Appendix for 'Variational Flow Graphical Model'

A. Derivation of the ELBO for both Tree and DAG structures

A.1. ELBO of Tree Models

Let each data sample has k sections, i.e., $\mathbf{x} = [\mathbf{x}^{(1)}, ..., \mathbf{x}^{(k)}]$. VFGs are graphical models that can integrate different sections or components of the dataset. We assume that for each pair of connected nodes, the edges are invertible flow functions. The vector of parameters for all the edges is denoted by θ. The forward message passing starts from $\mathbf{x}$ and ends at $\mathbf{h}^L$, and backward message passing in the reverse direction. We start with the hierarchical generative tree network structure illustrated by an example in Figure 8. Then the marginal likelihood term of the data reads

$$p(\mathbf{x}|\theta) = \sum_{\mathbf{h}^1,...,\mathbf{h}^L} p(\mathbf{h}^L|\theta)p(\mathbf{h}^{L-1}|\mathbf{h}^L,\theta)\cdots p(\mathbf{x}|\mathbf{h}^1,\theta)\,.$$

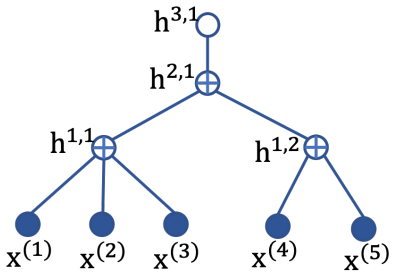


Figure 8. A VFG tree with $L = 3$.

The hierarchical prior distribution is given by factorization

$$p(\mathbf{h}) = p(\mathbf{h}^L)\mathbf{\Pi}_{l=1}^{L-1} p(\mathbf{h}^l|\mathbf{h}^{l+1}). \tag{19}$$

The probability density function $p(\mathbf{h}^{l-1}|\mathbf{h}^l)$ in the prior is modeled with one or multiple invertible normalizing flow functions. The hierarchical posterior (recognition network) is factorized as

$$q_\theta(\mathbf{h}|\mathbf{x}) = q(\mathbf{h}^1|\mathbf{x})q(\mathbf{h}^2|\mathbf{h}^1)\cdots q(\mathbf{h}^L|\mathbf{h}^{L-1}). \tag{20}$$

Draw samples from the prior (19) involves sequential conditional sampling from the top of the tree to the bottom, and computation of the posterior (20) takes the reverse direction. Notice that

$$q(\mathbf{h}|\mathbf{x}) = q(\mathbf{h}^1|\mathbf{x})q(\mathbf{h}^{2:L}|\mathbf{h}^1)\,.$$

With the hierarchical structure of a tree, we further have

$$q(\mathbf{h}^{l:L}|\mathbf{h}^{l-1}) = q(\mathbf{h}^l|\mathbf{h}^{l-1})q(\mathbf{h}^{l+1:L}|\mathbf{h}^l\mathbf{h}^{l-1}) = q(\mathbf{h}^l|\mathbf{h}^{l-1})q(\mathbf{h}^{l+1:L}|\mathbf{h}^l) \tag{21}$$

$$p(\mathbf{h}^{l:L}) = p(\mathbf{h}^l|\mathbf{h}^{l+1:L})p(\mathbf{h}^{l+1:L}) = p(\mathbf{h}^l|\mathbf{h}^{l+1})p(\mathbf{h}^{l+1:L}) \tag{22}$$

By leveraging the conditional independence in the chain structures of both posterior and prior, the derivation of trees' ELBO

becomes easier.

$$\begin{aligned}\log p(\mathbf{x}) &= \log \int p(\mathbf{x}|\mathbf{h})p(\mathbf{h})d\mathbf{h} \\ &= \log \int \frac{q(\mathbf{h}|\mathbf{x})}{q(\mathbf{h}|\mathbf{x})}p(\mathbf{x}|\mathbf{h})p(\mathbf{h})d\mathbf{h} \\ &\geqslant \mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\big[\log p(\mathbf{x}|\mathbf{h}) - \log q(\mathbf{h}|\mathbf{x}) + \log p(\mathbf{h})\big] = \mathcal{L}(x;\theta).\end{aligned}$$

The last step is due to the Jensen inequality. With $\mathbf{h} = \mathbf{h}^{1:L}$,

$$\begin{aligned}&\log p(\mathbf{x}) \geqslant \mathcal{L}(x;\theta) \\ =&\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log p(\mathbf{x}|\mathbf{h}^{1:L}) - \log q(\mathbf{h}^{1:L}|\mathbf{x}) + \log p(\mathbf{h}^{1:L})\big] \\ =&\underbrace{\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log p(\mathbf{x}|\mathbf{h}^{1:L})\big]}_{\text{(a) Reconstruction of the data}} - \underbrace{\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log q(\mathbf{h}^{1:L}|\mathbf{x}) - \log p(\mathbf{h}^{1:L})\big]}_{\mathbf{KL}^{1:L}}\end{aligned} \tag{23}$$

