3-2021

Bilateral variational autoencoder for collaborative filtering

Quoc Tuan TRUONG

Aghiles SALAH

Hady W. LAUW

Bilateral Variational Autoencoder for Collaborative Filtering

Quoc-Tuan Truong
School of Information Systems
Singapore Management University
qttruong.2017@smu.edu.sg

Aghiles Salah
School of Information Systems
Singapore Management University
asalah@smu.edu.sg

Hady W. Lauw
School of Information Systems
Singapore Management University
hadywlauw@smu.edu.sg

ABSTRACT

Preference data is a form of dyadic data, with measurements associated with pairs of elements arising from two discrete sets of objects. These are users and items, as well as their interactions, e.g., ratings. We are interested in learning representations for both sets of objects, i.e., users and items, to predict unknown pairwise interactions. Motivated by the recent successes of deep latent variable models, we propose Bilateral Variational Autoencoder (BiVAE), which arises from a combination of a generative model of dyadic data with two inference models, user- and item-based, parameterized by neural networks. Interestingly, our model can take the form of a Bayesian variational autoencoder either on the user or item side. As opposed to the vanilla VAE model, BiVAE is "bilateral", in that users and items are treated similarly, making it more apt for two-way or dyadic data. While theoretically sound, we formally show that, similarly to VAE, our model might suffer from an over-regularized latent space. This issue, known as *posterior collapse* in the VAE literature, may appear due to assuming an over-simplified prior (isotropic Gaussian) over the latent space. Hence, we further propose a mitigation of this issue by introducing constrained adaptive prior (CAP) for learning user- and item-dependent prior distributions. Empirical results on several real-world datasets show that the proposed model outperforms conventional VAE and other comparative collaborative filtering models in terms of item recommendation. Moreover, the proposed CAP further boosts the performance of BiVAE. An implementation of BiVAE is available on Cornac recommender library.

KEYWORDS

Collaborative Filtering, Variational Autoencoder, Dyadic Data

ACM Reference Format:
Quoc-Tuan Truong, Aghiles Salah, and Hady W. Lauw. 2021. Bilateral Variational Autoencoder for Collaborative Filtering. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3437963.3441759

1 INTRODUCTION

Preference data in collaborative filtering (CF) typically consists of a set of users, a set of items, and a set of interactions, e.g., ratings, clicks, purchases between some user-item pairs. Prevalent and pervasive, it is found in many user-oriented applications in e-commerce and social media. The goal of representation learning [2] is to capture and encode latent patterns of observed data that can prove useful in downstream tasks, e.g., clustering, prediction, visualisation. In the case of preference data, it is to learn suitable representations that would help in recommending items to users.

Preference data is quintessentially *dyadic data*. These are measurements associated with pairs of outcomes arising from categorical random variables [16]. Naturally we seek representations for both sides of dyadic data (users and items) whose combination would be capable of explaining user-item affinities. To tackle this objective, latent factor or matrix factorization models are predominant in the context of CF [11, 17, 23, 32, 40]. The latter owe their success mainly to their simplicity, efficiency, effectiveness, and extensibility— one can easily combine them, for instance, to incorporate side information [28, 35]. Nevertheless, this category of models is also known to suffer from a limited modeling capacity as it can only capture linear patterns both in the data and latent spaces. To go beyond this limitation, there has recently been a surge of interest in using non-linear neural-based approaches [14, 26, 41, 48, 51].

Although these models have shown promising improvements over traditional factorization models in many cases, most of them may turn out to be challenging to train on sparse CF data due to their complexity combined with their deterministic nature. In fact, contrary to the data (e.g., images) arising in domains such as computer vision where deep neural architectures are successful, preference data is usually sparse. That is, the numbers of users and items are large—ranging from tens of thousands to millions—while the observed *dyads* are relatively few, often less than 1% out of all possible interactions, posing serious difficulties for the estimation and generalization of deep neural networks.

Notably, Variational Autoencoder (VAE) model [22] has been recently applied to CF with strong performance improvements over several competitive approaches [26]. One plausible explanation for the good results achieved by VAE on the CF task is its probabilistic nature. Indeed, the key difference of this model with neural networks is that VAE does not seek to learn deterministic representations, but rather learns distributions over these representations, thereby allowing it to account for uncertainty in the latent space. The latter property is particularly beneficial when dealing with sparse data where few observations are available. Despite its remarkable performance, VAE was originally designed for vector based-data, and thus is not in complete fidelity to the two-way nature of dyadic data, i.e., only users are explicitly represented, while items are treated as features in a vector space of users. In consequence of this mismatch between VAE and the two-way nature of preference data, it is not clear how one would extend such model on the item side in a principled way, for example to represent side information such as item textual descriptions, images, etc.

WSDM '21, March 8–12, 2021, Virtual Event, Israel

ACM ISBN 978-1-4503-8297-7/21/03…$15.00
https://doi.org/10.1145/3437963.3441759

As remedy to the above drawback, we propose Bilateral Variational Autoencoder (BiVAE) – *our first contribution*. It consists of a generative model of user-item interactions (or dyads), and *a pair* of inference models (user- and item-based respectively) parameterized using multilayer neural networks, all combined together in a unified framework to *auto-encode* dyadic preference data. Interestingly, BiVAE can be expressed as a Bayesian variational autoencoder either on the user or item side. As opposed to the vanilla VAE, the proposed BiVAE is "bilateral" in that it treats users and items symmetrically, making it more apt for two-way or dyadic data. In particular, BiVAE can capture uncertainty on both sides of dyadic data, which would improve its robustness and performance on sparse preference data, compared to classical *one-sided* variational autoencoders.

Inherited from the variational autoencoders family is an issue of over-regularized latent space, known in the literature as *posterior collapse*, i.e., the posterior is set equal to the prior. This phenomenon is largely due to assuming an over-simplified prior – isotropic Gaussian – over the latent space. Its occurrence causes the model to use only some part of its capacity, which may lead to underfitting or even uninformative representations. In this work, we further propose a mitigation of this issue by adopting adaptive user- and item-dependent prior distributions – *our second contribution*. To fit BiVAE to observations, we derive a scalable stochastic alternating optimization procedure intertwining the optimization of a user- and an item-based objective. As *the third contribution*, we conduct extensive experiments on several real-world datasets, showing that the proposed model outperforms strong CF models, including VAE, in terms of item recommendation. Moreover, we empirically demonstrate that the proposed priors further boost the performance of BiVAE and alleviate the posterior collapse phenomenon.

2 RELATED WORK

Our contributions lie in the intersection of several research topics, including latent factor models, neural networks, and variational autoencoders for collaborative filtering.

Latent Factor Models. Matrix factorization (MF) models are extensively studied [11, 23, 40, 42] in collaborative filtering. Existing variants differ in several ways, including the assumptions on the latent and data spaces (e.g., Gaussian [40] or Poisson [9, 11]), the nature of training objective (pointwise (scoring) loss [17, 40] or pairwise (ranking) loss [32]), etc. The proposed model BiVAE, can be viewed as a generalization of MF models. That is, beginning from our formulation and making some restrictive assumptions one can recover instances of probabilistic MF [9, 11, 40] (see Section 3.3). In this work, we focus on the pointwise approach for learning.

Neural Networks for Collaborative Filtering. Several works have considered using neural networks in CF. For instance, while Sedhain et al. [41], who rely on standard (deterministic) autoencoders, point out that one could model either side of CF data, they consider two different models, namely user-based and item-based autoencoders. Our work is different from two perspectives. First, we consider a probabilistic approach to autoencoders. Second, our model BiVAE auto-encodes users and items simultaneously under a unified objective, while in [41], the user and item's autoencoders are independent and separate models. Along the same line, [48] uses a user-based denoising autoencoder (DAE), and further extends DAE with user-specific embeddings by introducing an additional input node. More recently, [14, 46] introduce a neural architecture for CF, which can be viewed as a non-linear extension of matrix factorization. They also explicitly model both users and items, and explore different ways of combining their representations to explain observations, including the widely used scalar product, multilayer perceptron, and the combination of both. In our experiments, we include [14] as a baseline. There are also methods that rely on neural network to represent auxiliary data [30, 47], but these approaches still rely on standard MF to model user-item interactions.

