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

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 logistic stick-breaking likelihood function for categorical LGMs that can exploit recently proposed linear and quadratic bounds on the logistic-log-partition function, leading to an effective variational inference and learning framework. We thoroughly compare the proposed approach to existing algorithms for both multinomial-probit and multinomial-logit likelihoods on several problems including inference in multinomial Gaussian process classification and learning in categorical latent Gaussian graphical models.

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 in a wide variety of areas from the analysis of discrete choice data and survey responses in social science and econometrics to predicting patient health outcomes in medical data and user preferences in recommender systems.

In this paper, we focus on the class of Latent Gaussian Models (LGMs). LGMs allow for a principled handling of missing data, can be used for dimensionality reduction and are highly effective models for prediction and visualisation. LGMs are also highly flexible models. They allow for the parameterization of arbitrary likelihood functions in terms of a linear projection of a vector of Gaussian-distributed latent variables.

In the case of LGMs for categorical data, the two most widely used likelihoods are the multinomial-probit likelihood and the multinomial-logit or softmax likelihood. 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 carried out analytically in Gaussian-likelihood LGMs like factor analysis because the model is jointly Gaussian in the latent factors and the observed variables. LGMs with logit and probit-based likelihoods lack this property, resulting in intractable integrals in 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 much 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. 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 including inference in multinomial Gaussian process classification and learning in categorical latent Gaussian graphical models.

2 Categorical Latent Gaussian Models

In this section, we describe a generic LGM for categorical data. We begin by introducing the required notation. We let $N$ be the number of data instances. We denote the $n$'th visible data vector by $y_n$ and the corresponding latent vector by $z_n$. In general, $y_n$ and
$z_n$ will have dimensions $D$ and $L$, respectively. Each element of $y_n$, 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 $K_d = K$ for all $d$. We use a dummy encoding for $y_{dn}$, that is, we encode it as a binary vector $y_{dn} \in \{0, 1\}^{K+1}$ where we set $y_{dn} = 1$ if $y_{dn} = C_k$.

In LGMs, the latent variables $z_n$ follow a Gaussian distribution with mean $\mu$ and covariance matrix $\Sigma$ as shown in Eq. 1. The probability of each categorical variable $y_{dn}$ is parameterized in terms of the linear projection $\eta_{dn} = W_d z_n + w_{0d}$ of the latent variables $z_n$ as seen in Eqs. 2 and 3. Here, $W_d \in \mathbb{R}^{(K+1) \times L}$ is the factor loading matrix and $w_{0d} \in \mathbb{R}^{K+1}$ is the offset vector, implying that $\eta_{dn} \in \mathbb{R}^{K+1}$. We consider the problem of choosing a parameterization for the categorical distribution $p(y_n|\eta_n)$ in the next section.

We denote the set of parameters by $\theta = \{\mu, \Sigma, W, w_0\}$ where $W$ and $w_0$ are the sets containing $W_d$ and $w_{0d}$ for all dimensions. To make the model identifiable, we set the last row of $W_d$ and last element of $w_{0d}$ to zero. Also, the prior mean $\mu$ and the offset $w_0$ are interchangeable in all the models we consider so we use the mean only.

$$p(z_n|\theta) = \mathcal{N}(z_n|\mu, \Sigma)$$  \hspace{1cm} (1)  

$$\eta_{dn} = W_d z_n + w_{0d}$$  \hspace{1cm} (2)  

$$p(y_n|\eta_n) = \prod_{d=1}^{D} p(y_{dn}|\eta_{dn})$$  \hspace{1cm} (3)  

Different models for categorical data can be obtained by restricting the generic model in different ways. We obtain the categorical factor analysis (cFA) model by assuming that $L \leq D$ and $\Sigma$ is the identity matrix, while $W$ and $\mu$ are unrestricted. Conversely, we obtain the categorical latent Gaussian graphical model (cLGGM) by assuming that $D = L$ and $W$ is the identity matrix, while $\mu$ and $\Sigma$ are unrestricted. We obtain a multi-class Gaussian process classification (mGPC) model by restricting $N = 1$ and $W$ to be the identity matrix, and specifying $\mu$ and $\Sigma$ 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

The generic categorical LGM defined in the previous section specializes into a number of different model types. However, it is only defined up to the probability $p(y|\eta)$. In this section, we review the multinomial-probit and multinomial-logit parameterizations for $p(y|\eta)$ and introduce our new stick-breaking parameterization.