With conditional independence in the hierarchical structure, we have

$$q(\mathbf{h}^{1:L}|\mathbf{x}) = q(\mathbf{h}^{2:L}|\mathbf{h}^1\mathbf{x})q(\mathbf{h}^1|\mathbf{x}) = q(\mathbf{h}^{2:L}|\mathbf{h}^1)q(\mathbf{h}^1|\mathbf{x}).$$

The second term of (23) can be further expanded as

$$\mathbf{KL}^{1:L} = \mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log q(\mathbf{h}^1|\mathbf{x}) + \log q(\mathbf{h}^{2:L}|\mathbf{h}^1) - \log p(\mathbf{h}^1|\mathbf{h}^{2:L}) - \log p(\mathbf{h}^{2:L})\big].$$

Similarly, with conditional independence of the hierarchical latent variables, $p(\mathbf{h}^1|\mathbf{h}^{2:L}) = p(\mathbf{h}^1|\mathbf{h}^2)$. Thus

$$\begin{aligned}\mathbf{KL}^{1:L} =&\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log q(\mathbf{h}^1|\mathbf{x}) - \log p(\mathbf{h}^1|\mathbf{h}^2) + \log q(\mathbf{h}^{2:L}|\mathbf{h}^1) - \log p(\mathbf{h}^{2:L})\big] \\ =&\underbrace{\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log q(\mathbf{h}^1|\mathbf{x}) - \log p(\mathbf{h}^1|\mathbf{h}^2)\big]}_{\mathbf{KL}^1} + \underbrace{\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log q(\mathbf{h}^{2:L}|\mathbf{h}^1) - \log p(\mathbf{h}^{2:L})\big]}_{\mathbf{KL}^{2:L}}.\end{aligned}$$

We can further expand the $\mathbf{KL}^{2:L}$ term following similar conditional independent rules regarding the tree structure. At level l, we get

$$\mathbf{KL}^{l:L} = \mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log q(\mathbf{h}^{l:L}|\mathbf{h}^{l-1}) - \log p(\mathbf{h}^{l:L})\big].$$

With (21) and (22), it is easy to show that

$$\mathbf{KL}^{l:L} = \underbrace{\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log q(\mathbf{h}^l|\mathbf{h}^{l-1}) - \log p(\mathbf{h}^l|\mathbf{h}^{l+1})\big]}_{\mathbf{KL}^l} + \underbrace{\mathbb{E}_{q(\mathbf{h}^{l:L}|\mathbf{x})}\big[\log q(\mathbf{h}^{l+1:L}|\mathbf{h}^l) - \log p(\mathbf{h}^{l+1:L})\big]}_{\mathbf{KL}^{l+1:L}}. \tag{24}$$

The ELBO (23) can be written as

$$\mathcal{L}(\mathbf{x};\theta) = \mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log p(\mathbf{x}|\mathbf{h}^{1:L})\big] - \sum_{l=1}^{L-1}\mathbf{KL}^l - \mathbf{KL}^L. \tag{25}$$

When $1 \leqslant l \leqslant L-1$

$$\mathbf{KL}^l = \mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log q(\mathbf{h}^l|\mathbf{h}^{l-1}) - \log p(\mathbf{h}^l|\mathbf{h}^{l+1})\big]. \tag{26}$$

As discussed in section 3.2, evaluation of the terms in (25) requires samples of both the posterior and the prior in each layer of the tree structure. According to conditional independence, the expectation regarding variational distribution layer l just depends on layer $l-1$. We can simplify the expectation each term of (25) with the default assumption that all latent variables are generated regarding data sample $\mathbf{x}$. Therefore the ELBO (25) can be simplified as

$$\mathcal{L}(\mathbf{x};\theta) = \mathbb{E}_{q(\mathbf{h}^1|\mathbf{x})}\big[\log p(\mathbf{x}|\widehat{\mathbf{h}}^1)\big] - \sum_{l=1}^{L}\mathbf{KL}^l. \tag{27}$$

The $\mathbf{KL}$ term (26) becomes

$$\mathbf{KL}^l = \mathbb{E}_{q(\mathbf{h}^l|\mathbf{h}^{l-1})}\big[\log q(\mathbf{h}^l|\mathbf{h}^{l-1}) - \log p(\mathbf{h}^l|\widehat{\mathbf{h}}^{l+1})\big].$$

When $l = L$,

$$\mathbf{KL}^L = \mathbb{E}_{q(\mathbf{h}^L|\mathbf{h}^{L-1})}\big[\log q(\mathbf{h}^L|\mathbf{h}^{L-1}) - \log p(\mathbf{h}^L)\big].$$