VAE for Collaborative Filtering. Despite its success in data generation and representation, VAE [22] receive relatively scant attention in the CF literature. A recent work of Liang et al. [26] ignites interest on VAE for CF[1], by demonstrating performance improvements over competitive baselines. Concurrently to [26], Lee et al. [24] also consider the VAE framework for CF, but the latter focus on conditional and joint VAE formulation to incorporate user auxiliary data. Karamanolakis et al. [19] also rely on VAE for personalized recommendation and further explore user-dependant priors. More recently, Kim and Suh [20] investigate VAE with the VampPrior [45]. Lobel et al. [27] propose an actor-critic reinforcement learning method to train VAE to approximately maximize a ranking-based metric. Shenbin et al. [43] explore different regularization strategies to improve VAE for collaborative filtering. These works differ from ours; they still adopt the original formulation of VAE, which is asymmetric and unfaithful to the two-way nature of dyadic data, i.e., only users are explicitly represented, while items are treated as features in a vector space of users. In principle, some of the above methods (e.g., [27, 43]) can also be applied to the proposed BiVAE model for further improvements, which we leave to future work.

Posterior Collapse in VAE Models. Significant efforts have been expended on alleviating the posterior collapse issue, which may affect VAE in practice [1, 8]. These include annealing the problematic KL term in the variational lower bound [6, 26, 44], weakening the decoder to force a reliance on the latent representations [6, 49], replacing the KL term by another regularizer, e.g., adversarial one [29, 50], or adopting rich priors [45, 50]. We follow the latter line of efforts and act on the priors to mitigate the posterior collapse issue. Differently from previous work using a shared prior across data points, we consider heterogeneous, i.e., user- and item-dependent priors, that can adapt during training. In particular we build such prior using external features, extracted from available user and item side information. In the context of one-sided VAE, Karamanolakis et al. [19] also explore the use of heterogeneous user priors built from external user features. However, in their case the priors are held fixed during learning. Moreover, while they make a mention that such priors may reduce the effect of posterior collapse, they have not reported any experimental verification on this aspect.

3 METHODOLOGY

We introduce a new hierarchical generative model of dyads (user-item interactions) and along with it the corresponding user and item inference models, parameterized using neural networks. When combined together, the generative and inference models give rise to a variational autoencoder either on the user or item side.

[1]Note that some works have considered VAE to represent auxiliary data [25, 36], but not to model user-item interactions. In this work, we are interested in the latter.

3.1 Bilateral Variational Autoencoder

The data that we seek to learn from is the user-item preference matrix, of size $U \times I$, denoted $\mathbf{R} = (r_{ui})$, where r_{ui} is the interaction, e.g., integer rating, between user u and item i. We use the notation $\mathbf{r}_{u*}$ to refer to the row in $\mathbf{R}$ corresponding to user u. Similarly, $\mathbf{r}_{*i}$ refers to the ith column of $\mathbf{R}$. The latent variables are the per user and item representations denoted respectively $\boldsymbol{\theta}_u$, $\boldsymbol{\beta}_i \in \mathbb{R}^K$.

Generative Model of Dyads. Figure 1 (middle) depicts our generative model in plate notations. The latent variables are drawn from prior distributions. Without loss of generality, we use Gaussian priors with diagonal covariance matrices. We further follow the common practice and adopt the standard multivariate isotropic Gaussian as the prior over all user/item latent variables. That is, $p(\boldsymbol{\theta}_u) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $p(\boldsymbol{\beta}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\forall i, u$.

Conditional on the latent variables, the observations are drawn from a univariate exponential family,

$$\begin{aligned} p(r_{ui}|\boldsymbol{\theta}_u, \boldsymbol{\beta}_i) &= \text{EXPFAM}(r_{ui}; \eta(\boldsymbol{\theta}_u; \boldsymbol{\beta}_i; \omega)) \\ &= h(r_{ui}) \exp\{\eta(\boldsymbol{\theta}_u; \boldsymbol{\beta}_i; \omega) r_{ui} - a(\eta(\boldsymbol{\theta}_u; \boldsymbol{\beta}_i; \omega))\} \end{aligned} \tag{1}$$

where $h(\cdot)$, $\eta(\cdot)$ and $a(\cdot)$ denote respectively the base measure, natural parameter and log-normalizer of the exponential family [3, 7]. For simplicity, we have assumed that r_{ui} is the sufficient statistic by itself. This form of the exponential family still encompasses many popular univariate distributions, including the Poisson, Bernoulli, Gaussian with unit variance, Gamma with fixed shape parameter, etc. Therefore, our framework can accommodate various types of preference data, such as counts, binary, continuous, etc. We further parameterize the conditional likelihood in such a way that,

$$\mathbb{E}(r_{ui}|\boldsymbol{\theta}_u, \boldsymbol{\beta}_i) = \frac{da(\eta)}{d\eta} = g_\omega(\boldsymbol{\theta}_u; \boldsymbol{\beta}_i) \tag{2}$$

where $g_\omega(\cdot)$ is some differentiable function (e.g., inner product, neural network, etc.) parameterized by ω, combining the latent representations to output the mean of the observation r_{ui}.

As a concrete example, consider the Poisson distribution, which we use in our experiment,

$$p(r_{ui}|\boldsymbol{\theta}_u, \boldsymbol{\beta}_i) = \frac{1}{r_{ui}!} \exp\{r_{ui} \log g_\omega(\boldsymbol{\theta}_u; \boldsymbol{\beta}_i) - g_\omega(\boldsymbol{\theta}_u; \boldsymbol{\beta}_i)\}. \tag{3}$$

We notice that $h(\cdot) = 1/(r_{ui}!)$, $\eta(\cdot) = \log g_\omega(\boldsymbol{\theta}_u; \boldsymbol{\beta}_i)$, and the log-normalizer $a(\eta) = \exp \eta(\cdot) = g_\omega(\boldsymbol{\theta}_u; \boldsymbol{\beta}_i)$.

Given some $\mathbf{R}$, the goal is to find the values of the parameters ω that would most likely have generated the observations, and to infer the posterior over the latent variables $p(\boldsymbol{\theta}_{1:U}, \boldsymbol{\beta}_{1:I}|\mathbf{R})$. The latter will allow us to make predictions about unknown preferences and form recommendations. However, the posterior and likelihood are intractable and thereby, exact inference and learning are infeasible. We therefore resort to variational Bayes (VB) [4, 18], a popular and efficient approach to deal with complex probabilistic models.

Inference Model. The starting point of VB is to introduce a tractable inference model q, governed by a set of *variational parameters* ν, which will be used as a proxy for the true but intractable posterior [3]. We choose a variational distribution that breaks the coupling between β and θ—a main source of intractability in our model, i.e., $q(\boldsymbol{\theta}_{1:U}, \boldsymbol{\beta}_{1:I}|\mathbf{R}) = q(\boldsymbol{\theta}_{1:U}|\mathbf{R}) q(\boldsymbol{\beta}_{1:I}|\mathbf{R})$, with $q(\boldsymbol{\theta}_{1:U}|\mathbf{R}) = \prod_u q(\boldsymbol{\theta}_u|\mathbf{r}_{u*})$, and $q(\boldsymbol{\beta}_{1:I}|\mathbf{R}) = \prod_i q(\boldsymbol{\beta}_i|\mathbf{r}_{*i})$. In our experiments, without loss of generality, we adopt factors of the following form:

$$q(\boldsymbol{\theta}_u|\mathbf{r}_{u*}) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_{\tilde{\psi}}(\mathbf{r}_{u*}), \tilde{\boldsymbol{\sigma}}_{\tilde{\psi}}(\mathbf{r}_{u*})),\ q(\boldsymbol{\beta}_i|\mathbf{r}_{*i}) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_{\tilde{\phi}}(\mathbf{r}_{*i}), \tilde{\boldsymbol{\sigma}}_{\tilde{\phi}}(\mathbf{r}_{*i})).$$

where $\nu = \{\tilde{\phi}, \tilde{\psi}\}$, $\tilde{\boldsymbol{\mu}}(\cdot)$ and $\tilde{\boldsymbol{\sigma}}(\cdot)$ are vector-valued functions – we use multilayer perceptrons (MLPs) in this work – parameterized by $\tilde{\phi}/\tilde{\psi}$, outputting respectively the mean and covariance parameters of the variational distributions.

