
Learning Correlated Latent Representations with Adaptive Priors

Da Tang

Columbia University
datang@cs.columbia.edu

Dawen Liang

Netflix Inc.
dliang@netflix.com

Nicholas Ruozzi

The University of Texas at Dallas
nicholas.ruozzi@utdallas.edu

Tony Jebara

Columbia University & Spotify Inc.
jebara@cs.columbia.edu

Abstract

Variational Auto-Encoders (VAEs) have been widely applied for learning compact, low-dimensional latent representations of high-dimensional data. When the correlation structure among data points is available, previous work proposed Correlated Variational Auto-Encoders (CVAEs), which employ a structured mixture model as prior and a structured variational posterior for each mixture component to enforce that the learned latent representations follow the same correlation structure. However, as we demonstrate in this work, such a choice cannot guarantee that CVAEs capture all the correlations. Furthermore, it prevents us from obtaining a tractable joint and marginal variational distribution. To address these issues, we propose Adaptive Correlated Variational Auto-Encoders (ACVAEs), which apply an adaptive prior distribution that can be adjusted during training and can learn a tractable joint variational distribution. Its tractable form also enables further refinement with belief propagation. Experimental results on link prediction and hierarchical clustering show that ACVAEs significantly outperform CVAEs among other benchmarks.

1 INTRODUCTION

Variational Auto-Encoders (VAEs) [13, 23] are a family of deep generative models that learn latent embeddings for data. By applying variational inference on the latent variables, VAEs learn a stochastic mapping from high-dimensional data to low-dimensional representations, which can be used for many downstream tasks, including classification, regression, and clustering.

VAEs assume the data points are *i.i.d.* generated and treat the model and posterior approximations as factorized over data points. However, if we know *a priori* that there is structured correlation between the data points, e.g., for graph-structured datasets [24, 3, 8, 27], correlated variational approximations can help. Tang et al. [27] proposed Correlated Variational Auto-Encoders (CVAEs), which take this kind of correlation structure as auxiliary information to guide the variational approximations for the latent embeddings by constructing a prior from a uniform mixture of tractable distributions on *maximal acyclic subgraphs* of the given undirected correlation graph.

However, there are several limitations that potentially prevent CVAEs from learning better correlated latent embeddings. First, it is possible that some of the maximal acyclic subgraphs of the given graph can, by themselves, well-capture the correlation between the data points while others may poorly capture the correlation. As a result, taking a uniform average may yield a sub-optimal result. Second, while the prior in CVAEs is over multiple subgraphs, each subgraph has a unique joint variational

distribution, and there is no single *global* joint variational distribution over the latent variables. CVAEs do learn pairwise variational approximation functions, but they are **not** exact pairwise marginal variational distributions on the latent variables. As a result, applying these variational approximation functions to some downstream tasks, e.g. link prediction, may result in poor performances due to the inexact approximations. In addition, CVAEs require a pre-processing step that takes an amount of time cubic in the number of vertices, which limits its applicability to smaller datasets.

To address these issues, we propose Adaptive Correlated Variational Auto-Encoders (ACVAEs), which chooses a non-uniform average over tractable distributions over the maximal acyclic subgraphs as a prior. This prior is adaptive, and will be adjusted during optimization. To learn the mixture weights, we provide two options, empirical Bayes or saddle-point optimization, both of which maximize the objective with respect to the model and variational parameters. The difference is that while empirical Bayes also maximizes the objective with respect to the prior structure, saddle-point optimization seeks to optimize the objective under the *worst* prior for more robust inference. In both cases, the non-uniform average converges to a tractable prior on a single graph, which ensures that we obtain a holistic tractable joint variational distribution. With this variational distribution, we obtain exact marginal evaluation using exact inference algorithms, e.g., belief propagation. Moreover, ACVAEs do not require the cubic time pre-processing step embedded in CVAEs, and they are generally faster for evaluation in practice. We demonstrate the superior empirical performance of ACVAEs for link prediction and hierarchical clustering on various real datasets.

2 VAES WITH CORRELATIONS

In this section, we provide a brief overview of Variational Auto-Encoders (VAEs) [13, 23] as well as Correlated Variational Auto-Encoders (CVAEs) [27], which take the correlation structure among data points into consideration.

2.1 Variational Auto-Encoders

We use a latent variable model to fit data $\mathbf{x} = \{x_1, \dots, x_n\} \subset \mathbb{R}^D$. The model assumes that there exist low-dimensional latent embeddings for each data point $\mathbf{z} = \{z_1, \dots, z_n\} \subset \mathbb{R}^d$ ($d \ll D$), which come from a prior distribution $p_0(\cdot)$, and x_i 's are drawn conditionally independently given z_i . Denote the model parameters as θ . The likelihood of this model is $p_\theta(\mathbf{x}) = \prod_{i=1}^n \int p_0(z_i) p_\theta(x_i | z_i) dz_i$.

To simultaneously learn the model parameters θ as well as a mapping from the observed data \mathbf{x} to the latent embeddings \mathbf{z} , Variational Auto-Encoders (VAEs) [13, 23] apply a data-dependent variational approximation $q_\lambda(\mathbf{z} | \mathbf{x}) = \prod_{i=1}^n q_\lambda(z_i | x_i)$, where λ denotes the variational parameters, and maximize the *evidence lower-bound* (ELBO) on the log-likelihood of the data, $\log p_\theta(\mathbf{x})$:

$$L(\lambda, \theta) = \mathbb{E}_{q_\lambda(\mathbf{z} | \mathbf{x})} [\log p_\theta(\mathbf{x} | \mathbf{z})] - \text{KL}(q_\lambda(\mathbf{z} | \mathbf{x}) || p_0(\mathbf{z})). \quad (1)$$

2.2 Correlated Variational Auto-Encoders

Standard VAEs are capable of learning compact low-dimensional embeddings for high-dimensional data. However, due to the i.i.d. assumption, they fail to account for the correlations between data points when *a priori* we know such correlations exist. Correlated Variational Auto-Encoders (CVAEs) [27] mitigate the issue by employing a structured prior as well as a structured variational posterior.

Formally, assume we are given an undirected correlation graph $G = (V, E)$, where $(v_i, v_j) \in E$ represents that the data points x_i and x_j are correlated. CVAEs apply a correlated prior $p_0^{\text{corr}}(\mathbf{z})$ on the latent variables \mathbf{z}_i 's which satisfies

$$\begin{cases} p_0^{\text{corr}}(z_i) = p_0(z_i) \text{ for all } v_i \in V \\ p_0^{\text{corr}}(z_i, z_j) = p_0(z_i, z_j) \text{ if } (v_i, v_j) \in E. \end{cases} \quad (2)$$

Here $p_0(\cdot)$ and $p_0(\cdot, \cdot)$ are parameter-free functions that capture the singleton and pairwise marginal distributions of the latent variables. For example, we can set $p_0(\cdot)$ to be the density of a standard multivariate normal distribution and $p_0(\cdot, \cdot)$ to be a multivariate normal density that has high values if

the two inputs are close to each other. With such a prior, we again assume \mathbf{x}_i 's are drawn conditionally independently given \mathbf{z}_i . When G is acyclic, such a prior $p_0^{\text{corr}}(\mathbf{z})$ does exist [30]:

$$p_0^{\text{corr}}(\mathbf{z}) = \prod_{i=1}^n p_0(\mathbf{z}_i) \prod_{(v_i, v_j) \in E} \frac{p_0(\mathbf{z}_i, \mathbf{z}_j)}{p_0(\mathbf{z}_i)p_0(\mathbf{z}_j)}. \quad (3)$$

However, when G is not acyclic, Eq. 3 is not necessarily a valid probability density function. To deal with this issue, CVAEs propose constructing a prior that is a mixture over the set \mathcal{A}_G of all of G 's maximal acyclic subgraphs, which are defined as follows.