A.2. Improve ELBO Estimation with Flows

To compute the EBLO, one way is to approximate $\mathbf{KL}$ terms with the latent values generated from a batch of training data samples. In this paper we follow the approach in (Rezende & Mohamed, 2015; Kingma et al., 2016; Berg et al., 2018) using normalizing flows to further improve posterior estimation. At each layer, minimizing $\mathbf{KL}$ term is to is to optimize the parameters of the network so that the posterior is closer to the prior. As shown in Figure 2, for layer l, we can take the encoding-decoding procedures (discussed in section 3.2) as transformation of the posterior distribution from layer l to $l+1$, and then transform it back. By counting in the transformation difference (Rezende & Mohamed, 2015; Kingma et al., 2016; Berg et al., 2018), the $\mathbf{KL}$ at layer l becomes

$$\begin{aligned}\mathbf{KL}^l =&\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\left[\log q(\mathbf{h}^l|\mathbf{h}^{l-1}) + \log\left|\det\frac{\partial \mathbf{h}^l}{\partial \mathbf{h}^{l+1}}\right| + \log\left|\det\frac{\partial \widehat{\mathbf{h}}^{l+1}}{\partial \widehat{\mathbf{h}}^l}\right| - \log p(\mathbf{h}^l|\widehat{\mathbf{h}}^{l+1})\right]\\ \simeq&\frac{1}{M}\sum_{m=1}^{M}\left[\log q(\mathbf{h}^l_m|\mathbf{h}^{l-1}) + \log\left|\det\frac{\partial \mathbf{h}^l_m}{\partial \mathbf{h}^{l+1}_m}\right| + \log\left|\det\frac{\partial \widehat{\mathbf{h}}^{l+1}_m}{\partial \widehat{\mathbf{h}}^l_m}\right| - \log p(\mathbf{h}^l_m|\widehat{\mathbf{h}}^{l+1}_m)\right].\end{aligned}$$

A.3. ELBO of DAG Models

Note that if we reverse the edge directions in a DAG, the resulting graph is still a DAG graph. The nodes can be listed in a topological order regarding the DAG structure as shown in Figure 9.

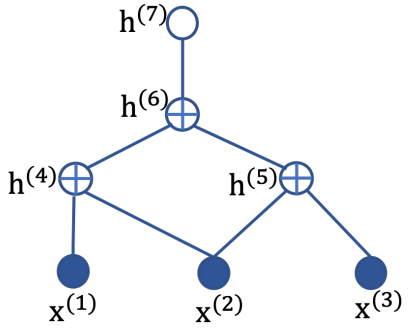


Figure 9. DAG structure. The inverse topology order is $\{$ {1,2,3}, {4,5}, {6}, {7} $\}$, and it corresponds to layers 0 to 3.

By taking the topology order as the layers in tree structures, we can derive the ELBO for DAG structures. Assume the DAG structure has L layers, and the root nodes are in layer L. We denote by $\mathbf{h}$ the vector of latent variables, then following (23) we develop the ELBO as

$$\begin{aligned}\log p(\mathbf{x}) \geqslant \mathcal{L}(x;\theta) =&\mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log\frac{p(\mathbf{x},\mathbf{h})}{q(\mathbf{h}|\mathbf{x})}\right] \\ =& \underbrace{\mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{h})\right]}_{\text{Reconstruction of the data}} \underbrace{-\mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\left[\log q(\mathbf{h}|\mathbf{x}) - \log p(\mathbf{h})\right]}_{\mathbf{KL}}.\end{aligned} \tag{28}$$

Similarly the KL term can be expanded as in the tree structures. For nodes in layer l

$$\mathbf{KL}^{l:L} =\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log q(\mathbf{h}^{l:L}|\mathbf{h}^{1:l-1}) - \log p(\mathbf{h}^{l:L})\big].$$

Note that $ch(l)$ may include nodes from layers lower than $l-1$, and $pa(l)$ may include nodes from layers higher than l. Some nodes in l may not have parent. Based on conditional independence with the topology order of a DAG, we have

$$q(\mathbf{h}^{l:L}|\mathbf{h}^{1:l-1}) = q(\mathbf{h}^l|\mathbf{h}^{1:l-1})q(\mathbf{h}^{l+1:L}|\mathbf{h}^l) = q(\mathbf{h}^l|\mathbf{h}^{1:l-1})q(\mathbf{h}^{l+1:L}|\mathbf{h}^{1:l}) \tag{29}$$

$$p(\mathbf{h}^{l:L}) = p(\mathbf{h}^l|\mathbf{h}^{l+1:L})p(\mathbf{h}^{l+1:L}) \tag{30}$$