With q in place, we proceed with approximate inference and learning by optimizing the Evidence Lower BOund (ELBO), w.r.t. the model ω and variational ν parameters, given in our case by,

$$\begin{aligned} \mathcal{L} = \sum_{u,i} &\mathbb{E}_{q(\boldsymbol{\theta}_u|\mathbf{r}_{u*})} \mathbb{E}_{q(\boldsymbol{\beta}_i|\mathbf{r}_{*i})} [\log p(r_{ui}|\boldsymbol{\theta}_u, \boldsymbol{\beta}_i)] \\ &- \sum_u \text{KL}(q(\boldsymbol{\theta}_u|\mathbf{r}_{u*})||p(\boldsymbol{\theta}_u)) - \sum_i \text{KL}(q(\boldsymbol{\beta}_i|\mathbf{r}_{*i})||p(\boldsymbol{\beta}_i)). \end{aligned} \tag{4}$$

As we shall see shortly, this objective is closely related to that of Variational Autoencoders (VAE) [22] either on the user or item side. For this reason, we refer to the model arising from the combination of the above generative and inference models, p and q respectively, as *Bilateral Variational Autoencoder (BiVAE)*.

3.2 Constrained Adaptive Priors (CAP)

While the objective (4) is theoretically sound, optimizing it in practice may result in over-simplified representations for users and items. This is due to the KL terms encouraging the posteriors to forget observations $\mathbf{R}$ by matching them to the same simple prior distribution. This is a known issue in the VAE literature [6, 8], often referred to as *posterior collapse* issue [31]. In the following, we characterize this phenomenon formally to gain more insights that motivate our solution, i.e., the use of constrained adaptive priors.

Consider Proposition 1 extending the scope of the result of Hoffman and Johnson [15] to BiVAE for dyadic data. The proof is at the end of this section. Eqs. (5) and (6) make clear the effect of the ELBO's KL terms on the latent variables. By minimizing the latter divergences, we are also minimizing the mutual information between the latent representations and the identity of each user/item.

Proposition 1 *Let $q(u, i) = \frac{1}{U \times I}$ denote the empirical distribution over user-item pairs, and $q(u) = \frac{1}{U}$, $q(i) = \frac{1}{I}$ the corresponding marginals. Define, $p(\boldsymbol{\theta}) \triangleq p(\boldsymbol{\theta}_u)$, $p(\boldsymbol{\beta}) \triangleq p(\boldsymbol{\beta}_i)$, $q(\boldsymbol{\theta}|r_{u*}) \triangleq q(\boldsymbol{\theta}_u|r_{u*})$, $q(\boldsymbol{\beta}|r_{*i}) \triangleq q(\boldsymbol{\beta}_i|r_{*i})$, and the user, item aggregated posteriors $q(\boldsymbol{\theta}) = \frac{1}{U} \sum_{u=1}^U q(\boldsymbol{\theta}|\mathbf{r}_{u*})$, $q(\boldsymbol{\beta}) = \frac{1}{I} \sum_{i=1}^I q(\boldsymbol{\beta}|\mathbf{r}_{*i})$. Then, the sums over the KL terms in the ELBO (4) can be expressed as follows,*

$$\sum_u KL(q(\boldsymbol{\theta}_u|\mathbf{r}_{u*})||p(\boldsymbol{\theta}_u)) = U \cdot [KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta})) + \mathbb{I}_q(\boldsymbol{\theta}, u)], \tag{5}$$

$$\sum_i KL(q(\boldsymbol{\beta}_i|\mathbf{r}_{*i})||p(\boldsymbol{\beta}_i)) = I \cdot [KL(q(\boldsymbol{\beta})||p(\boldsymbol{\beta})) + \mathbb{I}_q(\boldsymbol{\beta}, i)], \tag{6}$$

where $\mathbb{I}_q(\boldsymbol{\theta}, u)$ is the mutual information of $\boldsymbol{\theta}$ and u w.r.t. $q(\boldsymbol{\theta}, u)$, whose marginals are the $q(\boldsymbol{\theta})$ and $q(u)$ defined above. Similarly, $\mathbb{I}_q(\boldsymbol{\beta}, i)$ stands for the mutual information of $\boldsymbol{\beta}$ and i w.r.t. $q(\boldsymbol{\beta}, i)$.

In other words, the KL regularizers encourage the latent space to be independent from users and items. One easy way to achieve this is to set the posteriors equal to the prior, corresponding to the extreme scenario of posterior collapse. If this happens, it will cause the model not to use all its expressive capacity, or even lead to

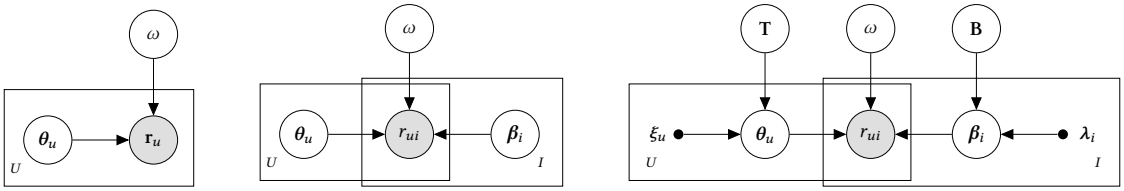


Figure 1: Graphical representations of VAE (left), BiVAE (middle), and BiVAE with Constrained Adaptive Priors (right).

useless representations for subsequent tasks such as personalized recommendation which we focus on.

To mitigate this potential issue, we propose to lower the effect of the KL regularization by adopting user- and item-dependant priors, which can adapt during learning. One might be tempted to choose priors that are expressive enough to perfectly match the posteriors, and thereby cancel the KL terms in the ELBO. However, such an extreme choice is not desirable either, as we will end up mainly optimizing the conditional likelihood $p(r_{ui}|\boldsymbol{\theta}_u, \boldsymbol{\beta}_u)$, which would lead to overfitting. To avoid falling in the latter case, we rather propose relatively weak priors (i.e., constrained on *a priori* knowledge), the only parameters adapted during learning are shared either across users and items. We further restrict the covariance matrix to be the identity for all priors, so as to reflect low confidence about preferences, as well as encourage the posteriors to spread their mass to capture uncertainty. Formally assume the following priors,

$$p(\boldsymbol{\theta}_u) = \mathcal{N}(\mathbf{T}\boldsymbol{\xi}_u, \mathbf{I}), \quad p(\boldsymbol{\beta}_i) = \mathcal{N}(\mathbf{B}\boldsymbol{\lambda}_i, \mathbf{I}), \tag{7}$$

where $\boldsymbol{\xi}_u \in \mathbb{R}^{k'}$ and $\boldsymbol{\lambda}_i \in \mathbb{R}^{k''}$ denote user- and item-specific fixed hyperparameters, while $\mathbf{T}$ and $\mathbf{B}$ of size $(K \times K')$ and $(K \times K'')$ are free learnable continuous parameters. The hyperparameters $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$ represent some prior knowledge about users and items reflecting their affinities. In this work, we rely on side information to compute them, for more details see Section 4. The graphical model of BiVAE augmented with these adaptive priors, is given in Figure 1 (right).