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 $j \neq k$.

$$p(y = k|\eta) = \int_{R_k} \prod_{j=0}^{K} p(u_j|\eta_j) du$$  \hspace{1cm} (8)  

$$p(y = k|\eta) = \frac{e^{\eta_k}}{\sum_{j=0}^{K} e^{\eta_j}} = e^{\eta_k} \text{lse}(\eta)$$  \hspace{1cm} (9)  

The form of the multinomial-logit function is given in Eq. 9. It is defined using the log-sum-exp (LSE) function $\text{lse}(\eta) = \log \sum_j \exp(\eta_j)$ 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 that we refer to as the stick-breaking parameterization. In this 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 $K+1$th category then receives the remaining stick length, as seen below.

$$p(y = C_0|\eta) = \sigma(\eta_0)$$  

$$p(y = C_k|\eta) = \prod_{j=k+1}^{K} (1-\sigma(\eta_j))\sigma(\eta_k), 0 < k < K$$  

$$p(y = C_K|\eta) = \prod_{j=1}^{K-1} (1-\sigma(\eta_j))$$  \hspace{1cm} (10)  

The probabilities (stick lengths) are all positive and sum to one. They thus define a valid probability distribution. The probability can be written more compactly as shown in Eq. 11.

$$p(y = C_k|\eta) = \exp(\eta_k - \sum_{j \leq k} \log(1 + e^{\eta_j}))$$  \hspace{1cm} (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); Hahn et al. (2009). It is understood that these parameterizations give similar performance and qualitative conclusions. By contrast, the stick-breaking parameterization is not well suited for the
parametric regression setting because it places restrictions on the form of the boundaries between categories in feature space. In the context of LGMs, however, there is no fixed feature space and the stick-breaking parameterization becomes a viable alternative to the standard multinomial-probit and logit models.

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 lse(η) function and, at present, it is not known how to obtain tight bounds on this function with more than two categories (Bohning, 1992; 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^y), This function is called the logistic-log-partition function. In stark contrast to the multinomial-logit case, extremely high quality 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 needed for learning 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 \( \mathcal{L}(\theta) \) given in Eq. 4 and introduce a variational posterior distribution \( q(z|\gamma_n) \) for each data case. We use a Gaussian posterior with mean \( m_n \) and covariance \( V_n \). The full set of variational parameters is thus \( \gamma_n = \{m_n, V_n\} \), and we use \( \gamma \) to denote the set of all \( \gamma_n \).

As log is a concave function, we obtain a lower bound \( \mathcal{L}_j(\theta, \gamma) \) using Jensen’s inequality as shown in Eq. 5. The first term is simply the Kullback–Leibler (KL) divergence from the variational Gaussian posterior \( q(z|m_n, V_n) \) to the Gaussian prior distribution \( p(z|\mu, \Sigma) \) which has a closed-form expression as shown in Eq. 7. In the second term, we substitute the likelihood definition from Eq. 3 and apply a change of variable from \( z \) to \( \eta \) to get Eq. 6. The new expectation is with respect to \( q(\eta)\gamma_{dn} \), where \( \gamma_{dn} = \{m_{dn}, V_{dn}\} \), \( m_{dn} = W_\eta m_n + w_0d \), and \( V_{dn} = W_\eta V_n W_\eta^T \).

The lower bound \( \mathcal{L}_j(\theta, \gamma) \) is still intractable as the expectation of \( \log p(y|\eta) \) is not available in closed form. We now describe the use of piecewise linear/quadratic bounds to derive a tractable lower bound. For simplicity, we suppress the dependence on \( d \) and \( n \) and consider the log-likelihood of a scalar observation \( y \) given a predictor \( \eta \sim q(\eta|\gamma) = \mathcal{N}(\eta|\tilde{m}, \tilde{V}) \) with \( \gamma = \{\tilde{m}, \tilde{V}\} \). 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.