Following (24) and with (29-30), we have

$$\mathbf{KL}^{l:L} = \mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[\log q(\mathbf{h}^l|\mathbf{h}^{1:l-1}) - \log p(\mathbf{h}^l|\mathbf{h}^{l+1:L})\big] + \underbrace{\mathbb{E}_{q(\mathbf{h}^{l:L}|\mathbf{x})}\big[\log q(\mathbf{h}^{l+1:L}|\mathbf{h}^{1:l}) - \log p(\mathbf{h}^{l+1:L})\big]}_{\mathbf{KL}^{l+1:L}}.$$

Furthermore,

$$q(\mathbf{h}^l|\mathbf{h}^{1:l-1}) = q(\mathbf{h}^l|\mathbf{h}^{ch(l)}), \qquad p(\mathbf{h}^l|\mathbf{h}^{l+1:L}) = p(\mathbf{h}^l|\mathbf{h}^{pa(l)}).$$

Hence,

$$\mathbf{KL}^{l:L} = \underbrace{\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[q(\mathbf{h}^l|\mathbf{h}^{ch(l)}) - p(\mathbf{h}^l|\mathbf{h}^{pa(l)})\big]}_{\mathbf{KL}^l} + \mathbf{KL}^{l+1:L} \tag{31}$$

For nodes in layer l,

$$\mathbf{KL}^l = \sum_{i\in l} \underbrace{\mathbb{E}_{q(\mathbf{h}^{1:L}|\mathbf{x})}\big[q(\mathbf{h}^{(i)}|\mathbf{h}^{ch(i)}) - p(\mathbf{h}^{(i)}|\mathbf{h}^{pa(i)})\big]}_{\mathbf{KL}^{(i)}}.$$

Recurrently applying (31) to (28) yields

$$\mathcal{L}(\mathbf{x};\theta) = \mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\big[\log p(\mathbf{x}|\mathbf{h})\big] - \sum_{i\in\mathcal{V}\setminus\mathcal{R}_\mathcal{G}} \mathbf{KL}^{(i)} - \sum_{i\in\mathcal{R}_\mathcal{G}} \mathbf{KL}\big(q(\mathbf{h}^{(i)}|\mathbf{h}^{ch(i)})||p(\mathbf{h}^{(i)})\big).$$

For node i,

$$\mathbf{KL}^{(i)} = \mathbb{E}_{q(\mathbf{h}|\mathbf{x})}\big[q(\mathbf{h}^{(i)}|\mathbf{h}^{ch(i)}) - p(\mathbf{h}^{(i)}|\mathbf{h}^{pa(i)})\big].$$

B. Theoretical Proofs

We present in this section the proofs for our Lemma 1 and Theorem 1.

B.1. Proof of Lemma 1

Lemma 1. *Let $\mathcal{G}$ be a well trained tree structured variational flow graphical model with L layers, and i and j are two leaf nodes with a as the closest common ancestor. Given observed value at node i, the value of node j can be approximated with $\widehat{\mathbf{x}}^{(j)} \approx \mathbf{f}_{(a,j)}(\mathbf{f}_{(i,a)}(\mathbf{x}^{(i)}))$. Here $\mathbf{f}_{(i,a)}$ is the flow function path from node i to node a. The conditional density of $\mathbf{x}^{(j)}$ given $\mathbf{x}^{(i)}$ can be approximated with*

$$\log p(\mathbf{x}^{(j)}|\mathbf{x}^{(i)}) \approx \log p(\widehat{\mathbf{h}}^L) - \frac{1}{2}\log\left(\det\left(\mathbf{J}_{\widehat{\mathbf{x}}^{(j)}}(\widehat{\mathbf{h}}^L)^\top \mathbf{J}_{\widehat{\mathbf{x}}^{(j)}}(\widehat{\mathbf{h}}^L)\right)\right).$$

Proof. Without loss generality, we assume that there are relationships among different data sections, and the value of one section can be partially or approximately imputed by other sections. According to the aggregation rule (b) discussed in section 3.3, at an aggregation node a, the latent value of a child node j has the same reconstruction value as the parent node. The reconstruction of the child node j can be approximated with the reconstruction of the parent node, i.e., $\widehat{\mathbf{h}}^{(j)} \approx \mathbf{f}_{(a,j)}(\widehat{\mathbf{h}}^{(a)})$. Recalling the reconstruction term in the ELBO (6), at each node we have $\mathbf{h}^{(a)} \approx \widehat{\mathbf{h}}^{(a)}$. Hence for node a's descendent j, we have $\widehat{\mathbf{h}}^{(j)} \approx \mathbf{f}_{(a,j)}(\mathbf{h}^{(a)})$, and $\mathbf{f}_{(a,j)}$ is the flow function path from a to j. The value of node a can be approximated by the value of its descendent i that has observation, i.e., $\mathbf{h}^{(a)} \approx \mathbf{f}_{(i,a)}(\mathbf{h}^{(i)})$. Hence, we have $\widehat{\mathbf{x}}^{(j)} \approx \mathbf{f}_{(a,j)}(\mathbf{f}_{(i,a)}(\mathbf{x}^{(i)}))$.