Proof. For brevity we prove eq. (5) for the user-based term only.

$$\begin{aligned}
\mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta})) + \mathbb{I}_q(\boldsymbol{\theta}, u) &= \mathbb{E}_{q(\boldsymbol{\theta})}[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta})] \\
&\quad + \mathbb{E}_{q(\boldsymbol{\theta},u)}[\log q(\boldsymbol{\theta}|u) + \log q(u) - \log q(\boldsymbol{\theta}) - \log q(u)] \\
&\overset{a}{=} \mathbb{E}_{q(u)q(\boldsymbol{\theta}|u)}[\log q(\boldsymbol{\theta}|u) - \log p(\boldsymbol{\theta})] \\
&\overset{b}{=} \sum_u \frac{1}{U} \mathbb{E}_{q(\boldsymbol{\theta}|u)}[\log q(\boldsymbol{\theta}|u) - \log p(\boldsymbol{\theta})],
\end{aligned} \tag{8}$$

where we have canceled some terms and rearranged the remaining one in a, and taken the expectation over $q(u)$ in b. Using the fact that, $q(\boldsymbol{\theta}|u) = \int q(\boldsymbol{\theta}|\mathbf{r})q(\mathbf{r}|u)d\mathbf{r} = \int q(\boldsymbol{\theta}|\mathbf{r})\delta(\mathbf{r}-\mathbf{r}_{u*})d\mathbf{r} = q(\boldsymbol{\theta}|\mathbf{r}_{u*})$, with $\delta(\cdot)$ denoting a Dirac distribution, noting that the notations $p(\boldsymbol{\theta})$, $p(\boldsymbol{\theta}_u)$ and $q(\boldsymbol{\theta}|\mathbf{r}_{u*})$, $q(\boldsymbol{\theta}_u|\mathbf{r}_{u*})$ are by definition equivalent and refer to a same prior, posterior respectively, and multiplying (8) by U completes the proof. ■

3.3 Connections to Existing Work

Matrix Factorization. The model we propose can be viewed as a generalization of different types of Matrix Factorization (MF) models. For instance, we can recover Probabilistic Gaussian MF (PMF) [40] from BiVAE, by choosing the exponential family in (1) to be a Gaussian, setting $g_\omega(\boldsymbol{\theta}_u, \boldsymbol{\beta}_i) = \boldsymbol{\theta}_u^\top \boldsymbol{\beta}_i$, and substituting free user and item variational parameters for the inference networks. As another example, we can obtain Bayesian Poisson Factorization [9, 11] with Gaussian latent factors, by assuming a Poisson likelihood, parameterizing the variation distribution without inference networks, and letting $g_\omega(\boldsymbol{\theta}_u, \boldsymbol{\beta}_i)$ to be some non-negative function of $\boldsymbol{\theta}_u^\top \boldsymbol{\beta}_i$. We include [11] and the latter as our baselines in the experiment.

Variational Autoencoder. As mentioned earlier, BiVAE corresponds to a Bayesian variational autoencoder [22], over users (resp., items) in which case the latent variables $\boldsymbol{\beta}_{1:I}$ (resp., $\boldsymbol{\theta}_{1:U}$) play the role of global parameters shared across users (resp., items). This can be noted by rewriting the ELBO as follows,

$$\begin{aligned}
\mathcal{L}^u = \sum_u \mathbb{E}_{q(\boldsymbol{\theta}_u|\mathbf{r}_{u*})} \mathbb{E}_{q(\boldsymbol{\beta}_{1:I}|\mathbf{R})} [\log p(\mathbf{r}_{u*}|\boldsymbol{\theta}_u, \boldsymbol{\beta}_{1:I})] \\
- \sum_u \mathrm{KL}(q(\boldsymbol{\theta}_u|\cdot)||p(\boldsymbol{\theta}_u)) - \mathrm{KL}(q(\boldsymbol{\beta}_{1:I}|\cdot)||p(\boldsymbol{\beta}_{1:I})),
\end{aligned} \tag{9}$$

where $p(\mathbf{r}_{u*}|\boldsymbol{\theta}_u, \boldsymbol{\beta}_{1:I}) = \prod_i p(\mathbf{r}_{ui}|\boldsymbol{\theta}_u, \boldsymbol{\beta}_i)$. A similar expression of the ELBO holds for the items,

$$\begin{aligned}
\mathcal{L}^i = \sum_i \mathbb{E}_{q(\boldsymbol{\beta}_i|\mathbf{r}_{*i})} \mathbb{E}_{q(\boldsymbol{\theta}_{1:U}|\mathbf{R})} [\log p(\mathbf{r}_{*i}|\boldsymbol{\theta}_{1:U}, \boldsymbol{\beta}_i)] \\
- \sum_i \mathrm{KL}(q(\boldsymbol{\beta}_i|\cdot)||p(\boldsymbol{\beta}_i)) - \mathrm{KL}(q(\boldsymbol{\theta}_{1:U}|\cdot)||p(\boldsymbol{\theta}_{1:U})),
\end{aligned} \tag{10}$$

where $p(\mathbf{r}_{*i}|\boldsymbol{\theta}_{1:U}, \boldsymbol{\beta}_i) = \prod_u p(\mathbf{r}_{ui}|\boldsymbol{\theta}_u, \boldsymbol{\beta}_i)$. As we shall see shortly, the above allows us to derive an efficient stochastic alternating procedure to fit BiVAE to observations, which intertwines the optimization of a user- and item-based VAEs.

Empirical Bayes. In statistics, Empirical Bayes (EB) refers to an inference method where the prior distribution is estimated from observations. Under the parametric EB scenario, we usually perform point estimate of a parameter α of a prior distribution $p(\boldsymbol{\theta}_u|\alpha)$ by maximizing the the corresponding marginal likelihood $p(\mathbf{r}_{u*}|\alpha)$ (or its approximation). This is reminiscent of our approach with adaptive priors, where we learn the prior parameters $\mathbf{T}$ and $\mathbf{B}$ from data by maximizing the ELBO. One key difference though is that we still hold some hyperparameters fixed to avoid overfitting as argued earlier. Hence, one can see our using adaptive prior as a form of EB treatment for BiVAE. Interestingly, since EB can be viewed as an approximation to a full Bayes approach, using adaptive prior constitutes a basis for a full variational Bayes treatment of BiVAE.

3.4 Optimization

In practice, we rely on stochastic optimization to fit BiVAE to observations. While the KL terms in the ELBO are available analytically, the expectations over the conditional log-likelihood are intractable

and thereby, the direct optimization of (4) is not possible. To overcome this difficulty, we rely on the *reparameterization trick* [22, 33] and build an unbiased Monte Carlo estimator of (4), which yields:

$$\tilde{\mathcal{L}} = \sum_{u,i} \log p(r_{ui}|\tilde{\boldsymbol{\theta}}_u, \tilde{\boldsymbol{\beta}}_i) - \sum_u \mathrm{KL}(q(\boldsymbol{\theta}_u|\mathbf{r}_{u*})||p(\boldsymbol{\theta}_u)) - \sum_i \mathrm{KL}(q(\boldsymbol{\beta}_i|\mathbf{r}_{*i})||p(\boldsymbol{\beta}_i)) \tag{11}$$

where $\tilde{\boldsymbol{\theta}}_u = \mathcal{T}(\boldsymbol{\epsilon}, \tilde{\boldsymbol{\psi}}) = \tilde{\boldsymbol{\mu}}_{\tilde{\psi}}(\mathbf{r}_{u*}) + \tilde{\boldsymbol{\sigma}}_{\tilde{\psi}}(\mathbf{r}_{u*}) \odot \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Similarly, $\tilde{\boldsymbol{\beta}}_i = \mathcal{T}(\boldsymbol{\epsilon}, \tilde{\boldsymbol{\phi}}) = \tilde{\boldsymbol{\mu}}_{\tilde{\phi}}(\mathbf{r}_{*i}) + \tilde{\boldsymbol{\sigma}}_{\tilde{\phi}}(\mathbf{r}_{*i}) \odot \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Now all the quantities involved in (11) are tractable. However, performing unbiased stochastic optimization over the above objective is not convenient, due to the mixing between r_{ui}, $\mathbf{r}_{u*}$ and $\mathbf{r}_{*i}$. To overcome this difficulty and ease subsampling of observations, we propose to exploit the two-way nature of our model and advocate alternate optimization in a Gauss-Seidel fashion. Precisely, we organize BiVAE parameters into two blocks consisting of user-related and item-related parameters respectively, then alternate the optimization of each block while holding the other one fixed.

