A Nonparametric Bayesian Poisson Gamma Model for Count Data

Sunil Kumar Gupta, Dinh Phung and Svetha Venkatesh
Centre for Pattern Recognition and Data Analytics, Deakin University, Australia
{*sunil.gupta,dinh.phung,svetha.venkatesh*}*@deakin.edu.au*

Abstract

We propose a nonparametric Bayesian, linear Poisson gamma model for count data and use it for dictionary learning. A key property of this model is that it captures the parts-based representation similar to nonnegative matrix factorization. We present an auxiliary variable Gibbs sampler, which turns the intractable inference into a tractable one. Combining this inference procedure with the slice sampler of Indian buffet process, we show that our model can learn the number of factors automatically. *Using synthetic and real-world datasets, we show that the proposed model outperforms other state-of-the-art nonparametric factor models.*

1 Introduction

Factor analysis has been widely used in different applications - dimensionality reduction, dictionary learning, collaborative filtering and so on. One popular instance of factor analysis is nonnegative matrix factorization (NMF) [4], which is developed for nonnegative data. The ability of NMF to capture parts-based representations has enabled modeling of such diverse nonnegative data - images, text, video. However, NMF needs *a priori* information of the number of parts (or factors), mostly unavailable information. Although model selection can provide an estimate of the number of such parts, it is often inefficient and sensitive to the data selection process. One solution to this problem is offered by Bayesian nonparametrics.

Most of the previous nonparametric matrix factorizations have focused on linear Gaussian models using a combination of Indian buffet process (IBP) and Gaussian distributions. Fewer attempts have been made towards nonnegative factor modeling. Santhanam et al [8] propose a nonnegative factor analysis using Poisson distribution, however, it is a *non-Bayesian* approach and subject to *overfitting*. In a recent work, Zhou et al [12] propose a Poisson factor analysis, where a Dirichlet distribution is used to model the factors. But the simplex support of Dirichlet prior restricts modeling of real-world data by imposing a fixed correlation. Paisley and Blei [6] propose a Poisson factor analysis, which uses gamma distributions for both the factors and the loadings and a Poisson model for the data. However, their model does not have a provision to separate noise, crucial for learning subspace dimensionality. Moreover, they use a variational scheme for inference, which instead of sampling from the true posterior, samples them from an approximate posterior. Thus the problem of nonnegative factor analysis for count data using a nonparametric Bayesian framework remains *open*.

Addressing this gap, we propose a nonparametric Bayesian, linear Poisson gamma model - a specialized model for count data. It provides nonnegative factors like NMF and automatically learns the number of factors from the data. For the proposed model, inference is intractable due to the non-conjugacy between the data likelihood and the parameters prior. To circumvent the problem, we present an auxiliary variable sampler that makes the problem tractable without approximations. Using synthetic and real-world datasets, we show that the proposed model outperforms other state-of-the-art nonparametric factor analysis models.

2 Background

2.1 Indian Buffet Process

Indian buffet process (IBP) is a Bayesian nonparametric prior [1] used to model infinite dimensional binary matrices. Let us assume a binary matrix $\mathbf{Z}_{K \times N}$ where K can be infinitely large. In modeling applications, N usually denotes the number of data points and K denotes the number of factors or dictionary elements which may be present/absent to represent a data point. IBP has been used in a range of applications e.g. nonparametric independent component analysis [3], collaborative filtering [5] and structure modeling [11].

The generative model behind IBP is a beta-Bernoulli process [1]. An extension was developed by [10], which can model integer valued matrices replacing the beta-Bernoulli process with a gamma-Poisson process.

2.2 Factor Modeling using IBP

The IBP had been widely used in nonparametric matrix factorizations and factor analysis applications where the main focus had been towards modeling the dyadic data using a linear Gaussian model. Consider a typical matrix factorization problem

$$\mathbf{X}_{D\times N} = \mathbf{W}_{D\times K}\mathbf{H}_{K\times N} + \mathbf{E} \tag{1}$$

where $\mathbf{X}_{D\times N}$ is a matrix containing N data points lying in D-dim Euclidean space. The matrix $\mathbf{W}_{D\times K}$ contains the factors or basis vectors of the transformed subspace (i.e, the subspace spanned by the columns of the matrix $\mathbf{W}$ and denoted as $\mathcal{W}$). The matrix $\mathbf{H}_{K\times N}$ contains co-ordinates of the data in $\mathcal{W}$ (usually $K < N$) . The matrix $\mathbf{E}$ denotes the factorization error.