To compute node j's conditional distribution given the observed node i, we can use the forward passing to compute the root's reconstruction value $\widehat{\mathbf{h}}^L$. Node j's reconstruction value $\widehat{\mathbf{x}}^{(j)}$ can be imputed by backward passing the message at the root. The density value of $\widehat{\mathbf{h}}^L$ can be computed with the prior distribution of the root. The conditional density of $\widehat{\mathbf{x}}^{(j)}$ can be computed using the change of variable theorem, and it is known in the context of geometric measure theory as the smooth coarea formula (Hanson, 1994; Krantz & Parks, 2008). It reads

$$p(\mathbf{x}^{(j)}|\mathbf{x}^{(i)}) \approx p(\widehat{\mathbf{h}}^L)\det\left(\mathbf{J}_{\widehat{\mathbf{x}}^{(j)}}(\widehat{\mathbf{h}}^L)^\top \mathbf{J}_{\widehat{\mathbf{x}}^{(j)}}(\widehat{\mathbf{h}}^L)\right)^{-\frac{1}{2}}.$$

Applying the logarithm operator on both sides concludes the proof of our Lemma.

□

B.2. Proof of Theorem 1

Theorem 1. *Assume that the observed data is distributed according to the model given by (17) and (18). Let the following assumptions holds,*

(a) The sufficient statistics $T_{ij}(h)$ are differentiable almost everywhere and their derivatives $\partial T_{i,j}/\partial_h$ are nonzero almost surely for all $h \in \mathcal{H}_i$, $1 \leqslant i \leqslant d$ and $1 \leqslant j \leqslant m$.

(b) There exist $(dm+1)$ distinct conditions $\mathbf{u}^{(0)}$, ..., $\mathbf{u}^{(dm)}$ such that the matrix

$$\mathbf{L} = [\lambda(\mathbf{u}^{(1)}) - \lambda(\mathbf{u}^{(0)}), ..., \lambda(\mathbf{u}^{(dm)}) - \lambda(\mathbf{u}^{(0)})]$$

of size $dm \times dm$ is invertible.

Then the model parameters $\mathbf{T}(\mathbf{h}^{(t)}) = \mathbf{A}\widehat{\mathbf{T}}(\mathbf{z}^{(t)}) + \mathbf{c}$. Here $\mathbf{A}$ is a $dm \times dm$ invertible matrix and $\mathbf{c}$ is a vector of size dm.

Proof. The conditional probabilities of $p_{\mathbf{T},\lambda,\mathbf{f}_t^{-1}}(\mathbf{x}^{(t)}|\mathbf{u})$ and $p_{\widehat{\mathbf{T}},\widehat{\lambda},\mathbf{g}}(\mathbf{x}^{(t)}|\mathbf{u})$ are assumed to be the same in the limit of infinite data. By expanding the probability density functions with the correct change of variable, we have

$$\log p_{\mathbf{T},\lambda}(\mathbf{h}^{(t)}|\mathbf{u}) + \log\left|\det \mathbf{J}_{\mathbf{f}_t}(\mathbf{x}^{(t)})\right| = \log p_{\widehat{\mathbf{T}},\widehat{\lambda}}(\mathbf{z}^{(t)}|\mathbf{u}) + \log\left|\det \mathbf{J}_{g^{-1}}(\mathbf{x}^{(t)})\right|.$$

Let $\mathbf{u}^{(0)}, ..., \mathbf{u}^{(dm)}$ be from condition (b). We can subtract this expression of $\mathbf{u}^{(0)}$ from some $\mathbf{u}^{(v)}$. The Jacobian terms will be removed since they do not depend $\mathbf{u}$,

$$\log p_{\mathbf{h}^{(t)}}(\mathbf{h}^{(t)}|\mathbf{u}^{(v)}) - \log p_{\mathbf{h}^{(t)}}(\mathbf{h}^{(t)}|\mathbf{u}^{(0)}) = \log p_{\mathbf{z}^{(t)}}(\mathbf{z}^{(t)}|\mathbf{u}^{(v)}) - \log p_{\mathbf{z}^{(t)}}(\mathbf{z}^{(t)}|\mathbf{u}^{(0)}). \tag{32}$$