\[
\begin{align*}
\mathcal{L}(\theta) &= \sum_{n=1}^{N} \log \int_z p(z|\theta)p(y_n|\eta)dz = \sum_{n=1}^{N} \log \int_z q(z|\gamma_n)\frac{p(z|\theta)p(y_n|\eta)}{q(z|\gamma_n)}dz \\
\mathcal{L}(\theta) &\geq \mathcal{L}_j(\theta, \gamma) := \sum_{n=1}^{N} -\int_z q(z|\gamma_n) \log \frac{q(z|\gamma_n)}{p(z|\theta)} dz + \int_z q(z|\gamma_n) \log p(y_n|\eta)dz \\
&= \sum_{n=1}^{N} -D_{KL}[q(z|\gamma_n)||p(z|\theta)] + \sum_{d=1}^{D} \mathbb{E}_{q(\eta|\gamma_{dn})} \log p(y_{dn}|\eta)d\eta \\
D_{KL}(q_n(z|\gamma_n)||p(z|\theta)) &= \frac{1}{2} \left[ -\log |V_n\Sigma^{-1}| + \text{tr}(V_n\Sigma^{-1}) + (m_n - \mu)^T\Sigma^{-1}(m_n - \mu) - L \right]
\end{align*}
\]
the expectation with respect to a truncated Gaussian distribution as shown in Eq. 14. We denote this lower bound by \( B(y, \gamma, \alpha) \). Note that the bound only depends on the diagonal elements of \( \mathbf{V} \). We denote these by \( \mathbf{v} \).

4.1 A Generalized EM Algorithm

We use the piecewise bound to obtain a final, tractable lower bound on \( L_j(\theta, \gamma) \), which we denote by \( \bar{L}_j(\theta, \gamma) \). To learn the parameters, we need to optimize the lower bound with respect to the variational posterior parameters \( \gamma \) and model parameters \( \theta \). 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 \( B(y, \gamma, \alpha) \) with respect to \( \mathbf{m} \) and \( \mathbf{v} \) which are also available in closed-form (see Marlin et al. (2011) for details). We denote these gradients by \( \mathbf{g}_m(y, \gamma, \alpha) \) and \( \mathbf{g}_v(y, \gamma, \alpha) \), both being real \( K + 1 \) length vectors. Also define \( \mathbf{G}_v \) to be a diagonal matrix with \( \mathbf{g}_v(y, \gamma, \alpha) \) as diagonal.

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 \( \alpha \) are computed in advance and fixed during learning and inference.

4.2 Computational Complexity

The computation of \( \mathbf{g}_m(y_{dn}, \gamma_{dn}, \alpha) \) and \( \mathbf{G}_v(y_{dn}, \gamma_{dn}, \alpha) \) is \( O(D K N R) \). To compute the sum over \( d \) in the E and M-steps costs \( O(N D K L^2) \) and inversion costs \( O(N L^3) \) (note that \( \mathbf{G}_v \) is a diagonal matrix). The total computational complexity of our algorithm is \( D K N R + (D K L^2 + L^3) N \). In the special case of multi-class Gaussian process classification, we have \( L = D K \) and \( N = 1 \) giving us complexity in \( O(D^3 K^3) \) and a straightforward implementation will not be efficient. However, the computation can be reduced to \( O(D^2 K^2) \) by reparameterizing the covariance matrix as suggested in Opper and Archambeau (2009) and then using rank-one updates. Complexity can be reduced further by assuming other sparse approximation.

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 and 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 alternative 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. In addition, these estimation methods are not easily adapted to handling missing data, suffer from overfitting and can exhibit sensitivity to regularisation.

Fully Bayesian approaches have been explored to overcome the above limitations. Albert and Chib (1993); Agresti (2002) describe Bayesian methods based on Gibbs sampling, but these are relevant to small scale-applications. Mohamed et al. (2008) described fully Bayesian approach based on Hybrid Monte Carlo for the exponential family factor analysis model. This approach can be highly accurate, but the sampler requires careful tuning. Holmes and Adams (2003); Scott and Carvalho (2008) describe a collapsed Gibbs sampler using auxiliary variables for inference in multinomial regression models. In all cases, it is difficult to assess the convergence of these MCMC methods and they are often slow.

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. Multi-class
expectation propagation (EP) is described by Seeger and Jordan (2004), but has problems that we discuss in section 7. Variational Bayesian multinomial probit regression is described by Girolami and Rogers (2006), and uses the auxiliary variable representation of the probit function, 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.

