaking parameterization also has important advantages over the multinomial-logit model in terms of variational approximations. The multinomial-logit parameterization requires bounding the $\mathrm{lse}(\boldsymbol{\eta})$ function and, at present, it is not known how to obtain tight bounds on this function with more than two categories (Bouchard, 2007; Khan et al., 2010). As we can see in Eq. 11, the stick-breaking parameterization only depends on functions of the form $\log(1+e^{\eta_j})$, known as the *logistic log-partition function*. In contrast to the multinomial-logit case, extremely accurate piecewise-linear and quadratic bounds are available for the logistic-log-partition function (Marlin et al., 2011).

4 Variational Learning and the Stick-Breaking Parameterization

Parameter estimation is always problematic in LGMs that use non-Gaussian likelihoods due to the fact that the marginal likelihood contains intractable integrals. In this section, we derive a tractable variational lower bound to the marginal likelihood for categorical LGMs using the stick-breaking parameterization, exploiting the availability of very tight bounds on the logistic-log-partition function.

We begin with the intractable log marginal likelihood $\mathcal{L}(\boldsymbol{\theta})$ in Eq. 4 and introduce a variational posterior distribution $q(\mathbf{z}|\boldsymbol{\gamma}_n)$ for each n. We use a Gaussian posterior with mean $\mathbf{m}_n$ and covariance $\mathbf{V}_n$. The full set of variational parameters is thus $\boldsymbol{\gamma}_n = \{\mathbf{m}_n, \mathbf{V}_n\}$ and we use $\boldsymbol{\gamma}$ to denote the set of all $\boldsymbol{\gamma}_n$.

As log is a concave function, we obtain a lower bound $\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})$ using Jensen's inequality, given by Eq. 5. The first term is simply the Kullback−Leibler (KL) divergence from the variational Gaussian posterior $q(\mathbf{z}|\mathbf{m}_n, \mathbf{V}_n)$ to the Gaussian prior distribution $p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and has a closed-form expression given by Eq. 7. In the second term, we substitute the likelihood definition from Eq. 3 and apply a change of variable from $\mathbf{z}$ to $\boldsymbol{\eta}$ to get Eq. 6. The new expectation is with respect to $q(\boldsymbol{\eta}|\tilde{\boldsymbol{\gamma}}_{dn})$, where $\tilde{\boldsymbol{\gamma}}_{dn} = \{\tilde{\mathbf{m}}_{dn}, \tilde{\mathbf{V}}_{dn}\}$, $\tilde{\mathbf{m}}_{dn} = \mathbf{W}_d\mathbf{m}_n + \mathbf{w}_{0d}$, and $\tilde{\mathbf{V}}_{dn} = \mathbf{W}_d\mathbf{V}_n\mathbf{W}_d^T$.

The lower bound $\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})$ is still intractable as the expectation of $\log p(y_{dn}|\boldsymbol{\eta})$ is not available in closed form. To derive a tractable lower bound, we make use of piecewise linear/quadratic bounds for this expectation. For simplicity, we suppress the dependence on d and n and consider the log-likelihood of a scalar observation y given a predictor $\boldsymbol{\eta} \sim q(\boldsymbol{\eta}|\tilde{\boldsymbol{\gamma}}) = \mathcal{N}(\boldsymbol{\eta}|\tilde{\mathbf{m}}, \tilde{\mathbf{V}})$

with $\tilde{\boldsymbol{\gamma}} = \{\tilde{\mathbf{m}}, \tilde{\mathbf{V}}\}$. The log likelihood is shown in Eq. 12. We see that the expectation of this term with respect to a Gaussian distribution is intractable due to the presence of $\log(1 + \exp(\eta_j))$ terms. We use the piecewise linear/quadratic upper bound of Marlin et al. (2011) to obtain a lower bound to the log-likelihood in Eq. 13.

$$\log p(y = C_k|\boldsymbol{\eta}) = \eta_k - \sum_{j \leq k} \log(1 + e^{\eta_j}) \tag{12}$$

$$\geq \eta_k - \sum_{j \leq k} \sum_{r=1}^{R} \mathbb{I}_{(t_{r-1}, t_r)}(\eta_j)[a_r \eta_j^2 + b_r \eta_j + c_r] \tag{13}$$