Both conditional distributions in equation 32 belong to the exponential family. Eq. (32) thus reads

$$\begin{aligned}
&\sum_{i=1}^{d}\left[\log\frac{Z_i(\mathbf{u}^{(0)})}{Z_i(\mathbf{u}^{(v)})} + \sum_{j=1}^{m} T_{i,j}(\mathbf{h}^{(t)})\big(\lambda_{i,j}(\mathbf{u}^{(v)}) - \lambda_{i,j}(\mathbf{u}^{(0)})\big)\right] \\
=&\sum_{i=1}^{d}\left[\log\frac{\widehat{Z}_i(\mathbf{u}^{(0)})}{\widehat{Z}_i(\mathbf{u}^{(v)})} + \sum_{j=1}^{m} \widehat{T}_{i,j}(\mathbf{z}^{(t)})\big(\widehat{\lambda}_{i,j}(\mathbf{u}^{(v)}) - \widehat{\lambda}_{i,j}(\mathbf{u}^{(0)})\big)\right].
\end{aligned}$$

Here the base measures Q_is are canceled out. Let $\bar{\lambda}(\mathbf{u}) = \lambda(\mathbf{u}) - \lambda(\mathbf{u}^{(0)})$. The above equation can be expressed, with inner products, as follows

$$\langle \mathbf{T}(\mathbf{h}^{(t)}), \bar{\lambda}\rangle + \sum_i \log\frac{Z_i(\mathbf{u}^{(0)})}{Z_i(\mathbf{u}^{(v)})} = \langle \widehat{\mathbf{T}}(\mathbf{z}^{(t)}), \bar{\widehat{\lambda}}\rangle + \sum_i \log\frac{\widehat{Z}_i(\mathbf{u}^{(0)})}{\widehat{Z}_i(\mathbf{u}^{(v)})}, \quad \forall\, v, 1 \leqslant v \leqslant dm.$$

Combine dm equations together and we can rewrite them in matrix equation form as following

$$\mathbf{L}^\top \mathbf{T}(\mathbf{h}^{(t)}) = \widehat{\mathbf{L}}^\top \widehat{\mathbf{T}}(\mathbf{z}^{(t)}) + \mathbf{b}.$$

Here $b_v = \sum_{i=1}^{d} \log\frac{\widehat{Z}_i(\mathbf{u}^{(0)})Z_i(\mathbf{u}^{(v)})}{\widehat{Z}_i(\mathbf{u}^{(v)})Z_i(\mathbf{u}^{(0)})}$. We can multiply $\mathbf{L}^\top$'s inverse with both sized of the equation,

$$\mathbf{T}(\mathbf{h}^{(t)}) = \mathbf{A}\widehat{\mathbf{T}}(\mathbf{z}^{(t)}) + \mathbf{c}. \tag{33}$$

Here $\mathbf{A} = \mathbf{L}^{-1\top}\widehat{\mathbf{L}}^\top$, and $\mathbf{c} = \mathbf{L}^{-1\top}\mathbf{b}$. By Lemma 1 from (Khemakhem et al., 2020), there exist m distinct values $h_1^{(t),i}$ to $h_m^{(t),i}$ such that $\left[\frac{dT_i}{dh^{(t),i}}(h_1^{(t),i}), ..., \frac{dT_i}{dh^{(t),i}}(h_m^{(t),i})\right]$ are linearly independent in $\mathbb{R}^m$, for all $1 \leqslant i \leqslant d$. Define m vectors $\mathbf{h}_v^{(t)} = [h_v^{(t),1}, ..., h_v^{(t),d}]$ from points given by this lemma. We obtain the following Jacobian matrix

$$\mathbf{Q} = \left[\mathbf{J}_{\mathbf{T}}(\mathbf{h}_1^{(t)}), ..., \mathbf{J}_{\mathbf{T}}(\mathbf{h}_m^{(t)})\right],$$

where each entry is the Jacobian of size $dm \times d$ from the derivative of Eq. (33) regarding the m vectors $\{\mathbf{h}_j^{(t)}\}_{j=1}^m$. Hence $\mathbf{Q}$ is a $dm \times dm$ invertible by the lemma and the fact that each component of $\mathbf{T}$ is univariate. We can construct a corresponding matrix $\widehat{\mathbf{Q}}$ with the Jacobian of $\widehat{\mathbf{T}}(\mathbf{g}^{-1} \circ \mathbf{f}_t^{-1}(\mathbf{h}^{(t)}))$ computed at the same points and get

$$\mathbf{Q} = \mathbf{A}\widehat{\mathbf{Q}}\,.$$

Here $\widehat{\mathbf{Q}}$ and $\mathbf{A}$ are both full rank as $\mathbf{Q}$ is full rank. □

According to Theorem 1, the proposed model not only can identify global latent factors, but also identify the latent factors for each section with enough auxiliary information. VFG provides a potential approach to learn the latent hierarchical structures from datasets.