Definition 1 (Maximal acyclic subgraph). *For an undirected graph $G = (V, E)$, an acyclic subgraph $G' = (V', E')$ is a maximal acyclic subgraph of G if:*

- $V' = V$, i.e., G' contains all vertices of G .
- Adding any edge from E/E' to E' will create a cycle in G' .

Each of the maximal acyclic subgraphs $G' \in \mathcal{A}_G$ partially approximates the correlation structure in G , and CVAEs set the prior $p_0^{\text{corr}_g}(\mathbf{z})$ to be the uniform average over all of these tractable densities.

$$p_0^{\text{corr}_g}(\mathbf{z}) \triangleq \frac{1}{|\mathcal{A}_G|} \sum_{G'=(V, E') \in \mathcal{A}_G} p_0^{G'}(\mathbf{z}), \quad (4)$$

where $p_0^{G'}(\mathbf{z}) = \prod_{i=1}^n p_0(\mathbf{z}_i) \prod_{(v_i, v_j) \in E'} \frac{p_0(\mathbf{z}_i, \mathbf{z}_j)}{p_0(\mathbf{z}_i)p_0(\mathbf{z}_j)}$ is a prior on a maximal acyclic subgraph $G' = (V, E')$ with the same form as in Eq. 3. For each $G' \in \mathcal{A}_G$, we can similarly define a structured variational approximation $q_{\lambda}^{G'}(\mathbf{z}|\mathbf{x})$ following the form of Eq. 3 (see Appendix A for details). With this structured prior and variational posterior, CVAEs optimize a different ELBO:

$$\begin{aligned} \log p_{\theta}(\mathbf{x}) &= \log \mathbb{E}_{p_0^{\text{corr}_g}(\mathbf{z})} [p_{\theta}(\mathbf{x}|\mathbf{z})] \\ &\geq \frac{1}{|\mathcal{A}_G|} \sum_{G' \in \mathcal{A}_G} \left(\mathbb{E}_{q_{\lambda}^{G'}(\mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] - \text{KL}(q_{\lambda}^{G'}(\mathbf{z}|\mathbf{x}) || p_0^{G'}(\mathbf{z})) \right). \end{aligned} \quad (5)$$

Even though empirically Tang et al. [27] show that CVAEs are capable of capitalizing on the correlation structure as auxiliary information when learning latent embeddings. As discussed in the introduction, there are a few limitations to this approach. In the next section, we will propose fixes to all of these limitations.

3 ADAPTIVE CORRELATED VAES

3.1 A Non-uniform Mixture Prior

As motivated in Section 2.2, rather than using a uniform average, we instead employ a categorical distribution $\pi \in \Delta^{|\mathcal{A}_G|-1}$ representing the normalized weights over all maximal acyclic subgraphs $G' \in \mathcal{A}_G$ of G . In the ELBO in Eq. 5, we can replace the uniform average in the prior $p_0^{\text{corr}_g}(\mathbf{z})$ in Eq. 4 with the non-uniform distribution π , which gives us the following ELBO:

$$\begin{aligned} &\mathbb{E}_{G' \sim \pi} \left[\mathbb{E}_{q_{\lambda}^{G'}(\mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] - \text{KL}(q_{\lambda}^{G'}(\mathbf{z}|\mathbf{x}) || p_0^{G'}(\mathbf{z})) \right] \\ &\leq \mathbb{E}_{G' \sim \pi} \left[\mathbb{E}_{p_0^{G'}(\mathbf{z})} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] \right] := \mathbb{E}_{p_{\pi}^{\theta}(\mathbf{z})} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] \\ &\leq \log p_{\pi, \theta}(\mathbf{x}). \end{aligned} \quad (6)$$

Here we define the non-uniform prior $p_{\pi}^{\theta}(\mathbf{z}) = \mathbb{E}_{G' \sim \pi} [p_0^{G'}(\mathbf{z})]$. From the above inequality we can see that, using the non-uniform prior p_{π}^{θ} , we are still able to obtain a lower bound of the log-likelihood $\log p_{\pi, \theta}(\mathbf{x})$, which is now also parametrized by the weight parameter π . If we optimize π together with all the other parameters, the above loss function implies that we are optimizing with an adaptive prior. Hence, we call the above model Adaptive Correlated Variational Auto-Encoders (ACVAEs). If we replace π with a uniform distribution over all subgraphs in \mathcal{A}_G , we recover CVAEs.

Plugging $q_{\lambda}^{G'}(z|\mathbf{x})$ and $p_0^{G'}(z)$ from Section 2.2 into Eq. 6, yields the following ELBO for ACVAEs:

$$\begin{aligned} \mathcal{L}^{\text{ACVAE}}(\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\theta}) := & \sum_{i=1}^n \left(\mathbb{E}_{q_{\lambda}(\mathbf{z}_i|\mathbf{x}_i)} [\log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_i)] - \text{KL}(q_{\lambda}(\mathbf{z}_i|\mathbf{x}_i)||p_0(\mathbf{z}_i)) \right) - \sum_{(v_i, v_j) \in E} w_{G, \boldsymbol{\pi}, (v_i, v_j)}^{\text{MAS}} \\ & \left(\text{KL}(q_{\lambda}(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j)||p_0(\mathbf{z}_i, \mathbf{z}_j)) - \text{KL}(q_{\lambda}(\mathbf{z}_i|\mathbf{x}_i)||p_0(\mathbf{z}_i)) - \text{KL}(q_{\lambda}(\mathbf{z}_j|\mathbf{x}_j)||p_0(\mathbf{z}_j)) \right). \end{aligned} \quad (7)$$

Similar to CVAEs, we have edge weights $w_{G, \boldsymbol{\pi}, (v_i, v_j)}^{\text{MAS}}$ representing the expected appearance probability for edge (v_i, v_j) over the set of maximal acyclic subgraphs \mathcal{A}_G given the distribution $\boldsymbol{\pi}$. In the following definition, we abusively write $\boldsymbol{\pi}(G')$ as the probability of G' being sampled from \mathcal{A}_G .