Alternative local variation methods, which improve on the factorial representation used by Girolami and Rogers (2006), are described by Blei et al. (2003); Khan et al. (2010); Braun and McAuliffe (2010); Bouchard (2007). These approaches are easy to generalise 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

Throughout this section, we use \( p_{\text{logit}}(y|\theta) \) to refer to the exact probability a data case \( y \) under the multinomial-logit LGM with parameters \( \theta \). Similarly, we use \( p_{\text{stick}}(y|\theta) \) 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 Monte-Carlo. Also, we will refer to multinomial-logit LGM as ‘logit’ model and stick-breaking LGM as ‘stick’ model in this section.

6.1 Synthetic Data Experiments

In this section, we generate data from a 2D categorical LGGM model. We assume that both dimensions have \( K \) categories, giving us \( K^2 \) unique data cases. We set the true parameters \( \theta^* \) to \( \mu^* = 0 \) and \( \Sigma^* = 20 \text{cov}(\mathbf{X}) + \mathbf{I}_K \), where \( \mathbf{X} = [I_{K-1} I_{K-1}] \). This choice of \( \Sigma^* \) forces both dimensions to take same value, resulting in high correlation. We sample \( 10^6 \) data cases from the logit model to get an estimate of \( p_{\text{logit}}(y|\theta^*) \). We estimate parameters \( \theta \) of logit and stick using a dataset consisting of all \( K^2 \) data cases \( y \) weighted by \( p_{\text{logit}}(y|\theta^*) \). For the logit model, we use variational EM algorithm based on the Bohning and the Blei bound. 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 for the stick model there is a modeling error in addition to the approximation in learning.

We first compare results for \( K = 4 \) in Fig. 1(a) which shows the true \( p_{\text{logit}}(y|\theta^*) \) as well as \( p_{\text{logit}}(y|\theta) \) for Logit-Blei and Logit-Bohning and \( p_{\text{stick}}(y|\theta) \) 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 \( K = 4, 5, 7, 8 \). Here we plot KL-divergence between true distribution \( p_{\text{logit}}(y|\theta^*) \) 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 \( \Sigma^* \) 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 Multi-class Gaussian process classification (mGPC) model as in this case the number of parameters is small, which makes it easy to investigate the approximation. We use hybrid Monte Carlo (HMC) sampling along with annealed importance sampling (AIS) to get “true” value of marginal likelihood. We obtain very similar estimates for logit/probit/stick model and hence only present the logit model. We refer to this approach as Logit-HMC. We consider the multinomial probit model of Girolami and Rogers (2006), which uses variational-Bayesian inference. We use the MATLAB code provided by the user. We call this the Probit-VB approach. A Multinomial-Logit model learned using variational E step similar to our approach. We use the bound proposed by Blei and Lafferty (2006) and the Bohning bound proposed by Bohning (1992); Khan et al. (2010). We refer to these models as Logit-Blei and Logit-Bohning respectively.

We apply mGPC model to the forensic glass data set (available in the UCI repository) which has \( D = 214 \) data examples with \( K = 6 \) categories with features \( \mathbf{x} \) of length 8. We use 80% of the dataset for training and rest for testing. We set \( \mu = 0 \) and use a squared-exponential kernel, under which \( (i, j) \)th entry of \( \Sigma \) is defined as \( \Sigma_{ij} = -\sigma^2 \exp(-\frac{1}{2}||\mathbf{x}_i - \mathbf{x}_j||^2/s) \). To compare the marginal likelihood, we fix \( \theta \) which consists of \( \sigma \) and \( s \) and compute a posterior distribution (or draw samples from it) and an approximation to the marginal likelihood using one of methods listed above. We compute prediction error on test data defined as follows: \( -\sum_{d=1}^{D_t} \log_2 p(y_d|\theta, \mathbf{y}_{1:d}, \mathbf{x}_d)/D \) where \( y_d \) is a test data point, \( p(\cdot|\cdot) \) is the marginal predictive distribution approximated with the corresponding method, and \( D_t \) is the size of test data.
Figure 1: (a) Comparison of the true probability distribution to the estimated distributions on a synthetic data with 4 categories. (b) KL divergence between the true distribution and estimated distributions for different categories.