While item related parameters remain unchanged, BiVAE's learning problem boils down to optimizing the following objective w.r.t. user-specific parameters,

$$\tilde{\mathcal{L}}^u = \sum_u \left[\log p(\mathbf{r}_{u*}|\tilde{\boldsymbol{\theta}}_u, \tilde{\boldsymbol{\beta}}_{1:I}) - \mathrm{KL}(q(\boldsymbol{\theta}_u|\mathbf{r}_{u*})||p(\boldsymbol{\theta}_u)) \right], \tag{12}$$

where $p(\mathbf{r}_{u*}|\tilde{\boldsymbol{\theta}}_u, \tilde{\boldsymbol{\beta}}_{1:I}) = \prod_i p(r_{ui}|\tilde{\boldsymbol{\theta}}_u, \tilde{\boldsymbol{\beta}}_i)$, and we have dropped constant terms. Analogously, holding user-related parameters fixed gives rise to the following item-based objective up to a constant,

$$\tilde{\mathcal{L}}^i = \sum_i \left[\log p(\mathbf{r}_{*i}|\tilde{\boldsymbol{\beta}}_i, \tilde{\boldsymbol{\theta}}_{1:U}) - \mathrm{KL}(q(\boldsymbol{\beta}_i|\mathbf{r}_{*i})||p(\boldsymbol{\beta}_i)) \right], \tag{13}$$

with $p(\mathbf{r}_{*i}|\tilde{\boldsymbol{\theta}}_{1:U}, \tilde{\boldsymbol{\beta}}_i) = \prod_u p(r_{ui}|\tilde{\boldsymbol{\theta}}_u, \tilde{\boldsymbol{\beta}}_i)$. Note that the above objectives (12) and (13) are not conflicting; the maximization of either of them corresponds to the maximization of BiVAE's ELBO (11).

In practice, we rely on stochastic gradient ascent and alternate the maximization of (12) and (13) w.r.t. user- and item- specific parameters, respectively. We evaluate the different gradients using automatic differentiation, and we further scale them using ADAM [21]. To nearly optimize the ELBO with respect to each block's parameters before moving to the next one, we take several stochastic gradient steps over each objective. Algorithm 1 summarizes our block stochastic optimization procedure.

Complexity Analysis. The computational bottleneck of Algorithm 1 is with the forward/backward passes – essentially matrix operations – of the neural networks $g_\omega(\cdot)$, $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\sigma}(\cdot)$ used to parameterize the different distributions. Assuming network architectures with only one hidden layer, the complexity of one optimization epoch over the user and item objectives is $O(U \cdot I \cdot K + I \cdot U \cdot K)$. Hence, the asymptotic time complexity of one epoch of Algorithm 1 is $O(U \cdot I \cdot K)$, which is the same as the standard VAE. By leveraging sparse matrix multiplication, the above complexity can be reduced to $O(N \cdot K)$, with N denoting the number of non-zeros entries in the preference matrix $\mathbf{R}$, in practice $N \ll I \cdot U$. Moreover, the two loops (user-based and item-based optimization) in Algorithm 1 are independent and can be parallelized.

Algorithm 1 Block Stochastic Optimization for BiVAE

Input: $\mathbf{R}$, $\boldsymbol{\xi}$, $\boldsymbol{\lambda}$, g_ω, $\tilde{\boldsymbol{\mu}}_{\tilde{\psi}}$, $\tilde{\boldsymbol{\sigma}}_{\tilde{\psi}}$, $\tilde{\boldsymbol{\mu}}_{\tilde{\phi}}$, $\tilde{\boldsymbol{\sigma}}_{\tilde{\phi}}$. m: mini-batch size.
Output: ω, $\tilde{\phi}$, $\tilde{\psi}$.
repeat
  User-based objective (12) optimization
  Sample $\{\mathbf{r}_{*1}, \ldots, \mathbf{r}_{*I}\}$ from observations.
  Sample $\{\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_I\}$ from posterior $q(\boldsymbol{\beta}|\mathbf{r})$.
  Sample $\{\mathbf{r}_{1*}, \ldots, \mathbf{r}_{m*}\}$ from observations.
  Sample $\{\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_m\}$ from posterior $q(\boldsymbol{\theta}|\mathbf{r})$.
  Update ω, $\tilde{\psi}$ by taking a gradient ascent step.
  Item-based objective (13) optimization
  Sample $\{\mathbf{r}_{1*}, \ldots, \mathbf{r}_{U*}\}$ from observations.
  Sample $\{\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_U\}$ from posterior $q(\boldsymbol{\theta}|\mathbf{r})$.
  Sample $\{\mathbf{r}_{*1}, \ldots, \mathbf{r}_{*m}\}$ from observations.
  Sample $\{\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_m\}$ from posterior $q(\boldsymbol{\beta}|\mathbf{r})$.
  Update ω, $\tilde{\phi}$ by taking a gradient ascent step.
until convergence

4 EXPERIMENTS

In this section we evaluate the performance of the proposed BiVAE model. We are interested in the following experimental objectives: (i) investigating the performance of BiVAE as compared to VAE and other competitive collaborative filtering models, and (ii) examining the effect of the proposed constrained adaptive priors (CAP) on recommendation and determining whether it can alleviate posterior collapse in BiVAE.

4.1 Setup

Datasets. We use a total of seven publicly available benchmark datasets exhibiting various characteristics. Table 1 provides the statistics of the different datasets, after preprocessing if applicable. For all datasets, we binarize the integer ratings by treating all available user-item interactions as positive feedback.

These include three MovieLens [2] datasets of varying sizes, namely *ML-100K*, *ML-1M*, and *ML-20M*. In the literature, the former two datasets are considered to be relatively dense, as every user has at least 20 ratings.

We also experiment with several datasets including side information that we can leverage to compute our adaptive priors. Three are from *Amazon.com* covering various product categories, namely *Office*, *Clothing*, and *Sports*, which are made available by He and McAuley [13]. In addition to user-item interactions, these datasets come with user review texts (*#docs*) as well as item relations/network (*#i-rels*) in the form of *Also-Viewed* information, which we leverage to compute the hyperparameters $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$ of our constrained adaptive priors. The last dataset is Epinions [3], containing a user social network (*#u-rels*), which we exploit to build adaptive user-dependent priors. We retain only users and items with at least five ratings.

Comparative Baselines. We benchmark the proposed BiVAE model with comparable collaborative filtering models:

[2] https://grouplens.org/datasets/movielens
[3] https://snap.stanford.edu/data/soc-Epinions1.html

Table 1: Data Statistics

Dataset	#users	#items	#feedback	density	#docs	#u-rels	#i-rels
ML-100K	943	1,682	100,000	6.30%	-	-	-
ML-1M	6,040	3,706	1,000,209	4.47%	-	-	-
ML-20M	138,493	26,744	20,000,263	0.54%	-	-	-
Office	4,855	2,359	51,453	0.45%	51,453	-	28,190
Clothing	39,145	22,275	272,765	0.03%	272,765	-	235,894
Sports	72,464	26,640	439,650	0.02%	439,650	-	602,624
Epinions	23,247	59,451	553,354	0.04%	-	374,009	-

- **HPF** (Hierarchical Poisson Factorization) [11] combines Gamma latent factors with Poisson likelihood. HPF has shown strong performance compared to other popular MF-based models.
- **Gauss-PF** (Gaussian-Poisson Factorization) is an MF-based model combining Gaussian latent factors with a Poisson likelihood. This form of MF arises as a special case from the proposed BiVAE as described in Section 3.3.
- **NeuMF** (Neural Matrix Factorization) [14]. This approach relies on both neural networks, MLPs, and scalar product to combine user and item embeddings and map them to preferences. It has shown promising improvements over competitive similarity-based as well as factorization-based approaches.
- **VAE** (Variational Autoencoders for Collaborative Filtering) [26] has recently shown strong performance on the item recommendation task. As discussed in the paper, BiVAE is more adequate for modeling dyadic data, such as CF data, than VAE. Hence, we consider VAE as the main baseline to assess our contributions.

Evaluation Metrics. We assess the item recommendation accuracy on the held-out test set with two standard measures, *NDCG* and *Recall*, for top-M recommendation [5, 38]. We have varied the value of $M = 10, 20, 50$ and observed similar trends. As representative, we report the results at $M = 50$ due to space limitation.