Definition 2 (Non-uniform maximal acyclic subgraph edge weight). *For an undirected graph $G = (V, E)$, an edge $e \in E$ and a distribution $\boldsymbol{\pi}$ on the set \mathcal{A}_G of maximal acyclic subgraphs of G , define $w_{G, \boldsymbol{\pi}, e}^{\text{MAS}}$ to be the expected appearance probability of the edge e in a random maximal acyclic subgraph $G' = (V, E') \sim \boldsymbol{\pi}$, i.e., $w_{G, \boldsymbol{\pi}, e}^{\text{MAS}} := \sum_{G' \in \mathcal{A}_G, e \in E'} \boldsymbol{\pi}(G')$.*

Similar to CVAEs, we can apply negative sampling (equivalent to applying a complete graph as a weak prior) to ACVAEs as regularization, which helps prevent overfitting on the learned pairwise variational approximation ($\gamma > 0$ is the regularization strength):

$$\begin{aligned} \mathcal{L}^{\text{ACVAE-NS}}(\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\theta}) := & \mathcal{L}^{\text{ACVAE}}(\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\theta}) - \gamma \cdot \left(\sum_{i=1}^n \text{KL}(q_{\lambda}(\mathbf{z}_i|\mathbf{x}_i)||p_0(\mathbf{z}_i)) \right. \\ & \left. + \frac{2}{n} \sum_{1 \leq i < j \leq n} \mathbb{E}_{q_{\lambda}(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j)} \log \frac{q_{\lambda}(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j)}{q_{\lambda}(\mathbf{z}_i|\mathbf{x}_i)q_{\lambda}(\mathbf{z}_j|\mathbf{x}_j)} \right). \end{aligned} \quad (8)$$

In what follows, we use $\mathcal{L}^{\text{ACVAE}}$ to refer to $\mathcal{L}^{\text{ACVAE-NS}}$ for notational brevity.

3.2 Learning the Non-uniform Mixture

With the loss function in Eq. 8, an intuitive direction for estimating $\boldsymbol{\pi}$ would be to perform empirical Bayes [7] and directly maximize $\mathcal{L}^{\text{ACVAE}}(\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\pi}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, as in Eq. 9:

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\theta}} \max_{\boldsymbol{\pi}} \mathcal{L}^{\text{ACVAE}}(\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\theta}). \quad (9)$$

Alternatively, we can consider a minimax saddle-point optimization, which may lead to more robust inference:

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\theta}} \min_{\boldsymbol{\pi}} \mathcal{L}^{\text{ACVAE}}(\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\theta}). \quad (10)$$

As Eq. 10 indicates, we are optimizing the ELBO under the prior that produces the *lowest* lower bound. The intuition is that if we can even optimize the worst lower bound well, the variational distribution and the model distribution we learn would be robust and generalize better. This is similar to the *least favorable prior*, under which a Bayes estimator can achieve minimax risk [15].

Empirical Bayes (Eq. 9) aims to find the best variational approximation, while the saddle-point option (Eq. 10) aims for robust inference. At first glance, the empirical Bayes option seems more reasonable since it gives us the tightest lower bound. However, a better ELBO does not necessarily translate into better predictive performance in the downstream task. In Section 4, we compare these two optimization options on various datasets, and discuss the pros and cons of each.

An important observation is that, no matter which option is applied, for fixed $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, the loss function $\mathcal{L}^{\text{ACVAE}}$ is linear w.r.t. the weight parameter $\boldsymbol{\pi}$. Therefore, if optima for $\mathcal{L}^{\text{ACVAE}}(\boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\theta})$ exist, then at least one optimum will have a $\boldsymbol{\pi}^*$ which puts all of its probability mass on a single subgraph G'^* .

Proposition 1 (Optimum for $\boldsymbol{\pi}$). *If the optimization in Eq. 9 or Eq. 10 has global optima, then at least one optimum $(\boldsymbol{\pi}^*, \boldsymbol{\lambda}^*, \boldsymbol{\theta}^*)$ will have a $\boldsymbol{\pi}^*$ that places all of its probability mass on a single maximal acyclic subgraph $G'^* \in \mathcal{A}_G$.*

From this proposition, we know both Eq. 9 and Eq. 10 return a single subgraph G'^* , which drastically simplifies the structured prior. At this optimum, the loss function becomes the ELBO on a single acyclic subgraph G'^* , with $q_{\lambda}^{G'^*}(z|\mathbf{x})$ as the variational distribution. Therefore, we have a holistic variational approximation, overcoming a limitations of CVAEs.

3.3 Learning with Alternating Updates

Direct optimization of either Eq. 9 or Eq. 10 is non-trivial. Following similar saddle-point optimization for a spanning tree structured upper bound for the log-partition function of undirected graphical models [31, 32], we perform an alternating optimization procedure on the parameters λ , θ and π . Details are shown in Algorithm 1.

Updates For π When the parameters λ and θ are fixed, the loss function $\mathcal{L}^{\text{ACVAE}}(\pi, \lambda, \theta)$ is linear in π . However, we cannot directly optimize over $\pi \in \Delta^{|\mathcal{A}_G|-1}$, as it may contain exponentially many dimensions. We can instead update the edge weights $w_{G,\pi,(v_i,v_j)}^{\text{MAS}}$ as the loss function is also linear in them.

By definition, we know that each maximal acyclic subgraph G' of G is a forest, consisting of one spanning tree for each connected component of G . Therefore, the domain for the edge weights $\bigcup_{e \in E} \{w_{G,\pi,e}^{\text{MAS}}\}$ is the projection of the Cartesian product of the spanning tree polytopes for all connected components of G [31, 32] to the edge weight space. This Cartesian product on the polytopes is convex and its boundary is determined by potentially exponentially many linear inequalities. Despite that, directly maximizing (or minimizing) $\mathcal{L}^{\text{ACVAE}}(\pi, \lambda, \theta)$ with respect to these weights $\bigcup_{e \in E} \{w_{G,\pi,e}^{\text{MAS}}\}$ is in fact tractable: the optimum for Eq. 9 or Eq. 10 is obtained at $\hat{\pi}$ that has all the mass on a single maximal acyclic subgraph \hat{G}' . This means the optimum for these edges weights can be obtained from a single subgraph \hat{G}' . By re-arranging terms in Eq. 8 with respect to $\bigcup_{e \in E} \{w_{G,\pi,e}^{\text{MAS}}\}$, it is not difficult to see that \hat{G}' should have the smallest (for empirical Bayes) or largest (for saddle-point) “edge mass” sum over all maximal acyclic subgraphs \mathcal{A}_G , where the “edge mass” $m_{(v_i,v_j)}$ of edge $e = (v_i, v_j)$ is:

$$m_{(v_i,v_j)} := \text{KL}(q_{\lambda}(z_i, z_j | \mathbf{x}_i, \mathbf{x}_j) || p_0(z_i, z_j)) - \text{KL}(q_{\lambda}(z_i | \mathbf{x}_i) || p_0(z_i)) - \text{KL}(q_{\lambda}(z_j | \mathbf{x}_j) || p_0(z_j)), \quad (11)$$

which means \hat{G}' is the combination of the minimum (for empirical Bayes) or maximum (for saddle-point) spanning trees of all connected components of the graph with $m_{(v_i,v_j)}$ as the weights.