C. Additional Numerical Experiments

In all the experiments of the paper, we use the same coupling block (Dinh et al., 2016) to construct different flow functions. The coupling block consists in three fully connected layers (of dimension 64) separated by two RELU layers along with the coupling trick. Each flow function has block number $b \geqslant 2$. All latent variables, $\mathbf{h}^i, i \in \mathcal{V}$ are forced to be non-negative via Sigmoid or RELU functions. Non-negativeness can help the model to identify sparse structures of the latent space.

C.1. California Housing Dataset

We further investigate the method on a real dataset. The California Housing (chs) dataset has 8 feature entries and 20 640 data samples. We use the first 20 000 samples for training and 100 of the rest for testing. We get 4 data sections, and each section contains 2 variables. In the testing set, the second section is assumed missing for illustration purposes, as the goal is to impute this missing section. Here, we construct a tree structure VFG with 2 layers. The first layer has two aggregation nodes, and each of them has two children. The second layer consists of one aggregation node that has two children connecting with the first layer. Each flow function has 4 coupling blocks. We can see Table 1 that our model yields significantly better results than any other method in terms of prediction error.

Methods	*Imputation MSE*
Mean Value	1.993
MICE	1.951
Iterative Imputation	1.966
KNN (k=3)	1.974
KNN (k=5)	1.969
VFG	**1.356**

Table 1. California Housing dataset: Imputation Mean Squared Error (MSE) results.

C.2. MNIST

For MNIST, we construct a tree structure VFG model depicted in Figure 10. In the first layer, there are 4 flow functions, and each of them takes 14×14 image blocks as the input. Thus a 28×28 input image is divided into four 14×14 blocks as the input of VFG model. The four nodes are aggregated as the input of the upper layer flow.

C.2.1. ELBO AND LOG-LIKELIHOOD OF MNIST

In Table 2, the negative evidence lower bound (-ELBO) and the estimated negative likelihood (NLL) for baseline methods. The results of the baseline methods are from (Berg et al., 2018). These methods are VAE based methods enhanced with normalizing flows. They use 16 flows to improve the posterior estimation. SNF is orthogonal sylvester flow method with a bottleneck of M = 32. We set the VFG coupling block number with two different values, and they achieve similar performance. Compared to VAE based methods, the proposed VFG model can achieve significant improvement on both ELBO and NLL values.

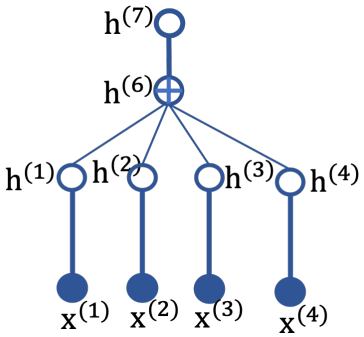


Figure 10. The tree structure for MNIST.

Model	-ELBO	NLL
VAE (Kingma & Welling, 2013)	86.55	82.14
Planer (Rezende & Mohamed, 2015)	86.06	81.91
IAF (Kingma et al., 2016)	84.20	80.19
SNF (Berg et al., 2018)	83.32	80.22
VFG (b=4)	74.01	**67.58**
VFG (b=6)	**73.21**	67.74

Table 2. Negative log-likelihood and free energy (negative evidence lower bound) for static MNIST.

C.2.2. Latent Representation Learning on MNIST

Figure 11 presents the t-SNE plot of the root latent variables from VFG trained without labels. The figure clearly shows that even without label information, different digits' representation are roughly scattered in different areas. Compared to Figure 7 in section 5.2, label information indeed can improve the latent representation learning.