Usually the dimensionality (K) of $\mathcal{W}$ is not known *a priori* and model-selection needs to be performed. The goal of using IBP is to automatically infer K using the data. The factorization of (1) can be re-written as

$$\mathbf{X}_{D\times N} = \mathbf{W}_{D\times K}\left(\mathbf{Z}_{K\times N} \odot \mathbf{F}_{K\times N}\right) + \mathbf{E} \tag{2}$$

where $\mathbf{H} \triangleq \mathbf{Z} \odot \mathbf{F}$ and $\odot$ denotes element-wise product of two matrices. The matrix $\mathbf{F}$ contains the co-ordinates while $\mathbf{Z}$ is drawn from IBP and $\mathbf{Z}^{kn}$ indicates the presence or absence of k-th basis vector in $\mathbf{W}$.

3 Model Description and Inference

We consider the factorization problem of (2) and propose a linear Poisson gamma model (LPGM) for the factorization ensuring that $\mathbf{W}$ and $\mathbf{F}$ are nonnegative. Given these parameters, both the data and the modeling error follow Poisson distribution (for details, refer Figure 1). Formally,

$$\text{LPGM}: \begin{cases} \mathbf{Z} \sim \text{StickIBP}\left(\alpha\right) \\ \mathbf{W}^{ik} \sim \text{gamma}\left(a, b\right) \\ \mathbf{F}^{kj} \sim \text{gamma}\left(c, d\right) \\ \mathbf{X}^{:,j} \mid \mathbf{W}, \mathbf{Z}^{:,j}, \mathbf{F}^{:,j} \\ \sim \text{Poisson}\left(\mathbf{W}\left(\mathbf{Z}^{:,j} \odot \mathbf{F}^{:,j}\right) + \lambda\right) \end{cases} \tag{3}$$

where $\text{StickIBP}\left(\alpha\right)$ denotes the stick-breaking construction proposed in Teh et al [9]. The symbols a, b, c and d denote the shape and scale parameters of the respective gamma distributions and λ is a parameter for the modeling error $\mathbf{E}$ such that $\mathbf{E}^{ij} \sim \text{Poisson}\left(\lambda\right)$.

For inference of the parameters $\mathbf{W}$ and $\mathbf{F}$, we use Gibbs sampling. The full condition posteriors for these parameters are intractable. Introduction of auxiliary variables, however, makes the inference tractable.

$$p\left(\mathbf{W}^{i,:} \mid \mathbf{Z}, \mathbf{F}, \mathbf{X}, \lambda, \mathbf{s_i}\right) \propto \Pi_k \left(\mathbf{W}^{ik}\right)^{a+\sum_j s_i^{jk}-1} \times \exp\left\{-\left(b + \sum_{l=1}^{N} \mathbf{H}^{kl}\right)\mathbf{W}^{ik}\right\} \tag{4}$$

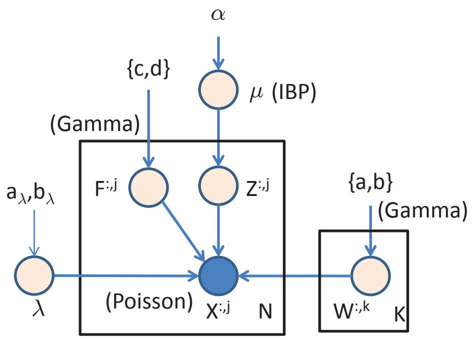


Figure 1: Directed graphical representation of LPGM.

where defining $\lambda_{\mathbf{X}} \triangleq \lambda + \sum_k \mathbf{W}^{ik}\mathbf{H}^{kj}$, the auxiliary variables $\mathbf{s_i} = \left\{s_i^{jk}, \forall j\right\}_{k=1}^{K+1}$ can be sampled as below

$$p\left(s_i^{j1}, \dots s_i^{jK}, s_i^{j(K+1)} \mid \text{rest}\right) = \text{Multinomial}\left(\frac{\mathbf{W}^{i1}\mathbf{H}^{1j}}{\lambda_{\mathbf{X}}}, \dots, \frac{\mathbf{W}^{iK}\mathbf{H}^{Kj}}{\lambda_{\mathbf{X}}}, \frac{\lambda}{\lambda_{\mathbf{X}}}\right) \tag{5}$$