Once we identify \hat{G}' , the optimal weights $\hat{w}_{G,\hat{\pi},e}^{\text{MAS}}$ are either 1 (if the edge e is selected) or 0 (otherwise). Instead of directly updating the weights to the optimal values, we perform a soft update with step size α^t at iteration t , similar to Wainwright [31], Wainwright et al. [32]:

$$w_{G,\pi,e}^{\text{MAS}^{t+1}} \leftarrow (1 - \alpha^t) w_{G,\pi,e}^{\text{MAS}^t} + \alpha^t \hat{w}_{G,\hat{\pi},e}^{\text{MAS}}. \quad (12)$$

This soft update helps prevent the algorithm from becoming trapped in bad local optima early in the optimization procedure. The step size α^t can be either a constant or dynamically adjusted during optimization. We set it to be a constant in our experiments.

One of the limitations of CVAEs mentioned in Section 2.2 is the $O(|V|^3)$ pre-processing step to compute all the edge weights $w_{G,e}^{\text{MAS}}$. We alleviate this bottleneck in ACVAEs, as it only takes $O(\min(|V|^2, |E| \log |V|))$ operations per initialization (details in Appendix B.2) and per update on the weights, which ensures that ACVAEs can scale to datasets with many more vertices than would be feasible with CVAEs.

Updates For λ And θ When π is fixed, λ and θ can be updated by taking a stochastic gradient step following $\nabla_{\lambda,\theta} \mathcal{L}^{\text{ACVAE}}(\pi, \lambda, \theta)$ with reparametrization gradient [13, 23], as done in standard VAEs.

If empirical Bayes (Eq. 9) is applied, Algorithm 1 will converge with properly selected learning rates. On the other hand, it is difficult to make any general statement about the convergence for saddle-point optimization (Eq. 10) since the objective is generally non-concave in (λ, θ) . However, as we show in Section 4, empirically we find that Algorithm 1 is stable for both options and performs well on multiple real datasets.

Algorithm 1 ACVAES learning

Input: data $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^D$, undirected graph $G = (V = \{v_1, \dots, v_n\}, E)$, parameter $\gamma > 0$. Initialize the parameters λ, θ . Initialize the weights $w_{G,\pi,e}^{\text{MAS}}$ for each $e \in E$.

while not converged **do**

Optimize the parameters (λ, θ) using the gradients $\nabla_{\lambda, \theta} \mathcal{L}^{\text{ACVAE}}(\pi, \lambda, \theta)$.

Compute the mass $m_{(v_i, v_j)}$ for each edge $e \in E$ with Eq. 11.

Compute a minimum (if applying the empirical Bayes option in Eq. 9) or maximum (if applying the saddle-point option in Eq. 10) spanning tree of the graph with the masses $m_{(v_i, v_j)}$'s as the edge weights and update the weights $w_{G,\pi,e}^{\text{MAS}}$ for each $e \in E$ according to Eq. 12.

end while

Return: The parameters λ, θ , the weights $w_{G,\pi,e}^{\text{MAS}}$ for each $e \in E$.

3.4 Exact Marginal Posterior Approximation with Belief Propagation

From Proposition 1, we know the weights $w_{G,\pi,e}^{\text{MAS}}$ returned from Algorithm 1 are from a single maximal acyclic subgraph $G' \in \mathcal{A}_G$. Consequently, we have a holistic variational approximation $q_\lambda^{G'}(z|\mathbf{x})$. However, by itself this variational approximation might not be necessarily better at the downstream predictive tasks than CVAES since it can only make use of the structure from *one* maximal acyclic subgraph G' .

On the plus side, the acyclic structure of G' makes it possible to compute the exact pairwise marginal variational distribution between any pair of vertices via a belief-propagation-style [21] message-passing algorithm, which is not possible for CVAES, as it does not have a single joint variational distribution on z . This can be crucial in tasks in which we need an accurate pairwise marginal approximation, e.g., link prediction and hierarchical clustering.

Consider any $v_i \neq v_j \in V$ that are in the same connected component of G' . Since G' is acyclic there is a unique path from v_i to v_j . Denote it as $v_i = u_0^{i,j} \rightarrow u_1^{i,j} \rightarrow \dots \rightarrow u_{k_{i,j}}^{i,j} = v_j$. The exact pairwise marginal $r_\lambda(z_i, z_j | \mathbf{x}_i, \mathbf{x}_j)$ equals

$$\int \prod_{l=0}^{k_{i,j}-1} q_\lambda(z_{u_l^{i,j}}, z_{u_{l+1}^{i,j}} | \mathbf{x}_{u_l^{i,j}}, \mathbf{x}_{u_{l+1}^{i,j}}) \prod_{l=1}^{k_{i,j}-1} \frac{dz_{u_l^{i,j}}}{q_\lambda(z_{u_l^{i,j}} | \mathbf{x}_{u_l^{i,j}})}.$$

The above pairwise marginal densities can be computed for all pairs of (v_i, v_j) by doing a depth- or breadth-first search starting from each $v_i \in V$ after we obtain the variational approximation $q_\lambda^{G'}(z|\mathbf{x})$ from Algorithm 1, which has a total complexity of $O(|V|^2)$. Note that the time complexity for evaluating every pairwise marginal in CVAES is also $O(|V|^2)$. But the belief propagation refinement computation is usually more efficient in practice, since it involves much less neural network function evaluations, which dominate the runtime.

4 EXPERIMENTS

In this section, we evaluate ACVAES on the task of link prediction and hierarchical clustering. We show that our method significantly outperforms various baselines. We attempt to identify the contributing factors for the gain, answering the following questions:

Q1: Uniform mixture (CVAE) *versus* non-uniform mixture (ACVAE), which one is better? (Section 4.2.1)

Q2: How important is the belief propagation refinement for ACVAE? (Section 4.2.1)

Q3: Empirical Bayes *versus* saddle-point, which one performs better? Can we select purely based on ELBO? (Section 4.2.2)

Q4: Does the learned single graph capture more information than singleton representations? What do the learned latent embeddings look like? (Section 4.2.3)

Q5: Can ACVAE scale to datasets that CVAE cannot? (Section 4.2.4)

4.1 Experiment Settings

Before presenting our experimental results, we describe the tasks, datasets, baselines, and metrics for evaluation. Additional details can be found in Appendix B.

4.1.1 Tasks

For each of the tasks, we are given a correlation graph $G = (V = \{v_1, \dots, v_n\}, E)$ and a feature vector $\mathbf{x}_i \in \mathbb{R}^N$ for each $i \in \{1, \dots, n\}$.

For the link prediction task, we keep consistent with the setting of Tang et al. [27]. For the hierarchical clustering experiments, we apply the **complete-linkage** algorithm [33], which is relatively more stable among common hierarchical clustering algorithms. We cluster all data points into $K = 5$ clusters.