Here a_r, b_r, c_r are the parameters of r'th quadratic piece in the interval (t_{r-1}, t_r), R is the total number of pieces, and $\mathbb{I}_{(t_{r-1},t_r)}(x) = 1$ if $x \in (t_{r-1}, t_r)$ and 0 otherwise. We denote the parameters of the piecewise bound by $\boldsymbol{\alpha}$ which can be computed beforehand by solving an optimization problem as shown by Marlin et al. (2011). The expectation of the lower bound with respect to a Gaussian is tractable as the expectation of each piece is just the expectation with respect to a truncated Gaussian distribution as shown in Eq. 14. We denote this lower bound by $B(y, \tilde{\boldsymbol{\gamma}}, \boldsymbol{\alpha})$. Note that the bound only depends on the diagonal elements of $\tilde{\mathbf{V}}$. We denote these by $\tilde{\mathbf{v}}$.

$$\mathbb{E}_{q(\boldsymbol{\eta}|\tilde{\boldsymbol{\gamma}})}[\log p(y = C_k|\boldsymbol{\eta})] \geq B(y, \tilde{\boldsymbol{\gamma}}, \boldsymbol{\alpha}) \tag{14}$$

$$:= \tilde{m}_k - \sum_{j \leq k} \sum_{r=1}^{R} \int_{t_{r-1}}^{t_r} (a_r x^2 + b_r x + c_r) \mathcal{N}(x|\tilde{m}_j, \tilde{v}_j) dx$$

An important property of the piecewise bound is that its maximum error is bounded and can be driven to zero by increasing the number of pieces. This means that the lower bound in Eq. 14 can be made arbitrarily tight by increasing the number of pieces.

4.1 A Generalized EM Algorithm

We substitute Eq. 14 into Eq. 6 to obtain a final, tractable lower bound on $\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})$, which we denote by $\underline{\mathcal{L}}_J(\boldsymbol{\theta}, \boldsymbol{\gamma})$. To learn the parameters, we optimize the lower bound with respect to the variational posterior parameters $\boldsymbol{\gamma}$ and model parameters $\boldsymbol{\theta}$. Some of the updates are not available in closed form and require numerical optimization, resulting in a generalized expectation-maximization algorithm. The generalized E-step requires numerically optimizing the variational posterior means and covariances. The generalized M-step consists of a mix of closed-form updates and numerical optimization. To derive the required gradients, we need the gradient of $B(y_{dn}, \tilde{\boldsymbol{\gamma}}_{dn}, \boldsymbol{\alpha})$ with respect to $\tilde{\mathbf{m}}_{dn}$ and $\tilde{\mathbf{v}}_{dn}$, which are also available in closed-form (see Marlin et al. (2011) for details).

Algorithm 1 Generalized EM Algorithm for SB-LGM

E-Step:

$$\frac{\partial \underline{\mathcal{L}}_J}{\partial \mathbf{m}_n} \leftarrow -\boldsymbol{\Sigma}^{-1}(\mathbf{m}_n - \boldsymbol{\mu}) + \sum_{d=1}^{D} \mathbf{W}_d^T \mathbf{g}_{dn}$$

$$\frac{\partial \underline{\mathcal{L}}_J}{\partial \mathbf{V}_n} \leftarrow \frac{1}{2}\left(\mathbf{V}_n^{-1} - \boldsymbol{\Sigma}^{-1}\right) + \sum_{d=1}^{D} \mathbf{W}_d^T \mathrm{diag}(\mathbf{h}_{dn}) \mathbf{W}_d$$

M-Step:

$$\boldsymbol{\mu} \leftarrow \frac{1}{N} \sum_{n=1}^{N} \mathbf{m}_n$$

$$\boldsymbol{\Sigma} \leftarrow \frac{1}{N} \sum_{n=1}^{N} \mathbf{V}_n + (\mathbf{m}_n - \boldsymbol{\mu})(\mathbf{m}_n - \boldsymbol{\mu})^T$$

$$\frac{\partial \underline{\mathcal{L}}_J}{\partial \mathbf{W}_d} \leftarrow \sum_{n=1}^{N} \mathbf{g}_{dn} \mathbf{m}_n^T + 2\mathbf{W} \mathrm{diag}(\mathbf{h}_{dn}) \mathbf{V}$$

We denote these gradients by $\mathbf{g}_{dn} := \partial B / \partial \tilde{\mathbf{m}}_{dn}$ and $\mathbf{h}_{dn} := \partial B / \partial \tilde{\mathbf{v}}_{dn}$.