Gibbs sampling update for $\mathbf{F}^{:,j}$ conditioned on the other variables can be derived similarly and given as

$$p\left(\mathbf{F}^{:,j} \mid \mathbf{Z}, \mathbf{W}, \mathbf{X}, \lambda, \mathbf{t_j}\right) \propto \Pi_{k=1}^{K}\left(\mathbf{F}^{kj}\right)^{c+\sum_i t_j^{ik}-1} \times \exp\left\{-\left(d + \sum_i \left(\mathbf{W}\mathbf{D}_j\right)^{ik}\right)\mathbf{F}^{kj}\right\} \tag{6}$$

where $\mathbf{D}_j$ denotes a diagonal matrix constructed from $\mathbf{Z}^{:,j}$, and the associated auxiliary variables $\mathbf{t_j} = \left\{t_j^{ik}, \forall i\right\}_{k=1}^{K+1}$ can be sampled similar to $\mathbf{s_i}$ as in (5).

The posterior updates of λ given $\mathbf{W}$ and $\mathbf{F}$ can be written as

$$p\left(\lambda \mid \mathbf{W}, \mathbf{Z}, \mathbf{F}, \mathbf{X}, \mathbf{r}\right) \propto \lambda^{a_\lambda + \sum_{i=1}^{D}\sum_{l=1}^{N} r^{ij}-1} \exp\left\{-\left(b_\lambda + DN\right)\right\} \tag{7}$$

and $\forall i, j$, the auxiliary variables r^{ij} can be sampled as

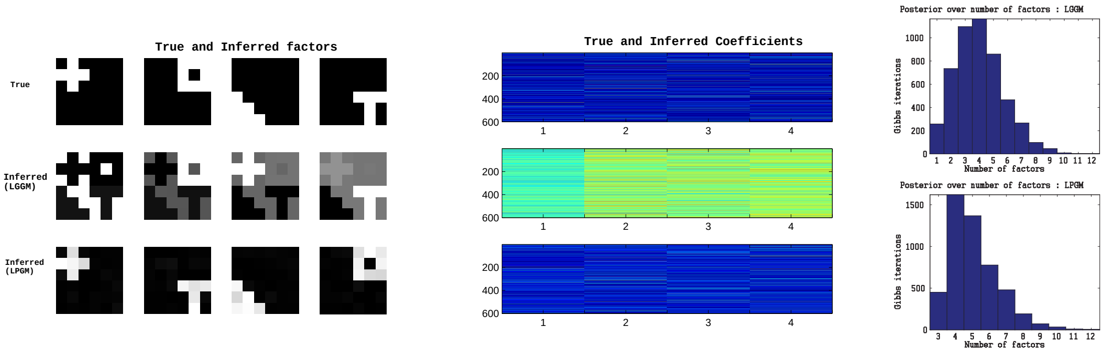
$$p\left(r^{ij} \mid \text{rest}\right) = \text{Binomial}\left(\mathbf{X}^{ij}, \frac{\lambda}{\mathbf{W}^{i,:}\mathbf{H}^{:,j} + \lambda}\right) \tag{8}$$

The matrix $\mathbf{Z}$ is inferred using the slice sampler of [9].

4 Demonstration

4.1 Synthetic Dataset

We first demonstrate the proposed model on a synthetic dataset used in [1]. This dataset has four non-overlapping binary factors with different shapes. These factors are linearly combined using random coefficients (uniformly distributed on $(0, 1)$) to generate 600 synthetic images. We further add a uniformly distributed random noise with support $(0, 0.1)$. Our goal is to learn



(a) (b) (c)

Figure 2: Synthetic dataset results (a) the true and inferred factors (b) the true and inferred coefficients; in both (a) and (b), the first row depicts the true values while the second and the third rows depict the inferred values for LGGM and LPGM respectively (c) Posterior over number of factors for LGGM (top) and LPGM (bottom).

the generative factors and their coefficients along with the number of factors automatically.