4.1.2 Datasets

We evaluate ACVAEs on the following 3 datasets. All of 3 datasets are tested for link prediction and in addition the LibraryThing dataset is tested for the hierarchical clustering experiment:

- Epinions¹ [19], a public product rating dataset that contains $\approx 49K$ users and $\approx 140K$ products. After pre-processing, the dataset contains $\approx 16K$ users.
- Citation² [16], a High-energy physics theory citation network dataset, which has a citation graph with $\approx 28K$ papers and $\approx 353K$ citation edges. After preprocessing, the dataset contains $\approx 2K$ users (for the results as in Section 4.2.1). We also perform an experiment in Section 4.2.4 on a larger version of this dataset, which contains $\approx 26K$ users.
- LibraryThing³ [14], a public book review data set that contains $\approx 73K$ users and $\approx 337K$ items. After pre-processing, the dataset contains $\approx 6K$ users.

For the hierarchical clustering task, the LibraryThing dataset does not contain cluster labels for users. We generate the cluster labels for each user by learning a standard VAE on the feature vectors \mathbf{x} , and perform the complete-linkage algorithm to cluster the data points into $K = 5$ clusters. This helps us generate a *semi-synthetic* dataset.

4.1.3 Baselines

We compare ACVAE with 4 baseline methods:

- VAE [13]: standard variational auto-encoders, with no information about the correlations.
- GraphSAGE [8]: the state-of-the-art method for learning latent embeddings that takes the correlation structure into account with graph convolutional neural networks.
- CVAE_{ind} and CVAE_{corr} [27]: Two variations of CVAEs with factorized and structured variational approximations, respectively.

4.1.4 Metrics

For all methods, we first learn latent embeddings $\mathbf{z}_1, \dots, \mathbf{z}_n$, which are deterministic for GraphSAGE and stochastic for the VAES-based methods. Then we compute the pairwise distance $\text{dis}_{i,j}$ between each pair $(\mathbf{z}_i, \mathbf{z}_j)$ of the latent embeddings as $\|\mathbf{z}_i - \mathbf{z}_j\|_2^2$. Recall that the embeddings are stochastic for the VAES-based methods, hence we use $\mathbb{E}[\|\mathbf{z}_i - \mathbf{z}_j\|_2^2]$ as the pairwise distance. The expectation is taken over the variational pairwise marginal $q(\mathbf{z}_i, \mathbf{z}_j | \mathbf{x}_i, \mathbf{x}_j)$ or the refined pairwise marginal $r(\mathbf{z}_i, \mathbf{z}_j | \mathbf{x}_i, \mathbf{x}_j)$ if we perform belief propagation (Section 3.4).

For the link prediction experiments, for each user u_i , we compute the *Cumulative Reciprocal Rank* (CRR) as follows.

$$\text{CRR}_i = \sum_{(v_i, v_j) \in E_{\text{test}}} \frac{1}{|\{k : (v_i, v_k) \notin E_{\text{train}}, \text{dis}_{i,k} \leq \text{dis}_{i,j}\}|}$$

¹http://www.trustlet.org/downloaded_epinions.html

²<http://snap.stanford.edu/data/cit-HepTh.html>

³https://cseweb.ucsd.edu/~jmcauley/datasets.html#social_data

Table 1: Link Prediction Normalized CRR

| Method | Epinions | Citation | LibraryThing |
|------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| VAE | 0.005 ± 0.001 | 0.064 ± 0.005 | 0.006 ± 0.000 |
| GraphSAGE | 0.012 ± 0.003 | 0.018 ± 0.003 | 0.004 ± 0.001 |
| CVAE _{ind} | 0.016 ± 0.000 | 0.046 ± 0.003 | 0.012 ± 0.001 |
| CVAE _{corr} | 0.017 ± 0.001 | 0.094 ± 0.002 | 0.020 ± 0.001 |
| ACVAE _{EB} | 0.013 ± 0.000 | 0.070 ± 0.004 | 0.018 ± 0.001 |
| ACVAE _{SP} | 0.010 ± 0.003 | 0.083 ± 0.001 | 0.018 ± 0.001 |
| ACVAE _{EB+BP} | 0.034 ± 0.003 | 0.122 ± 0.007 | 0.032 ± 0.002 |
| ACVAE _{SP+BP} | 0.035 ± 0.001 | 0.116 ± 0.006 | 0.032 ± 0.001 |

A larger CRR value indicates the heldout edges have a higher rank among all the candidates. We further normalize the CRR values to be in $[0, 1]$, and report the normalized CRR (NCRR).

For hierarchical clustering, we apply the normalized mutual-information scores [29] as the metric. These scores are in the range $[0, 1]$ and a larger score indicates better clustering performance.

4.2 Results

We show the heldout NCRR values for link predictions and the normalized MI scores in Table 1 and Table 2, respectively. ACVAE_{EB} and ACVAE_{SP} stand for empirical Bayes (Eq. 9) and saddle-point optimization (Eq. 10), respectively. The rows with BP mean we perform belief-propagation refinement (Section 3.4). We dissect the results in the following sections.

4.2.1 Advantages of the Non-uniform Mixture

As motivated in Section 2.2, ACVAEs improve over the limitations of CVAEs by providing a holistic variational approximation at the end of the empirical Bayes or saddle-point optimization, which further enables applying belief propagation for more accurate marginal approximation.

At first glance, the performance results in Table 1 for the single joint distribution (the rows ACVAE_{EB} and ACVAE_{SP}) is no better than that of CVAE_{corr}, which applies a uniform mixture. We speculate in Section 3.4 that by itself this holistic variational approximation might not necessarily be better at the downstream predictive tasks since it can only make use of the structure from *one* maximal acyclic subgraph, even though it sometimes has a higher ELBO (Table 4). However, we can observe a huge performance boost after applying the belief propagation refinement, which outperforms the baseline methods by a wide margin for link prediction and performs comparably better for hierarchical clustering.⁴

Recall that the prerequisite for applying the belief propagation is to have a variational distribution on a single acyclic subgraph (i.e., we **cannot** perform BP with CVAEs). This answers two questions we sought to answer: First, the non-uniform mixture is not necessarily better than the uniform mixture at the downstream task even when it has a higher ELBO, but it opens up the possibility to perform exact inference; Second, variational approximations has a lot of room for improvement when compared with exact inference (i.e., belief propagation) on an acyclic graph.

4.2.2 Empirical Bayes *versus* Saddle-Point

As shown in Table 1 and Table 2, both empirical Bayes and saddle-point optimization perform similarly on most tasks, though the saddle-point option is often more stable (especially on the clustering task). This is reasonable since the saddle-point objective optimizes the most conservative lower bound.

Moreover, we show that we should not select between these two methods purely based on ELBO: By definition, the saddle-point optimization will yield an ELBO lower than empirical Bayes. In Table 3

⁴VAE does not count as a baseline method for the clustering experiment since it is applied as an oracle in the pre-processing steps. We omitted the results for ACVAEs without belief propagation for the hierarchical clustering experiments since empirically we found their performance are much worse compared to the case of using belief propagation refinement.