We give the gradients or closed form updates as appropriate in Algorithm 1. We use limited memory BFGS to perform the updates that require numerical optimization. The piecewise bound parameters $\boldsymbol{\alpha}$ are computed in advance and are fixed during learning and inference.

4.2 Computational Complexity

The computation of $\mathbf{g}_{dn}$ and $\mathbf{h}_{dn}$ is $O(DKNR)$. To compute the sum over d in the E and M-steps costs $O(NDKL^2)$ and inversion costs $O(NL^3)$. The total computational complexity of one iteration of our algorithm is $O(DKNR + (DKL^2 + L^3)N)$. In the special case of multi-class Gaussian process classification, we have $L = DK$ and $N = 1$ giving us complexity in $O(D^3K^3)$ and a straightforward implementation will not be efficient. However, optimization can be made simpler by reparameterizing the covariance matrix as suggested in Opper and Archambeau (2009).

5 Related Work

There is a great deal of related work on learning standard multinomial-probit and multinomial-logit LGMs. Moustaki and Knott (2000) describe an EM algorithm for learning in exponential family factor analysis (EFA). The integration of the latent variables is achieved by quadrature, limiting the applicability of this approach. Collins et al. (2002) describe an al-

ternative method for learning EFA models based on an alternating optimization of the latent variables and parameters. This approach does not take into account the uncertainty in the latent variables and may not perform well in some cases (Khan et al., 2010). In addition, these estimation methods are not easily adapted to handling missing data, suffer from overfitting and can exhibit sensitivity to regularization.

Fully Bayesian approaches have been explored to overcome the above limitations. Albert and Chib (1993) describe Bayesian methods based on Gibbs sampling, but these do not scale to large-scale applications. Mohamed et al. (2008) describe a fully Bayesian approach for the exponential family factor analysis model based on Hybrid Monte Carlo. This approach can be quite accurate, but the sampler requires careful tuning. Holmes and Held (2006), Scott (2011), and Frühwirth-Schnatter and Frühwirth (2010) describe MCMC samplers using auxiliary variables for inference in multinomial-logit regression models. These methods generally tend to be slower than deterministic approaches, and it is usually difficult to assess their convergence.

The Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009) can also be used, but this approach is limited to six or fewer parameters and is thus not suitable for most problems that we consider. Multi-class expectation propagation (EP) is described by Seeger and Jordan (2004), but has issues that we discuss in section 7. A variational Bayesian multinomial probit regression is described by Girolami and Rogers (2006), and uses the auxiliary variable representation of the probit function (Eq. 8), with a factorial representation for the posterior distribution. This model was shown to be effective compared to Gibbs sampling and Laplace approximations, but the factorial representation limits the effectiveness of the inference procedure as we show in this paper.

Alternative local variational methods are described by Blei and Lafferty (2006); Khan et al. (2010), Braun and McAuliffe (2010), Bouchard (2007), and Ahmed and Xing (2007). These approaches are easy to generalize to different LGMs, and are amenable to developing efficient parameter learning algorithms. The key disadvantage is that the error due to the local approximation may result in a severe bias in the parameter estimates, as we show next.

6 Results

We use $p_{logit}(\mathbf{y}|\boldsymbol{\theta})$ to refer to the exact probability of a data vector $\mathbf{y}$ under the multinomial-logit LGM with parameters $\boldsymbol{\theta}$. Similarly, we use $p_{stick}(\mathbf{y}|\boldsymbol{\theta})$ to refer to the exact probability under the stick-breaking LGM. These exact probabilities remain intractable, but for small D we can compute them to reasonable accuracy using a Monte Carlo approximation to the integral, $p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{z})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{z}$, where the likelihood is either the multinomial-logit or stick-breaking. The MATLAB code to reproduce the results in this section can be found online[1].