Experimental Settings. For all the datasets, we randomly split observed preferences as follows, 80% for training set, 10% for validation set, and 10% for test set. We use Normalized Discounted Cumulative Gain (*NDCG*) to tune the different models based on the held-out validation sets. The number of latent dimensions K for user and item representations/factors is set to 20 for parity among comparative methods. We do not observe significant change when we go beyond that size. For the autoencoder family, we rely on multilayer perceptrons (MLPs) to parameterize the inference and generative models. In a pilot study, we explore MLPs with 0, 1, 2, and 3 hidden layers, finally retaining 1-hidden layer with 40 dimensions ($2 \times K$) for the inference models, and 0-hidden layer for the decoders. We find that going deeper does not improve performance, while introducing additional overhead in terms of tuning and computational time. This confirms the findings of Liang et al. [26], who also arrived at the same settings. We retain `Tanh` as a nonlinear activation function at every hidden layer, which we found to offer better performance than when using `ReLU` function. Since we focus on binary data, the outputs of the different decoders are passed through `Sigmoid` function. The search spaces are respectively: $\{1e^{-4}, \ldots, 1e^{-1}\}$ with multiples of 10 for learning rate, and $\{100, \ldots, 500\}$ with steps of 100 for number of epochs. Batch size is set to 128 for all methods optimized using Adam.

4.2 Comparisons with Baselines

Table 2 reports the performance of the different competing models across all datasets and metrics. BiVAE adopts a Poisson likelihood by default, unless stated otherwise. For VAE, in addition to a *multinomial* likelihood (VAE-*Mult*), which has been recommended in [26], we also report results with a Poisson likelihood (VAE-*Poiss*) for a fair comparison with BiVAE. Moreover, we conducted two-tailed paired t-tests to assess the statistical significance of the results. The main observations from the above table are summarized as follows.

BiVAE consistently outperforms the comparative baselines. From Table 2 we observe that BiVAE achieves the highest performance among all the methods. In particular, BiVAE's improvements over VAE are statistically significant in almost all cases, except on Epinions in terms of *Recall* where BiVAE and VAE are tight. These results provide positive support for the importance of our formulation taking into account the dyadic or two-way nature of preference data.

BiVAE is robust regarding likelihood choices. As mentioned earlier, we assume binary feedback. However so far we adopt a Poisson likelihood for BiVAE, a choice which is motivated by the success of this likelihood in the context of CF even under binary feedback scenarios [11]. Nevertheless, a Bernoulli likelihood would be more natural in such scenario, and we propose therefore to investigate this alternative. Table 3 summarizes the results of VAE and BiVAE with the two likelihoods. For both models, we observe the two likelihoods are very competitive. Moreover, we notice that regardless of the likelihood, BiVAE still outperforms VAE, and the results are consistent across likelihood types.

4.3 Effects of Priors

As discussed in Section 3.2 and in the VAE literature [45], the common practice of using a Standard Gaussian (SG) prior may cause the model to learn a latent space that does not use all its expressive capacity. We also observe this phenomenon with BiVAE as we shall see shortly. This motivates us to investigate an alternative choice for the priors, namely the Constrained Adaptive Prior (CAP) introduced in Section 3.2. In the subsequent experiments, we focus on assessing the impact of CAP on BiVAE's performance.

We compute CAP's hyperparameters $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$ a priori from side information using VAE. Hence, for the following experiments we only consider datasets which come with user/item side information, namely Office, Clothing, Sports, and Epinions. For the first three datasets, we have access to user review texts, as well as item network extracted from the *Also-Viewed* information. For every user we concatenate all his reviews associated with training data into one document[4], we then rely on VAE to represent the user documents with 20-dimensional embeddings, which are used as our fixed hyperparameters $\boldsymbol{\xi}_u$. We follow the same procedure to compute the hyperparameters $\boldsymbol{\lambda}_i$ from the item network. Similarly, for *Epinions* dataset, we compute $\boldsymbol{\xi}$ from adjacency matrix of the user social network. No item side information is available for *Epinions*. Note that, when side information is not available for a given user/item we simply use the standard Gaussian as the prior.

Table 4 depicts the performance of BiVAE under the SG and CAP priors. Focusing on item recommendation, the proposed CAP priors consistently improve the recommendation accuracy (*NDCG*

[4]We use a binary bag-of-word representations.

Table 2: Quantitative results. The notation * indicates that the improvement over the best VAE model (VAE-*Poiss* or VAE-*Mult*) are statistically significant (p-value < 0.01). NeuMF failed to scale to ML-20M, thus results are not available for this case.

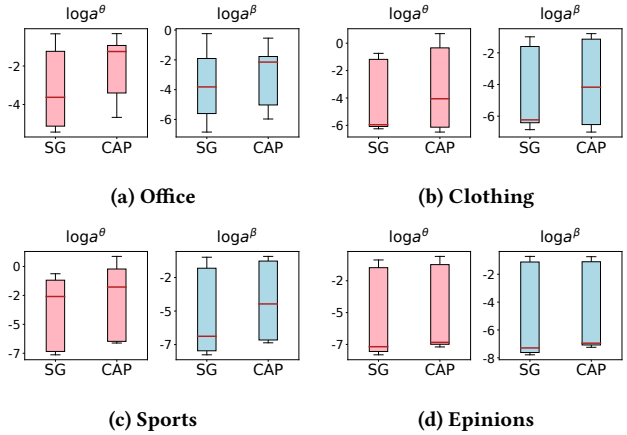
Method	ML-100K		ML-1M		ML-20M		Office		Clothing		Sports		Epinions	
	NDCG	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*
HPF	0.1837	0.3573	0.1379	0.2399	0.1382	0.2612	0.0514	0.1468	0.0084	0.0240	0.0157	0.0437	0.0365	0.0891
Gauss-PF	0.1918	0.3822	0.1395	0.2424	0.1559	0.2968	0.0637	0.1776	0.0172	0.0461	0.0179	0.0482	0.0231	0.0536
NeuMF	0.1803	0.3704	0.1408	0.2554	N/A	N/A	0.0630	0.1750	0.0179	0.0523	0.0202	0.0594	0.0403	0.1005
VAE-*Mult*	0.1838	0.3614	0.1466	0.2624	0.1561	0.3050	0.0593	0.1946	0.0230	0.0714	0.0365	0.1074	0.0473	0.1213
VAE-*Poiss*	0.1877	0.3749	0.1447	0.2666	0.1553	0.2987	0.0634	0.1894	0.0250	0.0729	0.0366	0.1077	0.0487	0.1218
BiVAE	**0.1947***	**0.3843***	**0.1539***	**0.2809***	**0.1602***	**0.3106***	**0.0700***	**0.2012***	**0.0284***	**0.0794***	**0.0386***	**0.1100***	**0.0498***	**0.1219**

Table 3: Comparison of *VAE* and *BiVAE* with *Bernoulli (Bern)* and *Poisson (Poiss)* likelihood functions.

Method	ML-100K		ML-1M		ML-20M		Office		Clothing		Sports		Epinions	
	NDCG	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*
VAE-*Bern*	0.1875	0.3709	0.1475	0.2629	0.1534	0.2947	0.0598	0.1845	0.0262	0.0754	0.0382	0.1082	0.0445	0.1146
VAE-*Poiss*	0.1877	0.3749	0.1447	0.2666	0.1553	0.2987	0.0634	0.1894	0.0250	0.0729	0.0366	0.1077	0.0487	0.1218
BiVAE-*Bern*	0.1890	0.3778	0.1499	0.2766	0.1589	0.3095	0.0699	0.1997	**0.0286**	**0.0798**	0.0380	0.1086	**0.0499**	**0.1243**
BiVAE-*Poiss*	**0.1947**	**0.3843**	**0.1539**	**0.2809**	**0.1602**	**0.3106**	**0.0700**	**0.2012**	0.0284	0.0794	**0.0386**	**0.1100**	0.0498	0.1219

Table 4: Comparison of the performance of *BiVAE* under different priors. For *NDCG* and *Recall*, the notation * indicates statistically significant improvements (p-values < 0.01).