Table 2: Hierarchical Clustering Normalized MI Scores

| Method | MI Scores |
|------------------------|-------------------------------------|
| GraphSAGE | 0.002 ± 0.000 |
| CVAE _{ind} | 0.010 ± 0.004 |
| CVAE _{corr} | 0.002 ± 0.000 |
| ACVAE _{EB+BP} | 0.012 ± 0.003 |
| ACVAE _{SP+BP} | 0.011 ± 0.002 |

and Table 4 we report ELBO as well as NCRP for 4 choices of the negative sampling parameter γ (Eq. 8) on Epinions and Citations with and without belief propagation refinement. We can see clearly that a better ELBO does not necessarily correlate with a better NCRP, regardless of belief propagation refinement or not.

Table 3: ELBO | Average NCRP Comparisons between ACVAE (with BP) on Epinions and Citation

| Epinions | ACVAE _{EB+BP} | ACVAE _{SP+BP} |
|------------------|------------------------|------------------------|
| $\gamma = 0.001$ | $-31.8 0.034$ | $-31.9 0.031$ |
| $\gamma = 0.1$ | $-36.4 0.031$ | $-38.3 0.035$ |
| $\gamma = 10.$ | $-61.3 0.028$ | $-119 0.034$ |
| $\gamma = 1000.$ | $-674 0.028$ | $-1535 0.037$ |
| Citation | ACVAE _{EB+BP} | ACVAE _{SP+BP} |
| $\gamma = 0.001$ | $-44.9 0.122$ | $-45.3 0.120$ |
| $\gamma = 0.1$ | $-44.7 0.121$ | $-44.9 0.122$ |
| $\gamma = 10.$ | $-52.1 0.124$ | $-63.8 0.116$ |
| $\gamma = 1000.$ | $-156 0.119$ | $-277 0.119$ |

Table 4: ELBO | Average NCRP Comparisons between ACVAE (without BP) and CVAE On Citation

| Citation | ACVAE _{EB} | ACVAE _{SP} | CVAE |
|------------------|---------------------|---------------------|-----------------|
| $\gamma = 0.001$ | $-44.9 0.021$ | $-45.1 0.016$ | $-45.1 0.016$ |
| $\gamma = 0.1$ | $-44.1 0.063$ | $-44.7 0.053$ | $-44.6 0.059$ |
| $\gamma = 10.$ | $-51.7 0.060$ | $-63.3 0.078$ | $-61.4 0.076$ |
| $\gamma = 1000.$ | $-153 0.055$ | $-275 0.080$ | $-248 0.088$ |

In general, both methods have their advantages. On simpler datasets, e.g., Citation, on which all methods perform well, empirical Bayes is preferred since it can easily capture the best correlation structure. On the other hand, with more complex datasets/difficult tasks, saddle-point optimization tends to provide more robust inference and stable results.

4.2.3 Learned Graph Structures

In Figure 1 we visualize part of the largest connected component of the maximal acyclic subgraph $\hat{G}' = (V, \hat{E}')$ that ACVAEs learn for the variational distribution on the Citation dataset with both empirical Bayes and saddle-point optimization (colors for better clarity only). The coordinates are t-SNE embeddings for the variational approximation mean of the latent variables. The edge widths are proportional to the strength of the learned correlations. We can see some of the learned embeddings are not necessarily close to each other even when they have high correlations. This indicates that the learned \hat{G}' provides some additional information that singleton marginals cannot provide.

Moreover, we can observe that the learned embeddings for empirical Bayes and saddle-point optimization exhibit different behaviors. We leave further investigation as future work.

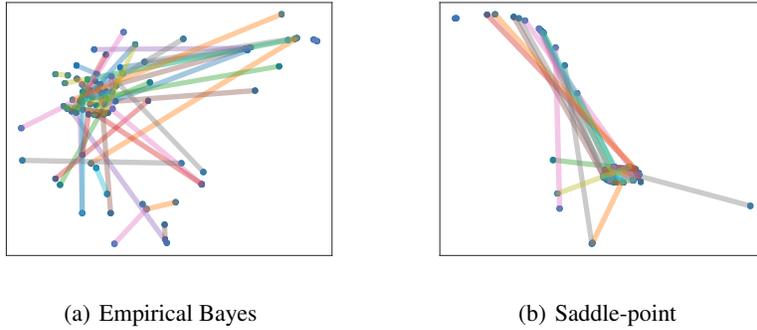


Figure 1: Embeddings of the Learned Maximal Acyclic Subgraphs \hat{G}'

4.2.4 Scalability to Large Datasets

To demonstrate the scalability of ACVAE compared to CVAE, we perform an experiment on a larger version of the Citation dataset with > 12 times more vertices, which CVAE can not easily scale to due to the cubic time initialization step and the quadratic pairwise marginal evaluations.

We compare the performance of ACVAE plus the belief propagation refinement on both the empirical Bayes and the saddle-point schemes with the other two baseline methods (VAE and GraphSAGE). As shown in Table 5, both schemes of ACVAE can significantly outperform the baseline methods.

Table 5: NCRR on The Larger Citation Dataset

| Method | NCRR |
|------------------------|-------|
| VAE | 0.011 |
| GraphSAGE | 0.002 |
| ACVAE _{EB+BP} | 0.075 |
| ACVAE _{SP+BP} | 0.068 |

5 RELATED WORK

This work extends CVAEs with the idea of learning a non-uniform average loss over some tractable loss functions on maximal acyclic subgraphs of the given graph. This is similar to the idea of obtaining a tighter upper bound on the log-partition function for an undirected graphical model by minimizing over a convex combination of spanning trees of the given graph [32]. To optimize the parameters, Wainwright et al. [32] also apply alternating updates on the parameters and the distributions over the spanning trees, similar to the approach in ACVAE learning. Alternating parameter updates are useful for many other cases. For example, Alternating Least Squares for matrix factorization [25] and Alternating Direction Method of Multipliers (ADMM) for convex optimization [4, 26, 10].

Some recent work also focuses on incorporating correlation structures over latent variables. For example, Hoffman and Blei [9] proposed structured variational families that can improve over traditional mean-field variational inference. Johnson et al. [11] proposed Structured VAEs that apply more complex forms for the priors on the latent embeddings. Recently in the NLP community, Yin et al. [34] proposed utilizing tree-structured latent variable models to deal with semantic parsing. However, most of these works focus on correlations *within* dimensions of latent variables whereas our work focus on correlations *between* latent variables, similar to the setting of CVAEs. In addition, Luo et al. [17] incorporated pairwise correlations between latent variables into deep generative models for semi-crowdsourced clustering.

Another line of related work appears in convolutional networks for graphs and their extensions [3, 6, 5, 20, 8, 28], which also take graph structure of data into considerations.