6.1 Synthetic Data Experiments

We generate data from a 2D categorical Latent Gaussian graphical model (cLGGM). A cLGGM is essentially a factor model in which $L = DK$ and $\mathbf{W}$ is the identity matrix, while $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unrestricted. We assume that both dimensions have K categories, giving us K^2 unique data cases. We set the true parameters $\boldsymbol{\theta}^*$ to $\boldsymbol{\mu}^* = 0$ and $\boldsymbol{\Sigma}^* = 20\text{cov}(\mathbf{X}) + I_L$, where $\mathbf{X} = [I_{K-1} I_{K-1}]$. This choice of $\boldsymbol{\Sigma}^*$ forces both dimensions to take the same value, resulting in high correlation. We sample 10^6 data cases from the logit model to get an estimate of $p_{logit}(\mathbf{y}|\boldsymbol{\theta}^*)$. We estimate parameters $\hat{\boldsymbol{\theta}}$ of logit and stick using this dataset. For the logit model, we use two versions of variational EM algorithms based on the Bohning bound (Khan et al., 2010) and the Blei bound (Blei and Lafferty, 2006) respectively. For the stick model, we use our proposed variational EM algorithm. We refer to these three methods as 'logit-Bohning', 'logit-Blei', and 'stick-PW' respectively. Note that since the data is generated from a multinomial-logit model, there is a modeling error for stick-PW in addition to the approximation in learning.

We first compare results for $K = 4$ in Fig. 1(a) which shows the true $p_{logit}(\mathbf{y}|\boldsymbol{\theta}^*)$ as well as $p_{logit}(\mathbf{y}|\hat{\boldsymbol{\theta}})$ for logit-blei and logit-Bohning, and $p_{stick}(\mathbf{y}|\hat{\boldsymbol{\theta}})$ for stick-PW. We see that stick-PW obtains a very close probability distribution to the true distribution, while other methods do not. Figure 1(b) shows results for $K = 4, 5, 6, 7, 8$. Here we plot KL-divergence between the true distribution $p_{logit}(\mathbf{y}|\boldsymbol{\theta}^*)$ and the estimated distributions for each method. We see that our method consistently gives very low KL divergence values (the values for other methods are decreasing because the entropy of the true distribution decreases since we have set the multiplying constant in $\boldsymbol{\Sigma}^*$ to 20 for all categories).

6.2 Multi-class Gaussian process classification

In this section, our goal is to compare the marginal likelihood approximation and its suitability for parameter estimation. We consider a multi-class Gaussian process classification (mGPC) model since the number

