Deep Generative Models of Sparse and Overdispersed Discrete Data

He Zhao* **Piyush Rai†** **Lan Du*** **Wray Buntine*** **Dinh Phung***

Mingyuan Zhou‡

*Monash University, Australia †IIT Kanpur, India ‡UT Austin, USA

Abstract

In this paper, we propose a variational autoencoder based framework that generates discrete data, including both count-valued and binary data, via negative-binomial distribution. We also examine the model's ability to capture self- and cross-excitations in discrete data, which are critical for modelling overdispersion. We conduct extensive experiments on text analysis and collaborative filtering. Compared with several state-of-the-art baselines, the proposed models achieve significantly better performance on the above problems. By achieving superior modelling performance with a simple yet effect Bayesian extension to VAEs, we demonstrate that it is feasible to adapt the knowledge and experience of Bayesian probabilistic matrix factorisation into newly-developed deep generative models.

1 Introduction

In this paper, we are interested in handling high-dimensional, sparse, discrete matrices, where Probabilistic Matrix Factorisation (PMF) [19] has been a key method of choice for such data. For example, Latent Dirichlet Allocation (LDA) [2], Poisson Factor Analysis (PFA) [5, 30], and their deep extensions [3, 20, 10, 31, 26, 27] are representative models that generate data samples using the multinomial or Poisson distributions. The recent success of deep generative models like Variational Autoencoders (VAEs) [13, 21] on modelling real-valued data such as images has motivated machine learning practitioners to adapt VAEs to dealing with discrete data as done in recent works [17, 18, 14, 15]. Instead of using the Gaussian distribution as the data distribution for real-valued data, the multinomial distribution has been used for discrete data [17, 14, 15]. Following Liang et al. [15], we refer to these VAE-based models as "MultiVAE" (Multi for multinomial). However, existing VAE models such as MultiVAE can lead to inferior modelling performance on discrete data due to: **1)** insufficient capability of modelling overdispersion in count-valued data, and **2)** model misspecification in binary data. Specifically, *overdispersion* (i.e., the variance larger than the mean) describes the phenomenon that the data variability is large, which is a key property for large-scale count-valued data. For example, overdispersion in text data can behave as *word burstiness* [6, 16, 8, 4]. Shown in Zhou [29], the deep-seated cause of insufficient capability of modelling overdispersion in existing PMF models with Poisson or multinomial is their limited ability of handling *self-* and *cross-excitation* [29]. For example, in text data, self-excitation captures the effect that if a word occurs in a document, it is likely to occur more times in the same document, while cross-excitation models the effect that if a word such as "puppy" occurs, it will likely to excite the occurrences of related words such as "dog." On the other hand, model misspecification means that it may not be proper to directly apply multinomial or Poisson to binary data, which is a common misspecification in many existing models. This is because multinomial and Poisson may assign more than one count to one position, ignoring the fact that the data are binary, which could result in inferior performance [28].

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

Table 1: Comparison of the data distributions, model parameters, predictive rates, and posteriors. $\mathrm{q}(\cdot)$ denotes the encoder in VAE models.