6 CONCLUSION

In this paper, we introduce ACVAEs, which learn a joint variational distribution on the latent embeddings of input data via optimizing loss function that is a non-uniform average over some tractable correlated ELBOs. To learn the mixture weights, we provide two different options, and compare them on various datasets and tasks. The learned joint variational distribution can be used to perform efficient evaluation using belief propagation. Experiment results show that ACVAEs can outperform existing methods for link prediction and hierarchical clustering on three real datasets. Future work will include better understanding the learned graph structures from both options and learning higher-order correlations between latent variables.

References

- [1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.
- [2] Stefan Behnel, Robert Bradshaw, Craig Citro, Lisandro Dalcin, Dag Sverre Seljebotn, and Kurt Smith. Cython: The best of both worlds. *Computing in Science & Engineering*, 13(2):31, 2011.
- [3] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. In *3rd International Conference on Learning Representations*, 2015.
- [4] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.
- [5] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pages 3844–3852, 2016.
- [6] David Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.
- [7] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.
- [8] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1025–1035, 2017.
- [9] Matthew Hoffman and David Blei. Stochastic structured variational inference. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 361–369, 2015.
- [10] Mingyi Hong and Zhi-Quan Luo. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1-2):165–199, 2017.
- [11] Matthew Johnson, David Duvenaud, Alex Wiltschko, Ryan Adams, and Sandeep Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.
- [12] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.
- [13] Diederik Kingma and Max Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations*, 2014.
- [14] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. What’s in a name? understanding the interplay between titles, content, and communities in social media. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

- [15] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.
- [16] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. <http://snap.stanford.edu/data>, June 2014.
- [17] Yucen Luo, Tian Tian, Jiaxin Shi, Jun Zhu, and Bo Zhang. Semi-crowdsourced clustering with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3212–3222, 2018.
- [18] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [19] Paolo Massa and Paolo Avesani. Trust-aware recommender systems. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 17–24. ACM, 2007.
- [20] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023, 2016.
- [21] Judea Pearl. Reverend bayes on inference engines: A distributed hierarchical approach. In *AAAI*, 1982.
- [22] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
- [23] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1278–1286, 2014.
- [24] Tianlin Shi, Da Tang, Liwen Xu, and Thomas Moscibroda. Correlated compressive sensing for networked data. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 722–731, 2014.
- [25] Yoshio Takane, Forrest W Young, and Jan De Leeuw. Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42(1):7–67, 1977.
- [26] Da Tang and Tong Zhang. On the duality gap convergence of admm methods. *arXiv preprint arXiv:1508.03702*, 2015.
- [27] Da Tang, Dawen Liang, Tony Jebara, and Nicholas Ruoizzi. Correlated variational auto-encoders. In *International Conference on Machine Learning*, 2019.
- [28] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *7th International Conference on Learning Representations*, 2019.
- [29] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.
- [30] Martin Wainwright and Michael Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 2008.
- [31] Martin J. Wainwright. *Stochastic processes on graphs with cycles: geometric and variational approaches*. PhD thesis, Massachusetts Institute of Technology, 2002.
- [32] Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51(7):2313–2335, 2005.
- [33] Odilia Yim and Kylee T Ramdeen. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *The quantitative methods for psychology*, 11(1):8–21, 2015.

- [34] Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. Structvae: Tree-structured latent variable models for semi-supervised semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 754–765, 2018.

Appendix

In the appendix, we provide more details on our baseline method CVAE [27] as well as the experiment data pre-processing and protocols.

A MORE DETAILS ON CVAES

CVAES set the prior $p_0^{\text{corr}g}(\mathbf{z})$ to be the uniform average over all of these tractable densities:

$$p_0^{\text{corr}g}(\mathbf{z}) = \frac{1}{|\mathcal{A}_G|} \sum_{G'=(V,E') \in \mathcal{A}_G} p_0^{G'}(\mathbf{z}), \quad (13)$$

where $p_0^{G'}(\mathbf{z}) = \prod_{i=1}^n p_0(\mathbf{z}_i) \prod_{(v_i, v_j) \in E'} \frac{p_0(\mathbf{z}_i, \mathbf{z}_j)}{p_0(\mathbf{z}_i)p_0(\mathbf{z}_j)}$ is a prior on a maximal acyclic subgraph $G' = (V, E')$ with the same form as in Eq. 3. For each $G' \in \mathcal{A}_G$, we can similarly define a structured variational approximation $q_\lambda^{G'}(\mathbf{z}|\mathbf{x})$ following the form of Eq. 3:

$$q_\lambda^{G'}(\mathbf{z}|\mathbf{x}) = \prod_{i=1}^n q_\lambda(\mathbf{z}_i|\mathbf{x}_i) \prod_{(v_i, v_j) \in E'} \frac{q_\lambda(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j)}{q_\lambda(\mathbf{z}_i|\mathbf{x}_i)q_\lambda(\mathbf{z}_j|\mathbf{x}_j)},$$

where $q_\lambda(\cdot|\cdot)$ and $q_\lambda(\cdot, \cdot|\cdot, \cdot)$ are two conditional density functions that captures the singleton and pairwise variational approximation densities. These two functions need to satisfy the symmetry and consistency properties:

$$\begin{cases} q_\lambda(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j) = q_\lambda(\mathbf{z}_j, \mathbf{z}_i|\mathbf{x}_j, \mathbf{x}_i) & \text{for all } \mathbf{z}_i, \mathbf{z}_j, \mathbf{x}_i, \mathbf{x}_j, \\ \int q_\lambda(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j) d\mathbf{z}_j = q_\lambda(\mathbf{z}_i|\mathbf{x}_i) & \text{for all } \mathbf{z}_i, \mathbf{x}_i, \mathbf{x}_j. \end{cases}$$

The ELBO in Eq. 5 is an average over potentially exponential many ELBOs. To make computations tractable, Tang et al. [27] simplifies this lower bound and represent it as

$$\begin{aligned} \mathcal{L}^{\text{CVAE}}(\lambda, \theta) := & \sum_{i=1}^n \left(\mathbb{E}_{q_\lambda(\mathbf{z}_i|\mathbf{x}_i)} [\log p_\theta(\mathbf{x}_i|\mathbf{z}_i)] - \text{KL}(q_\lambda(\mathbf{z}_i|\mathbf{x}_i)||p_0(\mathbf{z}_i)) \right) - \sum_{(v_i, v_j) \in E} w_{G, (v_i, v_j)}^{\text{MAS}} \\ & \left(\text{KL}(q_\lambda(\mathbf{z}_i, \mathbf{z}_j|\mathbf{x}_i, \mathbf{x}_j)||p_0(\mathbf{z}_i, \mathbf{z}_j)) - \text{KL}(q_\lambda(\mathbf{z}_i|\mathbf{x}_i)||p_0(\mathbf{z}_i)) - \text{KL}(q_\lambda(\mathbf{z}_j|\mathbf{x}_j)||p_0(\mathbf{z}_j)) \right). \quad (14) \end{aligned}$$

Where $w_{G, e}^{\text{MAS}} := \frac{|\{G' \in \mathcal{A}_G: e \in G'\}|}{|\mathcal{A}_G|}$ for each edge $e = (v_i, v_j)$ represents the fraction of G 's maximal acyclic subgraphs of G that contain e . These weights can be computed easily from the Moore-Penrose inverse of the Laplacian matrix of G .