[1] www.cs.ubc.ca/∼emtiyaz/software/catLGM.html

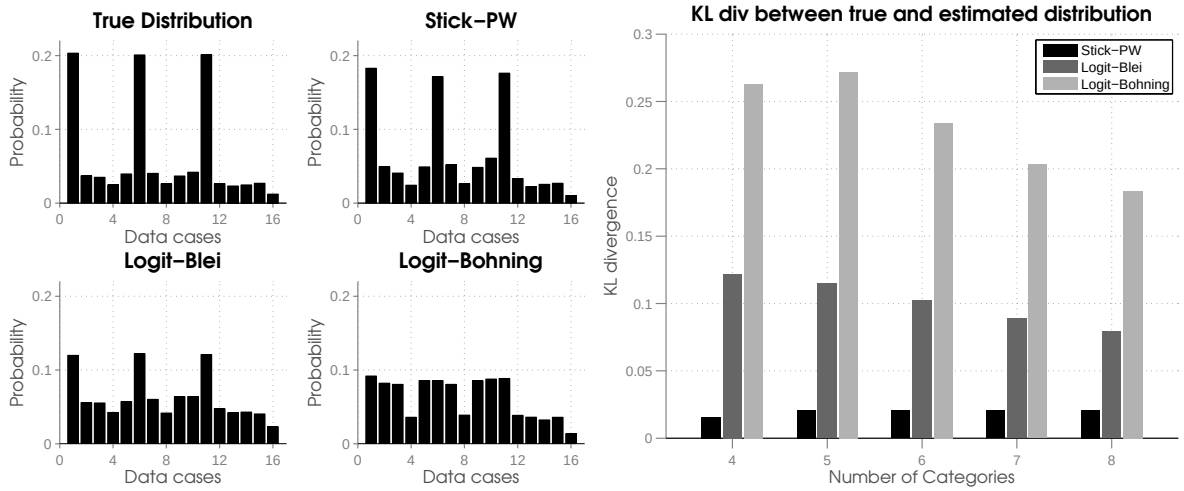


Figure 1: (a) Comparison of the true probability distribution to the estimated distributions on synthetic data with 4 categories. (b) KL divergence between the true and estimated distributions for different categories.

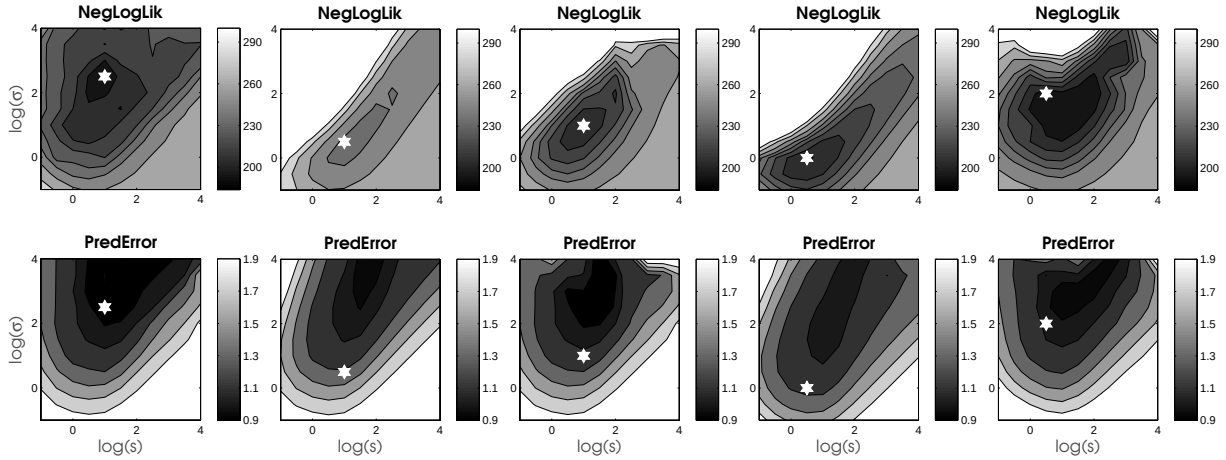
of parameters is small, which makes it easier to investigate the approximation. We use hybrid Monte Carlo (HMC) sampling along with annealed importance sampling (AIS) to get the 'true' value of marginal likelihood. We present results for the multinomial-logit link function and refer to this as logit-HMC. We compare to the multinomial probit model of Girolami and Rogers (2006), which uses variational-Bayesian inference. For this method, we use the MATLAB code provided by the authors[2]. We refer to this as the 'probit-VB' approach. We also compare to a multinomial-logit model learned using a variational EM algorithm based on the Blei bound proposed by Blei and Lafferty (2006) and the Bohning bound proposed in Bohning (1992). We refer to these models as the 'logit-Blei' and 'logit-Bohning' respectively.

We apply the mGPC model to the forensic glass data set (available from the UCI repository) which has $D = 214$ data examples, $K = 6$ categories, and features $\mathbf{x}$ of length 8. We use 80% of the dataset for training and the rest for testing. We set $\boldsymbol{\mu} = 0$ and use a squared-exponential kernel, for which the (i, j)'th entry of $\boldsymbol{\Sigma}$ is defined as: $\boldsymbol{\Sigma}_{ij} = -\sigma^2 \exp[-\frac{1}{2}||\mathbf{x}_i - \mathbf{x}_j||^2/s]$. To compare the marginal likelihood, we fix $\boldsymbol{\theta}$ which consists of σ and s and compute a posterior distribution (or draw samples from it), and an approximation to the marginal likelihood using one of the methods mentioned above. We compute the prediction error defined as $-\log_2 \tilde{p}(y_{test}|\boldsymbol{\theta}, \mathbf{y}_{train}, \mathbf{x}_{train}, \mathbf{x}_{test})$, where $(\mathbf{y}_{train}, \mathbf{x}_{train})$ and $(y_{test}, \mathbf{x}_{test})$ are training and testing data, respectively. Here, $\tilde{p}(y_{test}|\cdot)$ is the marginal predictive distribution approximated using the Monte Carlo method.