Dataset	Prior	Metric		Active Units	
		NDCG	*Recall*	θ *(user)*	β *(item)*
Office	SG	0.0700	0.2012	10	10
	CAP	**0.0736***	**0.2078***	**14**	**13**
Clothing	SG	0.0284	0.0794	8	9
	CAP	**0.0313***	**0.0894***	**10**	**10**
Sports	SG	0.0356	0.1100	10	9
	CAP	**0.0410***	**0.1184***	**11**	**10**
Epinions	SG	0.0498	0.1219	8	8
	CAP	**0.0505***	**0.1223**	8	8



(a) Office **(b) Clothing**

(c) Sports **(d) Epinions**

Figure 2: Distribution of the statistic $\log a^{\theta/\beta}$ under the two types of priors *SG* and *CAP* (higher box is more active).

and *Recall*) across all datasets. These improvements are statistically significant, as measured by two-tailed paired t-tests, except in terms of *Recall* on Epinions. One possible explanation for that could be the lack of availability of side information on Epinions as we do not have access to item side information while social information is only available for about half of the users (density is 0.069%).

4.4 Assessing Posterior Collapse

We now analyze latent variable collapse in BiVAE and examining whether using *CAP* alleviates this issue.

4.4.1 Latent Variable Activity. Measuring latent variable collapse is a challenging task. Here we follow the same approach as in [8] and assess the "activity" of every user and every item latent dimension θ_k, β_k using the statistic $a_k^\theta = \mathrm{Cov}_{p(\mathbf{r}_u)}(\mathbb{E}_{q(\boldsymbol{\theta}|\mathbf{r}_u)}[\theta_k])$. For items, we define a_k^β analogously. The above statistic measures the variance of the expectation of the latent dimension θ_k (resp., β_k) across users (resp., items). Hence, if a dimension θ_k/β_k encodes useful information, we expect it to vary across users/items, i.e., a_k^θ would be relatively high. Figure 2 shows the distributions –in the form of boxplots– of $\log a_k^\theta$ and $\log a_k^\beta$, with $k \in \{1, \ldots, 20\}$, under the two types of priors, namely *SG* and *CAP*. We notice a global trend of increase (higher box and median) in the variance a_k of the latent dimensions when using *CAP*. In other words, *CAP* induces greater activity and more informative priors.

4.4.2 Number of Active Units (AU). We also observe that the distribution of a_k (for both users and items) is bi-modal, with widely separated modes as illustrated by Figure 3 on various datasets. This pattern shows that there are two natural clusters of latent dimensions. In particular, one group seems to have low variances a_k suggesting that these dimensions have collapsed to their prior means. Based on this observation we can define the number of Active Units (AU) of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ as follows [8, 10]: $\mathrm{AU}^{\theta/\beta} = \sum_k \mathbb{1}[a_k^{\theta/\beta} > th]$. We set $th = 10^{-3}$ as the evident bi-modality pattern in Figure 3 suggests that AU is not very sensitive to this threshold.

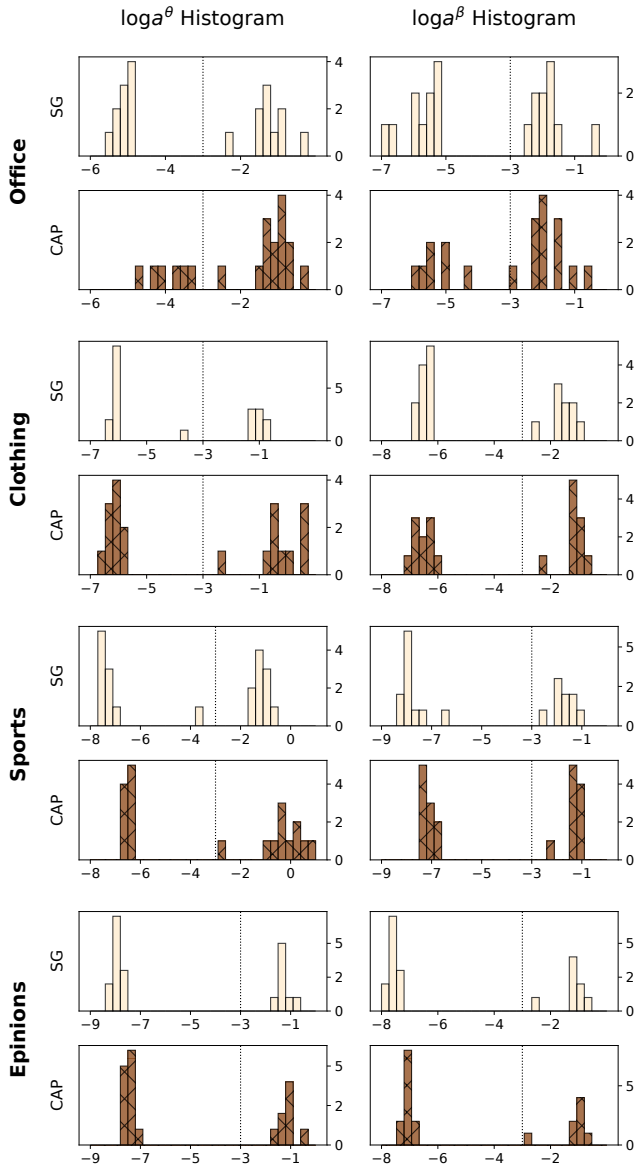


Figure 3: Histogram of $\log a^{\theta/\beta}$ under the two types of priors *SG* and *CAP*. The dash lines indicate cutoff threshold th separating active and inactive units.

The numbers of active units or dimensions for both the user and item latent representations are depicted in Table 4 (right side), under different prior choices. In all cases we observe that AU is substantially lower (≤ 14) than the total number of latent dimensions (20). Using *CAP* results in more active latent dimensions in most situations, except on Epinions. Recall for the latter dataset, the CAP priors are leveraged by only about half of users, the ones for which we have access to their social network. These results show that *CAP* effectively lessens the posterior collapse issue. As this correlates positively with the improvements in recommendation accuracy (as measured by *NDCG* and *Recall*), we argue that using *CAP* priors allows BiVAE to learn richer latent spaces.

Table 5: *BiVAE* performance under different priors. The notation * indicates statistically significant improvements (p-values < 0.01) over the second best priors.

Dataset	Metric	Mog	VampPrior	HPrior	CAP
Office	*NDCG*	0.0714	0.0712	0.0703	**0.0736***
	Recall	0.2030	0.2018	0.2033	**0.2078***
Clothing	*NDCG*	0.0271	0.0261	0.0301	**0.0313***
	Recall	0.0792	0.0737	0.0882	**0.0894**
Sports	*NDCG*	0.0364	0.0363	0.0394	**0.0410***
	Recall	0.1031	0.1026	0.1145	**0.1184***
Epinions	*NDCG*	0.0486	0.0456	0.0490	**0.0505***
	Recall	0.1173	0.1134	0.1201	**0.1223**

4.4.3 Alternative Priors. In addition to improving upon the Standard Gaussian (*SG*) prior, *CAP* offers competitive performance compared to other types of priors, most of which have been introduced recently to tackle posterior collapse. In particular, we consider the MoG (Mixture of Gaussians) prior [45], the VampPrior (Variational Mixture of Posteriors Prior) [45], which forms a rich prior distribution by mixing the variational posteriors using learnable pseudo inputs, and the HPrior (Heterogeneous Prior) [19] using fixed user- and item-dependent priors estimated from side information.

We set all the covariance matrices to the identity, for uncertainty purposes as discussed in Section 3.2. For *VampPrior* and *MoG*, we set the number of mixture components to 20, higher numbers result in decreased performances. As we focus on binary data, for *VampPrior* the pseudo inputs are constrained to lie in $[0, 1]$ using the Sigmoid.

The results are depicted in Table 5, *CAP* consistently outperforms the other priors across all datasets. These result reveal that learning priors solely from sparse user-item interactions may be challenging, as reflected by the lower performance of *MoG* and *VampPrior*. Please refer to the analysis in Section 3.2 for more insights on this aspect. Additional user/item features are promising, *HPrior* outperforms the other baselines in many cases. *CAP* further improves upon *HPrior* by adapting the user/item features, thus mitigating potential mismatches between external features and user-item signals.