Model	Data distribution	Model parameter	Predictive rate	Posterior
PFA	$\boldsymbol{y}_j \sim \text{Poisson}(\boldsymbol{l}_j)$	$\boldsymbol{l}_j = \boldsymbol{\Phi}\boldsymbol{\theta}_j$	$l'_{vj} \propto \sum_k^K \phi_{vk}\theta_{kj}$	$\boldsymbol{\Phi}, \boldsymbol{\theta}_j \sim \mathrm{p}(\boldsymbol{\Phi}, \boldsymbol{\theta}_j \mid \mathbf{Y}^{-ij})$
LDA	$\boldsymbol{y}_j \sim \text{Multi}(y_{\cdot j}, \boldsymbol{l}_j)$	$\boldsymbol{l}_j = \boldsymbol{\Phi}\boldsymbol{\theta}_j / \theta_{\cdot j}$	$l'_{vj} \propto \sum_k^K \phi_{vk}\theta_{kj}/\theta_{\cdot j}$	$\boldsymbol{\Phi}, \boldsymbol{\theta}_j \sim \mathrm{p}(\boldsymbol{\Phi}, \boldsymbol{\theta}_j \mid \mathbf{Y}^{-ij})$
MultiVAE	$\boldsymbol{y}_j \sim \text{Multi}(y_{\cdot j}, \boldsymbol{l}_j)$	$\boldsymbol{l}_j = \text{softmax}(f_\theta(\boldsymbol{z}_j))$	$l'_{vj} \propto \text{softmax}(f_\theta(\boldsymbol{z}_j))_v$	$\boldsymbol{z}_j \sim \mathrm{q}(\boldsymbol{z}_j \mid \mathbf{Y}^{-ij})$
NBFA	$\boldsymbol{y}_j \sim \text{NB}(\boldsymbol{l}_j, p_j)$	$\boldsymbol{l}_j = \boldsymbol{\Phi}\boldsymbol{\theta}_j$	$l'_{vj} \propto (y_{vj}^{-i} + \sum_k^K \phi_{vk}\theta_{kj})p_j$	$\boldsymbol{\Phi}, \boldsymbol{\theta}_j, p_j \sim \mathrm{p}(\boldsymbol{\Phi}, \boldsymbol{\theta}_j, p_j \mid \mathbf{Y}^{-ij})$
NBVAE	$\boldsymbol{y}_j \sim \text{NB}(\boldsymbol{r}_j, \boldsymbol{p}_j)$	$\boldsymbol{r}_j = \exp\left(f_{\theta^r}(\boldsymbol{z}_j)\right)$ $\boldsymbol{p}_j = \text{sigmoid}\left(f_{\theta^p}(\boldsymbol{z}_j)\right)$	$l'_{vj} \propto (y_{vj}^{-i} + \exp\left(f_{\theta^r}(\boldsymbol{z}_j)\right)_v) \cdot \text{sigmoid}\left(f_{\theta^p}(\boldsymbol{z}_j)\right)_v$	$\boldsymbol{z}_j \sim \mathrm{q}(\boldsymbol{z}_j \mid \mathbf{Y}^{-ij})$

In this paper, we show the above two issues on modelling discrete data can be addressed in a principled manner using the negative-binomial (NB) distribution as the data distribution in a VAE-based framework, called **N**egative-**B**inomial **V**ariational **A**uto**E**ncoder (**NBVAE** for short). Extensive experiments have been conducted on two important problems of discrete data analysis: text analysis on bag-of-words data and collaborative filtering on binary data. Compared with several state-of-the-art baselines, NBVAE achieves significantly better performance on the above problems.

2 Analytical study and model details

Here we start with the introduction of our proposed NBVAE model for count-valued data, and then give a detailed analysis on why NBVAE is capable of better handling self- and cross-excitations, and finally describe the variant of NBVAE for modelling binary data. Note that we focus on the generative process of the models and omit the details of the inference process due to the space limit.

Negative-binomial variational autoencoder (NBVAE): Without loss of generality, we present our model in the case of bag-of-word data for a text corpus, but the model can generally work with any kind of count-valued matrices. Suppose the bag-of-word data are stored in a V by N count matrix $\mathbf{Y} \in \mathbb{N}^{V\times N} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N]$, where $\mathbb{N} = \{0, 1, 2, \cdots\}$; N and V are the number of documents and the size of the vocabulary, respectively. To generate the occurrences of the words for the j^{th} ($j \in \{1, \cdots N\}$) document, $\boldsymbol{y}_j \in \mathbb{N}^V$, we draw a K dimensional latent representation $\boldsymbol{z}_j \in \mathbb{R}^K$ from a standard multivariate normal prior. After that, $\boldsymbol{y}_j$ is drawn from a (multivariate) negative-binomial distribution with $\boldsymbol{r}_j \in \mathbb{R}_+^V$ ($\mathbb{R}_+ = \{x : x \geq 0\}$) and $\boldsymbol{p}_j \in (0, 1)^V$ as the parameters. Moreover, $\boldsymbol{r}_j$ and $\boldsymbol{p}_j$ are obtained by transforming $\boldsymbol{z}_j$ from two nonlinear functions, $f_{\theta^r}(\cdot)$ and $f_{\theta^p}(\cdot)$, parameterised by θ^r and θ^p, respectively. The above process can be formulated as follows:

$$\boldsymbol{z}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \boldsymbol{r}_j = \exp\left(f_{\theta^r}(\boldsymbol{z}_j)\right), \boldsymbol{p}_j = \text{sigmoid}(f_{\theta^p}(\boldsymbol{z}_j)), \boldsymbol{y}_j \sim \text{NB}(\boldsymbol{r}_j, \boldsymbol{p}_j). \tag{1}$$

How NBVAE captures self- and cross-excitations: We now compare NBVAE and other PMF models in terms of their ability in capturing self- and cross-excitations in count-valued data. We first reformulate the models related to NBVAE into the above framework, including Poisson Factor Analysis (PFA) [5, 30], Latent Dirichlet Allocation (LDA) [2], MultiVAE [17, 14, 15], and Negative-Binomial Factor Analysis (NBFA) [29] into a unified presentation, shown in Table 1. In particular, we can show a model's capacity of capturing self- and cross-excitations by analysing its predictive distribution. Note that y_{vj}^{-i} denotes the number of v's occurrences in document j excluding the i^{th} word. If we compare PFA, LDA, MultiVAE V.S. NBFA, NBVAE, NBVAE_{dm}, it can be seen that the latter three models with NB as their data distributions explicitly capture self-excitation via the term y_{vj}^{-i} in the predictive distributions. Moreover, NBFA applies a single-layer linear combination of the latent representations, i.e., $\sum_k^K \phi_{vk}\theta_{kj}$, while NBVAE can be viewed as a deep extension of NBFA, using a deep neural network to conduct multi-layer nonlinear combinations of the latent representations, i.e, $r_{vj} = \exp\left(f_{\theta^r}(\boldsymbol{z}_j)\right)_v$ and $p_{vj} = \text{sigmoid}(f_{\theta^p}(\boldsymbol{z}_j))_v$. Therefore, NBVAE enjoys richer modelling capacity than NBFA on capturing cross-excitation.

NBVAE for binary data: Previous models like MultiVAE [14, 15] treat such binary data as counts, which is a model misspecification that is likely to result in inferior performance. Here we develop a simple yet effective method that links NBVAE and the Bernoulli distribution, as follows:

$$\boldsymbol{m}_j \sim \text{NB}(\boldsymbol{r}_j, \boldsymbol{p}_j), \;\; \boldsymbol{y}_j = \mathbf{1}(\boldsymbol{m}_j \geq 1), \tag{2}$$

where $\boldsymbol{r}_j$ and $\boldsymbol{p}_j$ have the same construction of the original NBVAE. Here we refer to this extension of NBVAE as NBVAE_{b} (b for binary).

Table 2: Perplexity comparisons. "Layers" indicate the architecture of the hidden layers (for VAE models, it is the hidden layer architecture of the encoder.). Best results for each dataset are in boldface. TLASGR and SGNHT are the algorithms of SGMCMC, detailed in the papers of DLDA [7] and DPFA [9]. Some results of the models with Gibbs sampling on RCV and Wiki are not reported because of the scalability issue.