Figure 2 shows the contour plots for all the methods over a range of settings for the hyperparameter values of the Gaussian process. The top row shows the negative log marginal likelihood approximation and the bottom row shows the prediction error. The star indicates the hyperparameter value at the minimum of the negative log marginal likelihood. The first column is the 'true' marginal likelihood obtained by sampling for logit-HMC. This plot shows the expected behavior of the true marginal likelihood. As we increase σ^2, we move from Gaussian-like posteriors to a posterior that is highly non-Gaussian. The posterior in the high σ^2 region is effectively independent of σ^2 and thus we see contours of marginal likelihood that remain constant (this has also been noted by Nickisch and Rasmussen (2008)). Importantly for model selection, there is a correspondence between the minimum value of the marginal likelihood (or evidence) and the region of minimum prediction error. Thus optimizing the hyperparameters and performing model selection by minimizing the marginal likelihood gives optimal prediction. In our experience, tuning HMC parameters is a tedious task for this model as these parameters depend on $\boldsymbol{\theta}$. In addition, convergence is difficult to assess. Both HMC and AIS samplers need to be run for many iterations to get reasonable estimates.

Columns 2 and 3 show the log marginal likelihood and prediction for the Bohning bound, and Blei's bound for the logit model. As we increase σ^2, the posterior becomes highly non-Gaussian and the variational bounds strongly underestimate the marginal likelihood

[2]We corrected a bug in this code for marginal likelihood computation. The corrected code can be found online.



(a) Logit-HMC (b) Logit-Bohning (c) Logit-Blei (d) Probit-VB (e) Stick-PW

Figure 2: Comparison of methods using multi-class GP classification on the glass dataset. The top row shows the negative log marginal likelihood approximations and the bottom row shows the prediction errors. Each column is a different method. The first column can be considered as ground-truth.

Table 1: Performance of methods at the best parameter setting (a star in Fig. 2)

Method	s	σ	negLogLik	predError
Logit HMC	1	2.5	198.63	0.92
Logit-Boh	1	0.5	239.28	1.31
Logit-Blei	1	1	208.26	1.13
Probit-VB	0.5	0	203.59	1.23
Stick-PW	0.5	2	194.16	1.07

.

in these regions (upper left corner of plots). The variational approximation also reduces the correspondence between the marginal likelihood and the test prediction, thus the minimum of the marginal likelihood is not useful in finding regions of low prediction error (high information score), resulting in suboptimal performance. The Blei-bound, being a tighter bound than the Bohning bound, provides improved marginal likelihood estimates as expected, and a better correspondence between the prediction error and the marginal likelihood. The 4th column is the behavior of the multinomial probit model and confirms the behavioral similarity of the logit and probit likelihoods.

The behavior of the stick likelihood is shown in the 5th column. The piecewise bound is highly effective for this model and the model provides good estimates even in the highly non-Gaussian posterior regions. An important appeal of this model is that the correspondence between the marginal likelihood and the prediction is better maintained than the logit or probit models, and thus parameters obtained by optimizing the marginal likelihood will result in good predictive performance.

Performance of all methods at the best parameter setting is summarized in Table 1 showing the best parameter values, an approximation to the negative marginal log-likelihood, and prediction error.

6.3 Categorical Latent Gaussian Graphical Model (cLGGM)

We compare Blei's bound to our piecewise bound for a latent Gaussian graphical model, using the tic-tac-toe data set, which consists of 958 data examples with 10 dimensions each. All dimensions have 3 categories except the last one which is binary (thus the sum of categories used in the cLGGM is 29). We use 80% for training and 20% for testing. The ASES data set consists of survey data from respondents in different countries (available online[3]). We select one country (UK) and only the categorical responses, resulting in 17 response fields from 913 people; 9 response fields have 4 categories and the remainder have 3 categories. We compute the imputation accuracy on the test data. Basically, we run inference on the test data using the estimated parameters and compute the predictive probability for each missing value. The prediction error is computed similar to Khan et al. (2010).