B EXPERIMENT DETAILS

B.1 Dataset Pre-processing Details

Epinions We follow the same pre-processing scheme as Tang et al. [27]: binarize the rating data and create a bag-of-words binary feature vector for each user. We only retain the items that have been rated for at least 100 times. We construct the graph $G = (V, E)$ and only keep an edge (v_i, v_j) to be in E if both $v_i \rightarrow v_j$ and $v_j \rightarrow v_i$ appear in the original directed graph. At last, we only retain users that have at least one edge in E (i.e. having at least one bi-directional edge in the original dataset).

Citations This dataset includes the abstract and the citation information for high-energy physic theory papers on arXiv from 1992 to 2003. We work on all papers from 1998 in this dataset (in total $\approx 2.8\text{K}$ papers). We treat all citation edges as undirected edges and build the graph $G = (V, E)$. We only retain papers that cite or are cited by at least one of the other papers within this subset (for year 1998) of the dataset. We compute word counts (with stop words removed) for the abstract of each paper as feature vector.

For the larger experiment on this dataset, we apply the same pre-processing steps but work on the whole dataset (instead of the subset for year 1998).

LibraryThing For the link prediction experiment, We follow the same pre-processing scheme as for the Epinions dataset, except that we only retain the items that have been rated for at least 200 times (since this dataset is larger than the Epinions dataset). For the clustering experiment, we follow the same scheme to get a graph $G = (V, E)$, but we do not split the edges to training/testing (since clustering is unsupervised), and apply a normal VAE to generate the labels (as mentioned in the main paper). This normal VAE has the same hidden layer size with the one used in testing, but has a smaller latent representation (we use 10) to avoid generating non-reasonable labels due to overfitting.

B.2 Experimental Protocol

We run 3 runs for each methods for the Epinions experiments, and 5 runs for the other experiments. This is since the Epinions experiments work more stable empirically.

For VAE, CVAE and ACVAE, we apply a two-layer feed-forward neural inference network for the singleton variational distribution $q_{\lambda}(z_i|\mathbf{x}_i)$'s and a two-layer feed-forward neural generative network for the model distribution $p_{\theta}(\mathbf{x}|\mathbf{z})$'s. $q_{\lambda}(z_i|\mathbf{x}_i)$ is a diagonal normal distribution with the mean and standard deviation outputted from the inference network and $p_{\theta}(\mathbf{x}|\mathbf{z})$ is a multinomial distribution with the logits outputted from the generative networks. The latent dimensionality d is 100 for the Epinions experiments and the LibraryThing clustering, and 10 for the other two link prediction experiments. The hidden layer dimensionality h_1 is 300 for the Epinions experiments and 30 for the other experiments.

For GraphSAGE, we choose to use $K = 2$ aggregations, the mean aggregator, and $Q = 20$ negative samples to optimize the loss function. The hidden layer size and latent dimensionality we apply to GraphSAGE are the same with that of the standard VAE.

For CVAE and ACVAE, we set the pairwise marginal prior density function to be $p_0(\cdot) = \mathcal{N}\left(\boldsymbol{\mu} = \mathbf{0}_{2d}, \Sigma = \begin{pmatrix} I_d & \tau \cdot I_d \\ \tau \cdot I_d & I_d \end{pmatrix}\right)$ with $\tau = 0.99$. For CVAE_{corr} and ACVAE, we model the pairwise variational approximations $q(z_i, z_j|\mathbf{x}_i, \mathbf{x}_j)$ to be a multi-variate normal distribution that can be factorized across the d dimensions as the product of d independent bi-variate normal distributions. The correlation coefficients of these bi-variate normal distributions are computed from two-layer feed-forward neural networks that taking \mathbf{x}_i and \mathbf{x}_j as inputs. These two-layer neural networks have latent dimensionality h_2 to be 1000 for the Epinions experiments and 100 for the other experiments. For CVAE and ACVAE on link prediction experiments, we select the negative sampling parameter γ from set of choices, and report the performances with the best average train NCRR metrics. This parameter is selected from $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ for the LibraryThing dataset and the Citation dataset, and $\{0.001, 0.1, 10., 1000.\}$ for the Epinions dataset. For the clustering experiments, we select $\gamma = 1$ for CVAE and ACVAE since empirically we found this choice gives us a reasonably good performance.

For link prediction, for all methods, we look into the performances for every fixed number of iterations (the specific numbers depend on models) and update the current best test NCRR values if both the train ELBO and the train NCRR reach better values. We report the final current best test NCRR values as the results. For clustering, we update the current best normalized MI scores if the train ELBO reaches better values and report the final current best normalized MI scores.

For ACVAE, we set the step size parameter (in Eq. 12) $\alpha^t = 0.1$ to be a constant. We train the parameters using alternating updates as in Algorithm 1. We switch between updates on the parameters $\boldsymbol{\lambda}, \boldsymbol{\theta}$ for an epoch of the edges in E , and a single update on the weights $w_{G,\pi,e}^{\text{MAS}}$ according to Eq. 12. For the random initialization on the tree weights $w_{G,\pi,e}^{\text{MAS}}$, we just assign random weights to the graph $G = (V, E)$. Then we use Kruskal's algorithm to compute the maximal acyclic subgraph $\tilde{G} = (V, \tilde{E})$ according to these random weights, and set $w_{G,\pi,e}^{\text{MAS}} = I[e \in \tilde{E}]$. It is straightforward to see that this is a valid initialization for the weights $w_{G,\pi,e}^{\text{MAS}}$'s since these weights relate to the distribution $\tilde{\pi}$ that has all of its mass on the single subgraph \tilde{G} .

For ACVAE, after running the algorithm for some iterations, we use Kruskal's algorithm to compute the maximal acyclic subgraph $\hat{G} = (V, \hat{E})$ on the converged edge weights $\hat{w}_{G,\pi,e}^{\text{MAS}}$ to find the learned single graph \hat{G}' . This heuristic helps us solve the issues of finding the converged maximal acyclic

subgraph if we want to perform an early stopping (recall that we evaluate our metrics for every fixed number of iterations) or if there is a numerical issue.

For all methods, we apply stochastic gradient optimizations and use Adam [12] to adjust the learning rates. We set the step size to be 10^{-3} . For all methods, we use a batch size $B_1 = 64$ for sampling the vertices. For CVAE and ACVAE, we use a batch size $B_2 = 256$ for sampling the edges and non-edges.

All experiments are done using Python. The training and evaluations are done with TensorFlow [1] and Numpy. The word count values and the t-SNE embeddings [18] in the visualization (Figure 1) are computed using Scikit-learn [22]. For faster computations, we call C++ functions to do belief propagation and the Kruskal's algorithm using Cython [2].