Model	Inference	Layers	20NG	RCV	Wiki
DLDA	TLASGR	128-64-32	757	815	786
DLDA	Gibbs	128-64-32	752	802	-
DPFM	SVI	128-64	818	961	791
DPFM	MCMC	128-64	780	908	783
DPFA-SBN	Gibbs	128-64-32	827	-	-
DPFA-SBN	SGNHT	128-64-32	846	1143	876
DPFA-RBM	SGNHT	128-64-32	896	920	942
NBFA	Gibbs	128	690	702	-
MultiVAE	VAE	128-64	746	632	629
MultiVAE	VAE	128	772	786	756
NBVAE	VAE	128-64	688	**579**	**464**
NBVAE	VAE	128	714	694	529

(a) ML-10M (b) ML-20M (c) Netflix (d) MSD

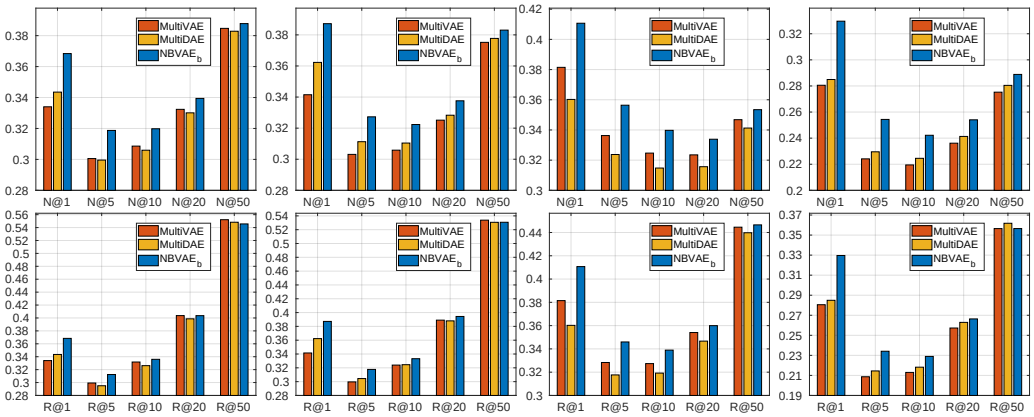

Figure 1: Comparisons of NDCG@R (N@R) and Recall$R$ (R@R). Standard errors in multiple runs are generally less than 0.003 for all the models on all the datasets, which are too tiny to show in the figures.

3 Experiments and conclusion

Experiments on text analysis: We used three widely-used corpora [22, 9, 12, 7]: 20 News Group (20NG), Reuters Corpus Volume (RCV), and Wikipedia (Wiki). Following Wallach et al. [24] we report per-heldout-word perplexity of all the models, which is a widely-used metric for text analysis, shown in Table 2. We compared ours with three categories of models for text analysis: **1)** Bayesian deep extensions of PFA and LDA: DLDA [7], DPFM [12], DPFA [10]; **2)** NBFA [29]; is a recently-proposed single-layer PMF with negative-binomial likelihood; **3)** MultiVAE [15, 14], a recent VAE model for discrete data with the multinomial distribution as the data distribution.

Experiments on collaborative filtering: We evaluate our models' performance on four user-item consumption datasets: MovieLens-10M (ML-10M), MovieLens-20M (ML-20M), Netflix Prize (Netflix), and Million Song Dataset (MSD) [1]. Following Liang et al. [15], we report two evaluation metrics: Recall@R and the truncated normalized discounted cumulative gain (NDCG@R), shown in Figure 1. As datasets used here are binary, we compared NBVAE$_\text{b}$, with the recent VAE models: **1)** MultiVAE. **2)** MultiDAE [15], a denoising autoencoder (DAE) with multinomial likelihood, which introduces dropout [23] at the input layer. MultiVAE and MultiDAE are the state-of-the-art VAE models for collaborative filtering and they have been reported to outperform several recent advances such as Wu et al. [25] and He et al. [11].

Conclusion: In this paper, we have proposed NBVAE and its variant to address the two issues of PMF and VAE models on discrete data: insufficient capability of modelling overdispersion in count-valued data and model misspecification in binary data. Our proposed models have achieved the state-of-the-art performance on text analysis and collaborative filtering. Longer version of our paper and the code is at `https://arxiv.org/abs/1905.00616` and `https://github.com/ethanhezhao/NBVAE`, respectively.