Figure 3(a) shows the error versus time for one split, for the tic-tac-toe data. The plot shows that the stick-PW is a better method to use, since it gives much lower error when the two methods are run for the same amount of time. Figure 3(b) compares the error of the Blei-bound and the piecewise bound for all 20 data splits used. For all splits, the points lie below the

[3] www.cs.ubc.ca/∼emtiyaz/datasets.html

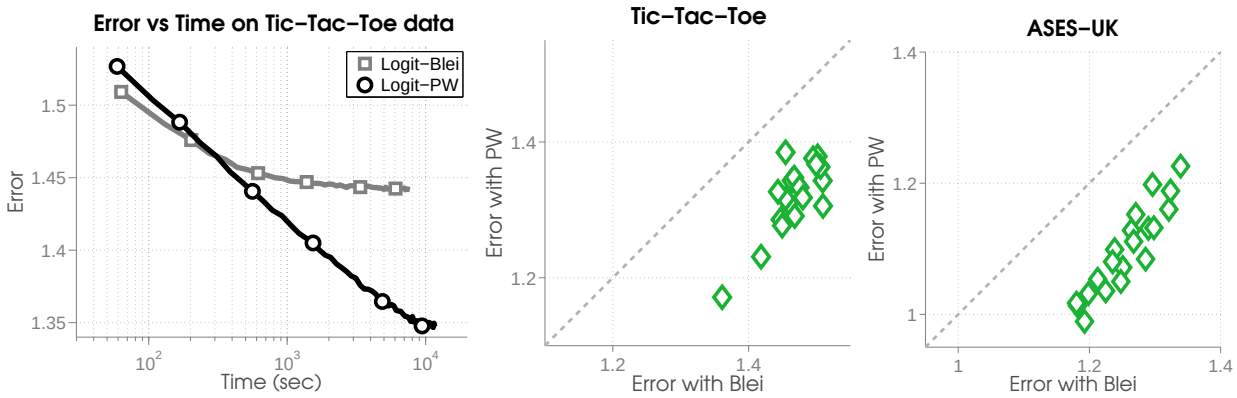


Figure 3: Results on cLGGM model: (a) Imputation error vs time for tic-tac-toe data (b) Imputation error for different splits for tic-tac-toe data (c) Imputation error for different splits for ASES-UK data.

diagonal line, indicating that the piecewise bound has better performance. We show a similar plot for the ASES data set in figure 3(c), which more markedly shows the improvement in prediction when using the piecewise bound over Blei's bound.

7 Discussion and Conclusion

We have presented a new stick-breaking latent Gaussian model for the analysis of categorical data. We also derived an accurate and efficient variational EM algorithm using piecewise linear and quadratic bounds. Due to the bounded error of the piecewise bounds, we are able to reduce the error in the lower bound to the marginal likelihood, up to the error introduced by Jensen's inequality. This leads to accurate estimates of the marginal likelihood and parameters, resulting in improved prediction accuracy. In contrast, variational learning in existing logit/probit based LGMs gives poor parameter estimates due to inaccurate bounds for the log-sum-exp function. Our extensive comparison with existing logit/probit based LGMs demonstrated that the proposed stick-breaking model effectively captures correlation in discrete data and is well suited to the analysis of categorical data.

A likelihood similar to our stick-breaking model has been proposed for probabilistic language modeling in Mnih and Hinton (2009) where the probability of a word is expressed as a product of sigmoids. A similar idea using a product of sigmoids has been applied by Bouchard (2007) to build efficient variational bounds for the log-sum-exp function.