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 of the marginal likelihood approximation and the bottom row the prediction error. The star indicates the hyperparameter value at the minimum negative marginal likelihood value. The first column shows the ‘true’ marginal likelihood obtained by sampling for Logit-HMC. This plot shows the expected behavior of the true marginal likelihood. As we increase $\sigma^2$, we move from Gaussian-like posteriors to a posterior that is highly non-Gaussian. The posterior in the high $\sigma^2$ is effectively independent of $\sigma^2$ and thus we see contours of marginal likelihood that remain constant (this has also also been noted in 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.

Columns 2 and 3 show the marginal likelihood and prediction for the Bohning bound and Blei’s bound for the logit model. As we increase $\sigma^2$, the posterior becomes highly non-Gaussian and the variational bounds strongly underestimate the marginal likelihood 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 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 information score and the evidence. The 4th column is the behavior of the multinomial probit model and confirms the behavioral similarity of the logit and probit likelihoods. The MCMC plot for the probit model shows the expected behavior and is not shown here.

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 and looks similar to the MCMC plot (not shown here). 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 evidence maximization will results in good predictive performance.

6.3 Categorical Latent Gaussian Graphical Model

We compare Blei’s bound to our piecewise bound for a latent Gaussian graphical model (LGGM). An LGGM is essentially a factor model in which the latent dimension is equal to the sum of the categories for an observation vector.

The tic-tac-toe data set has 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 LGGM is 29). We use 80% for training and 20% for testing. The ASES data set consists of survey data from respondents in different countries. 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 remainder has 3 categories. We compute the imputation accuracy on test data. Basically, we run inference on the test data using the estimated parameters and compute predictive probability of each missing value. The prediction error is computed similar to Khan et al. (2010).

Figure 3(a) shows the error vs time for one split, for the tic-tac-toe data. The plot shows that the piecewise bound is a better bound 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 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

In this paper, we have presented a new stick-breaking latent Gaussian model for the analysis of categorical data. We 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 the Jensen 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 to the log-sum-exp function. Our extensive comparison with existing logit/probit based LGMs demonstrated that the proposed stick-breaking model effectively captures the correlation in the discrete data and is well suited to the analysis of categorical data.

A popular alternative approach is Expectation Propagation (EP) (Minka, 2001), and shows good performance for binary Gaussian process classification (Nickisch and Rasmussen, 2008). An extension to multi-class EP is derived by Seeger and Jordan (2004), but as noted by the authors themselves their approach is “fundamentally limited by the requirement of an efficient numerical integration in $K$ dimensions” (Seeger and Jordan, 2004, §4.3.1). Furthermore, we are not aware of any work on parameter learning with EP, for example, for categorical factor analysis. The difficulty in parameter learning with EP is also discussed by Seeger and Jordan (2004, §§5) in the context of multi-class Gaussian process classification. They suggest the use of a lower bound similar to ours, 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 as 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 also used by Rattray et al. (2009) who also discuss the difficulty in computing EP approximations for the parameter learning setting.

In a separate study, we compared our approach to EP on binary data and found that our approach gave comparable performance, in a comparable computation time. Note that for binary observations, the stick-breaking likelihood is identical to the multinomial-logit likelihood, and our variational approach is then the same as that discussed in Marlin et al. (2011). We do not discuss this aspect in detail as it is not the focus of our paper. We found that if the posterior distribution is close to Gaussian, the posterior approximation obtained by our variational approach is almost identical to that obtained by EP. However, if the posterior distribution is skewed, the “zero-enforcing” property of our lower bound shifts the mean away from zero and shrinks the variance. This posterior approximation is thus different from EP, but is still as good for practical purposes. We obtain very similar estimates of the lower bound and the prediction score, as also has been shown by Nickisch and Rasmussen (2008). In addition, we obtain very similar estimates for the EP approximation to the marginal likelihood given in Minka (2001). Similarly, the lower bound described by Seeger and Jordan (2004) and the prediction score of Nickisch and Rasmussen (2008) are almost identical. Thus, we are able to demonstrate that our piecewise variational approach gives highly comparable performance to that of EP for binary classification. The advantage of our approach, however, is that it can be easily extended to multi-class case and also to parameter learning as we show in this paper.

References