Since the factors can be thought of as different parts that are added via a linearly weighted combination, we expect that a model, which enforces *nonnegativity constraints* on the factors and the coefficients would perform better than a model without constraints. For this, we run two nonparametric models - Linear Gaussian-Gaussian model (LGGM) [9] and the proposed LPGM. Gibbs sampler of both models were run for 5000 iterations. While both models converged to four factors, it can be seen from Figure 2 that factors inferred by LPGM are *correct* (up to a permutation) while those learnt by LGGM are *mixed-up* due to the Gaussian-based modeling.

4.2 CBCL Face Image dataset

Our second dataset is a publicly available[1] MIT CBCL face image dataset. We use this dataset to show the utility of our model for dictionary learning. This dataset consists of 2429 grayscale face images (19×19 pixels) for training and 472 face images for testing. We map the pixel intensities on the scale between 0 and 255.

4.2.1 Dictionary Learning

We run our proposed model with a single Gibbs chain for 2500 iterations, which converges with 72 nonnegative factors (see the mode value in Figure 3d). Figure 3a shows the first 35 factors, which can be seen to be parts-based and sparse.

To compare the performance of our method, we use two recently developed nonparametric Bayesian techniques [9, 7] that use Gaussian distributions for $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{E}$. For inferring the number of factors, [9] uses IBP whereas [7] uses a truncated beta process. We refer to these models as LGGM and T-LGGM respectively. We run a single Gibbs chain of LGGM for 2500 iterations, which converges with 60 mixed-sign factors (see the mode value in Figure 3f). The experiments with T-LGGM[2] were conducted with identical settings, leading to 63 mixed-sign factors and these factors qualitatively look similar to the factors learnt using LGGM.

It can be seen from Figure 3b (showing first 35 factors) that the factors learnt using LGGM are of holistic nature and do not capture parts of the face. Additionally, as noted from Figure 3c and 3e, the inference for LGGM takes *longer* to converge than that of LPGM.

4.2.2 Generalization and Perplexity

We use the dictionary learnt above for generalization over new faces. For evaluation, we use *perplexity per image* - a measure that expresses the degree of surprise for a new image. Given the training set $\mathbf{X}_{D\times N}$ and a test set $\tilde{\mathbf{X}}_{D\times\tilde{N}}$, perplexity per image (PPI) is defined as

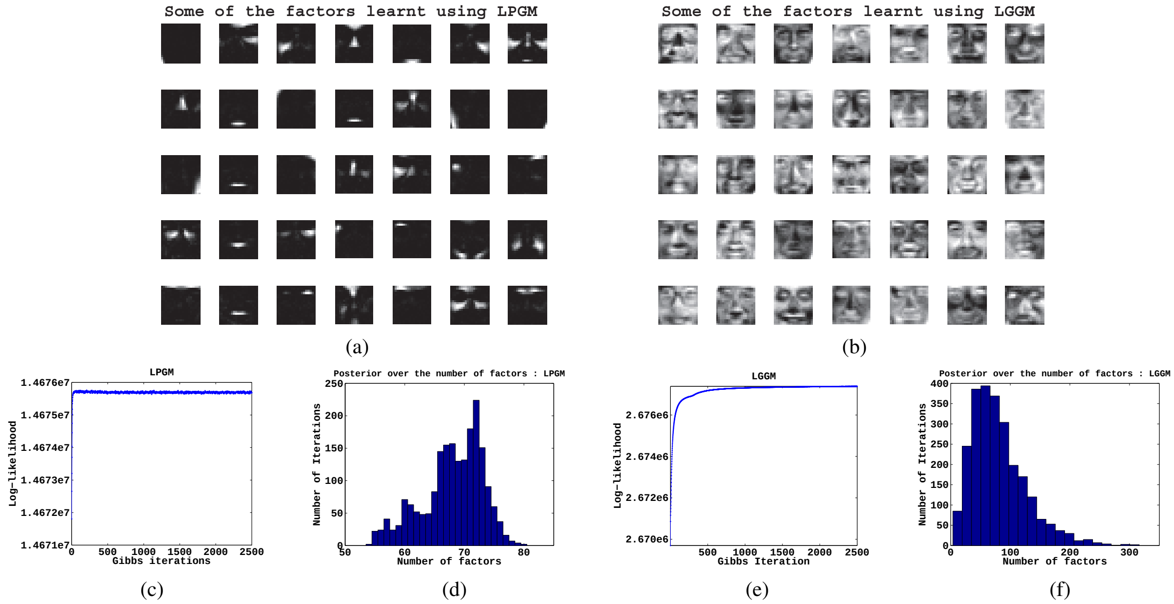
$$\text{PPI}\left(\tilde{\mathbf{X}}\right)=\exp\left(-\frac{1}{\tilde{N}}\log p\left(\tilde{\mathbf{X}}\mid\mathbf{X}\right)\right)\qquad(9)$$