A popular alternative approach to ours is Expectation Propagation (EP) (Minka, 2001), which has been shown to give good performance for binary Gaussian process classification (Nickisch and Rasmussen, 2008). An extension of EP to multi-class Gaussian process classification for the multinomial-logit link is derived by Seeger and Jordan (2004), but they state that their approach is "fundamentally limited by the requirement of an efficient numerical integration in K dimensions" (Seeger and Jordan, 2004, §4.3.1). For the multinomial-probit link, this is not a limitation since the numerical integration can be done efficiently as described in Seeger et al. (2006). The EP updates, however, are usually complicated for these methods. A more important problem with EP is parameter learning in models, such as categorical factor analysis, for which we are not aware of any work. The difficulty in parameter learning with EP is discussed by Seeger and Jordan (2004, §5) in the context of multi-class Gaussian process classification. They suggest the use of a lower bound based on KL divergence, since an EP approximation is not easy to obtain in the multi-class case. This leads to a non-standard and usually non-descending optimization since the inference and learning steps do not optimize the same lower bound. These lower bounds are usually not convex, which further adds to the difficulty. Such a hybrid VB-EP is used by Rattray et al. (2009), who also discuss the difficulty in computing EP approximations for the parameter learning setting.

Acknowledgments

We thank Guillaume Bouchard (XRCE) for encouraging us to pursue the idea of the stick-breaking likelihood. SM is supported by the Canadian Institute for Advanced Research (CIFAR). We thank the reviewers for their valuable suggestions. This work was supported in part by the Institute for Computing, Information and Cognitive Systems (ICICS) at UBC.

References

A. Ahmed and E. Xing. On tight approximate inference of the logistic-Normal topic admixture model. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2007.

J. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the Americal Statistical Association*, 88(422):669–679, 1993.

C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

D. Blei and J. Lafferty. Correlated topic models. In *Advances in Neural Information Proceeding Systems*, 2006.

D. Bohning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44:197–200, 1992.

G. Bouchard. Efficient bounds for the softmax and applications to approximate inference in hybrid models. In *NIPS workshop on approximate inference in hybrid models*, 2007.

M. Braun and J. McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

M. Collins, S. Dasgupta, and R.E. Schapire. A generalization of principal component analysis to the exponential family. In *Advances in neural information processing systems*, pages 617–624, 2002.

S. Frühwirth-Schnatter and R. Frühwirth. Data augmentation and MCMC for binary and multinomial logit models. *Statistical Modelling and Regression Structures*, pages 111–132, 2010.

M. Girolami and S. Rogers. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Comptuation*, 18(8):1790 – 1817, 2006.

C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 2006.

M. Khan, B. Marlin, G. Bouchard, and K. Murphy. Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Proceeding Systems*, 2010.

M. Knott and D.J. Bartholomew. *Latent variable models and factor analysis*. Number 7. 1999.

B. Marlin, M. Khan, and K. Murphy. Piecewise bounds for estimating Bernoulli-logistic latent Gaussian models. In *International Conference on Machine Learning*, 2011.

T. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, 2001.

A. Mnih and G.E. Hinton. A scalable hierarchical distributed language model. *Advances in Neural Information Processing Systems*, 21:1081–1088, 2009.

S. Mohamed, K. Heller, and Z. Ghahramani. Bayesian Exponential Family PCA. In *Advances in Neural Information Proceeding Systems*, 2008.

I. Moustaki and M. Knott. Generalized latent trait models. *Psychometrika*, 65(3):391–411, 2000.

H. Nickisch and C.E. Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9(10), 2008.

M. Opper and C. Archambeau. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

M. Rattray, O. Stegle, K. Sharp, and J. Winn. Inference algorithms and learning theory for Bayesian sparse factor analysis. *Journal of Physics: Conference Series*, 197(1):012002, 2009.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of Royal Statistical Society Series B*, 71:319–392, 2009.

S. L. Scott. Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models. *Statistical Papers*, 52(1):87–109, 2011.

M. Seeger and M. I. Jordan. Sparse Gaussian process classification with multiple classes. Technical Report Department of Statistics TR 661, University of California, Berkeley, 2004.

M. Seeger, N. Lawrence, and R. Herbrich. Efficient nonparametric Bayesian modelling with sparse Gaussian process approximations. Technical report, Max Planck Institute, 2006.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

M. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of Royal Statistical Society Series B*, 21(3):611–622, 1999.

P. M. Vijverberg. Betit: A family that nests probit and logit. Technical Report DP N 222, IZA, 2000.

M. Wedel and W. Kamakura. Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika*, 66(4):515–530, December 2001.