5 CONCLUSION

We present BiVAE, a new variational autoencoder tailored for dyadic data, where observations consist of measurements associated with two sets of objects, e.g., users, items and corresponding ratings. It combines a generative model of dyads with two inference models parameterized using neural networks, to autoencode users and items under a unified framework. Interestingly, it can take the form of a Bayesian VAE either on user or item side. We provide strong theoretical foundations for the proposed model and discuss the connections to other existing work. We further propose a simple way to mitigate posterior collapse, by using constrained adaptive priors (*CAP*). Extensive experiments on seven real-world datasets show that BiVAE achieves significant improvements over a number of approaches, including conventional VAE, which has recently proven strong performance on item recommendation. Future work could be investigating other ways of building informative priors, applying BiVAE to other types of dyadic data such as document-word matrices, and other tasks such as co-clustering [12, 34, 37]. BiVAE's implementation is available on Cornac [39].

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its NRF Fellowship Programme (Award No. NRF-NRFF2016-07).

REFERENCES

[1] Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. 2018. Fixing a Broken ELBO. In *International Conference on Machine Learning, ICML*. 159–168.

[2] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 8 (2013), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

[3] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational Inference: A Review for Statisticians. *J. Amer. Statist. Assoc.* 112, 518 (2017), 859–877.

[5] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowl.-Based Syst.* 46 (2013), 109–132.

[6] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating Sentences from a Continuous Space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 10–21.

[7] L. D. Brown. 1986. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Institute of Mathematical Statistics.

[8] Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. 2016. Importance Weighted Autoencoders. In *International Conference on Learning Representations, ICLR*.

[9] Ali Taylan Cemgil. 2009. Bayesian inference for nonnegative matrix factorisation models. *Comp. Int. and Neurosc.* 2009 (2009), 785152:1–785152:17.

[10] Adji B Dieng, Yoon Kim, Alexander M Rush, and David M Blei. 2019. Avoiding latent variable collapse with generative skip models. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (2019).

[11] Prem Gopalan, Jake M. Hofman, and David M. Blei. 2015. Scalable Recommendation with Hierarchical Poisson Factorization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI*. 326–335.

[12] G. Govaert and M. Nadif. 2013. *Co-clustering*. John Wiley & Sons.

[13] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 144–150.

[14] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW*. 173–182.

[15] Matthew D Hoffman and Matthew J Johnson. 2016. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, Annual Conference on Neural Information Processing Systems*.

[16] Thomas Hofmann, Jan Puzicha, and Michael I. Jordan. 1998. Learning from Dyadic Data. In *Advances in Neural Information Processing Systems, NeurIPS*. 466–472.

[17] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining ICDM*. 263–272.

[18] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. An Introduction to Variational Methods for Graphical Models. *Machine Learning* 37, 2 (1999), 183–233.

[19] Giannis Karamanolakis, Kevin Raji Cherian, Ananth Ravi Narayan, Jie Yuan, Da Tang, and Tony Jebara. 2018. Item Recommendation with Variational Autoencoders and Heterogeneous Priors. In *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems, DLRS@RecSys*. 10–14.

[20] Daeryong Kim and Bongwon Suh. 2019. Enhancing VAEs for collaborative filtering: flexible priors & gating mechanisms. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys*. 403–407.

[21] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations, ICLR*.

[22] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations, ICLR*.

[23] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer* 42, 8 (2009), 30–37.

[24] Wonsung Lee, Kyungwoo Song, and Il-Chul Moon. 2017. Augmented Variational Autoencoders for Collaborative Filtering with Auxiliary Information. In *Proceedings of the ACM Conference on Information and Knowledge Management, CIKM*. 1139–1148.

[25] Xiaopeng Li and James She. 2017. Collaborative Variational Autoencoder for Recommender Systems. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 305–314.

[26] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the World Wide Web Conference on World Wide Web, WWW*. 689–698.

[27] Sam Lobel, Chunyuan Li, Jianfeng Gao, and Lawrence Carin. 2020. RaCT: Toward Amortized Ranking-Critical Training For Collaborative Filtering. In *International Conference on Learning Representations, ICLR*.

[28] Hao Ma, Haixuan Yang, Michael R. Lyu, and Irwin King. 2008. SoRec: social recommendation using probabilistic matrix factorization. In *Proceedings of the ACM Conference on Information and Knowledge Management, CIKM*. 931–940.

[29] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2016. Adversarial autoencoders. In *4th International Conference on Learning Representation (ICLR), Workshop*.

[30] Trong T Nguyen and Hady W Lauw. 2017. Collaborative topic regression with denoising autoencoder for content and community co-representation. In *ACM on Conference on Information and Knowledge Management, CIKM*. 2231–2234.

[31] Ali Razavi, Aäron van den Oord, Ben Poole, and Oriol Vinyals. 2019. Preventing Posterior Collapse with delta-VAEs. In *International Conference on Learning Representations, ICLR*.

[32] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence, UAI*. 452–461.

[33] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *International Conference on Machine Learning, ICML*. 1278–1286.

[34] Aghiles Salah, Melissa Ailem, and Mohamed Nadif. 2018. Word co-occurrence regularized non-negative matrix tri-factorization for text data co-clustering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[35] Aghiles Salah and Hady W. Lauw. 2018. A Bayesian Latent Variable Model of User Preferences with Item Context. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*. 2667–2674.

[36] Aghiles Salah and Hady W. Lauw. 2018. Probabilistic Collaborative Representation Learning for Personalized Item Recommendation. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*.

[37] Aghiles Salah and Mohamed Nadif. 2017. Model-based von Mises-Fisher Coclustering with a Conscience. In *Proceedings of the 2017 SIAM International Conference on Data Mining, SDM*. SIAM, 246–254.

[38] Aghiles Salah and Mohamed Nadif. 2017. Social regularized von Mises–Fisher mixture model for item recommendation. *Data Mining and Knowledge Discovery* 31, 5 (2017), 1218–1241.

[39] Aghiles Salah, Quoc-Tuan Truong, and Hady W. Lauw. 2020. Cornac: A Comparative Framework for Multimodal Recommender Systems. *Journal of Machine Learning Research* 21, 95 (2020), 1–5.

[40] Ruslan Salakhutdinov and Andriy Mnih. 2007. Probabilistic Matrix Factorization. In *Advances in Neural Information Processing Systems NeurIPS*. 1257–1264.

[41] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. 2015. AutoRec: Autoencoders Meet Collaborative Filtering. In *Proceedings of the International Conference on World Wide Web (WWW)*. 111–112.

[42] Hanhuai Shan and Arindam Banerjee. 2010. Generalized Probabilistic Matrix Factorizations for Collaborative Filtering. In *ICDM, The 10th IEEE International Conference on Data Mining*. 1025–1030.

[43] Ilya Shenbin, Anton Alekseev, Elena Tutubalina, Valentin Malykh, and Sergey I Nikolenko. 2020. RecVAE: A New Variational Autoencoder for Top-N Recommendations with Implicit Feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM*. 528–536.

[44] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. 2016. Ladder Variational Autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3738–3746.

[45] Jakub M. Tomczak and Max Welling. 2018. VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics, AISTATS*. 1214–1223.

[46] Quoc-Tuan Truong and Hady Lauw. 2019. Multimodal review generation for recommender systems. In *The World Wide Web Conference*. 1864–1874.

[47] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative Deep Learning for Recommender Systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1235–1244.

[48] Yao Wu, Christopher DuBois, Alice X. Zheng, and Martin Ester. 2016. Collaborative Denoising Auto-Encoders for Top-N Recommender Systems. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*. 153–162.

[49] Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved Variational Autoencoders for Text Modeling using Dilated Convolutions. In *International Conference on Machine Learning, ICML*. 3881–3890.

[50] Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2018. Adversarially Regularized Autoencoders. In *International Conference on Machine Learning, ICML*. 5897–5906.

[51] Yin Zheng, Bangsheng Tang, Wenkui Ding, and Hanning Zhou. 2016. A Neural Autoregressive Approach to Collaborative Filtering. In *International Conference on Machine Learning, ICML*. 764–773.