A low value of perplexity implies a better generalization over the test data. It can seen from Table 1 that the proposed LPGM achieves much lower perplexity than LGGM and T-LGGM. This clearly indicates that LPGM has better modeling power for count data compared to the baselines.

We also analyze the *sparsity* of the factors obtained using the three models. For this, we use a sparsity index defined in [2]. This index maps the sparsity on the scale of 0 to 1 and a higher value implies more sparse factors. It can be seen from the Table 1 that sparsity index

[1] A normalized version is available at http://cbcl.mit.edu/cbcl/software-datasets/FaceData2.html

[2] Due to space limitation, we do not show the factors for T-iLGGM, however, we list the perplexity and sparsity results in Table 1.



(a) (b)

(c) (d) (e) (f)

Figure 3: Dictionary learning using LPGM and LGGM (a) the factors learnt using LPGM (b) the factors learnt using LGGM (c) Joint log-likelihood plot for LPGM (d) the posterior over the number of factors for LPGM (e) Joint log-likelihood plot for LGGM (f) the posterior over the number of factors for LGGM.

for LPGM is higher compared to LGGM and T-LGGM. Although, due to using beta process, T-LGGM factors are slightly more sparse than those of LGGM.

5 Conclusion

We have proposed a nonparametric Bayesian, linear Poisson gamma model for count data and applied it for dictionary learning. Crucially, this model captures parts-based representations in a similar vein to nonnegative matrix factorization. To make the inference tractable, we present an auxiliary variable Gibbs sampler. Our algorithm combines this inference procedure with a slice sampler [9] to learn the number of factors *automatically*. We demonstrate the model on both synthetic and real-world datasets for dictionary learning applications. Although, we have demonstrated our model on image data, it is generic and widely applicable to modeling of other count data e.g. text, video etc.

Table 1: A comparison by perplexity and sparsity .

Method	*Average Log Perplexity*	*Sparsity*
T-LGGM [7]	$2.413e5 \pm 759.8$	0.631
LGGM [9]	$2.267e5 \pm 1034.2$	0.546
LPGM	$\mathbf{1.512e5 \pm 987.5}$	**0.897**

References

[1] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. *NIPS*, 18:475, 2006.

[2] P. Hoyer. Non-negative matrix factorization with sparseness constraints. *JMLR*, 5:1457–1469, 2004.

[3] D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. *ICA*, pages 381–388, 2007.

[4] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[5] E. Meeds, Z. Ghahramani, R. Neal, and S. Roweis. Modeling dyadic data with binary latent factors. *NIPS*, 19:977–984, 2007.

[6] J. Paisley and D. Blei. Latent factor topic models with rank-reducing beta process priors. *NIPS Workshop on Low-rank Methods for Large-Scale Machine Learning, Whistler, BC*, 2010.

[7] J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. *ICML*, pages 777–784, 2009.

[8] G. Santhanam, B. Yu, K. Shenoy, and M. Sahani. Factor analysis with Poisson output. Technical report, NPSL-TR-06-1. Stanford Univ, CA, 2006.

[9] Y. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. *JMLR - Proceedings Track*, 2:556–563, 2007.

[10] M. Titsias. The infinite gamma-Poisson feature model. *NIPS*, 20:1513–1520, 2007.

[11] F. Wood, T. Griffiths, and Z. Ghahramani. A non-parametric Bayesian method for inferring hidden causes. In *UAI*, volume 22, 2006.

[12] M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative Binomial process and Poisson factor analysis. *Arxiv preprint arXiv:1112.3605*, 2011.