References

[1] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *International Conference on Music Information Retrieval*, 2011.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *JMLR*, 3: 993–1022, 2003.

[3] David M Blei, Thomas L Griffiths, and Michael I Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):7, 2010.

[4] Wray L Buntine and Swapnil Mishra. Experiments with non-parametric topic models. In *SIGKDD*, pages 881–890, 2014.

[5] John Canny. Gap: a factor model for discrete data. In *SIGIR*, pages 122–129, 2004.

[6] Kenneth W Church and William A Gale. Poisson mixtures. *Natural Language Engineering*, 1 (2):163–190, 1995.

[7] Yulai Cong, Bo Chen, Hongwei Liu, and Mingyuan Zhou. Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC. In *ICML*, pages 864–873, 2017.

[8] Gabriel Doyle and Charles Elkan. Accounting for burstiness in topic models. In *ICML*, pages 281–288, 2009.

[9] Zhe Gan, Changyou Chen, Ricardo Henao, David Carlson, and Lawrence Carin. Scalable deep Poisson factor analysis for topic modeling. In *ICML*, pages 1823–1832, 2015.

[10] Zhe Gan, R. Henao, D. Carlson, and Lawrence Carin. Learning deep sigmoid belief networks with data augmentation. In *AISTATS*, pages 268–276, 2015.

[11] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *WWW*, pages 173–182, 2017.

[12] Ricardo Henao, Zhe Gan, James Lu, and Lawrence Carin. Deep Poisson factor modeling. In *NIPS*, pages 2800–2808, 2015.

[13] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[14] Rahul Krishnan, Dawen Liang, and Matthew Hoffman. On the challenges of learning with inference networks on sparse, high-dimensional data. In *AISTATS*, pages 143–151, 2018.

[15] Dawen Liang, Rahul G Krishncan, Matthew D Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *WWW*, pages 689–698, 2018.

[16] Rasmus E Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the Dirichlet distribution. In *ICML*, pages 545–552, 2005.

[17] Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *ICML*, pages 1727–1736, 2016.

[18] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *ICML*, pages 2410–2419, 2017.

[19] Andriy Mnih and Ruslan R Salakhutdinov. Probabilistic matrix factorization. In *NIPS*, pages 1257–1264, 2008.

[20] John Paisley, Chong Wang, David M Blei, and Michael I Jordan. Nested hierarchical Dirichlet processes. *TPAMI*, 37(2):256–270, 2015.

[21] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286, 2014.

[22] Nitish Srivastava, Ruslan Salakhutdinov, and Geoffrey Hinton. Modeling documents with a deep Boltzmann machine. In *UAI*, pages 616–624, 2013.

[23] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

[24] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *ICML*, pages 1105–1112, 2009.

[25] Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. Collaborative denoising auto-encoders for top-n recommender systems. In *WSDM*, pages 153–162, 2016.

[26] He Zhao, Lan Du, and Wray Buntine. Leveraging node attributes for incomplete relational data. In *ICML*, pages 4072–4081, 2017.

[27] He Zhao, Lan Du, Wray Buntine, and Mingyuan Zhou. Dirichlet belief networks for topic structure learning. In *NeurIPS*, pages 7966–7977, 2018.

[28] MingYuan Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *AISTATS*, pages 1135–1143, 2015.

[29] Mingyuan Zhou. Nonparametric Bayesian negative binomial factor analysis. *Bayesian Analysis*, 2018.

[30] Mingyuan Zhou, Lauren Hannah, David B Dunson, and Lawrence Carin. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, pages 1462–1471, 2012.

[31] Mingyuan Zhou, Yulai Cong, and Bo Chen. Augmentable gamma belief networks. *JMLR*, 17 (163):1–44, 2016.