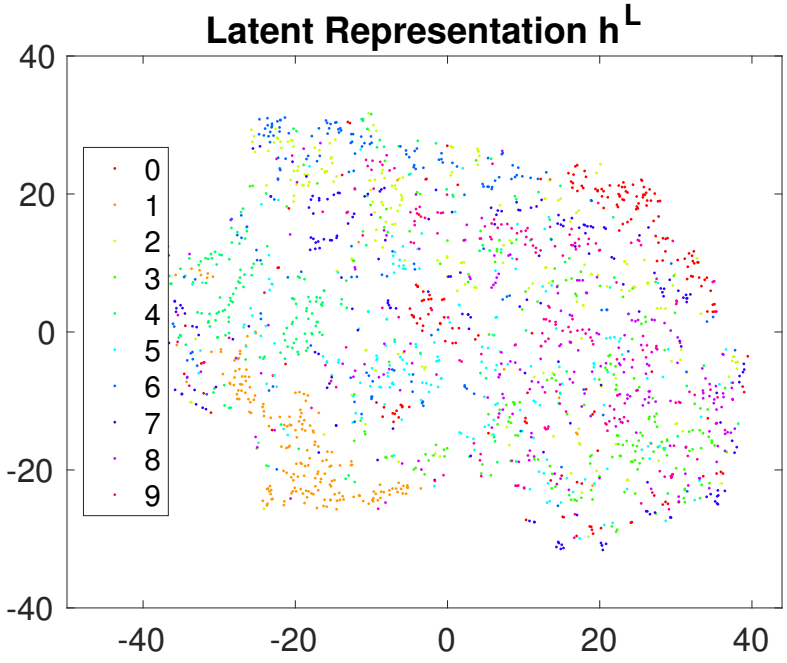
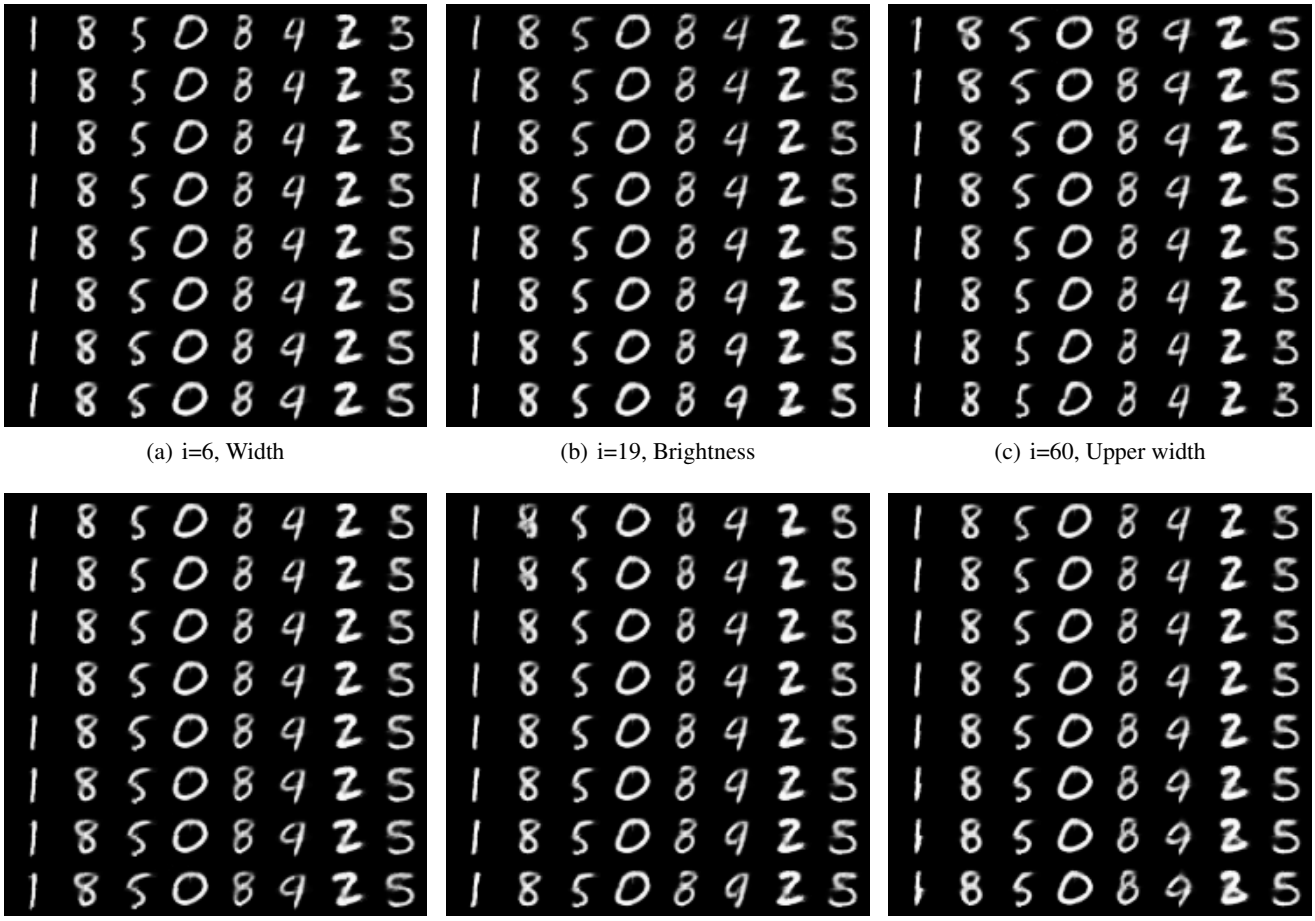


Figure 11. MNIST: t-SNE plot of latent variables from VFG learned without labels.

C.2.3. Disentanglement on MNIST

We study disentanglement on MNIST with our proposed VFG model introduced in section 5.2. But different from the model in section 5.2, here, the distribution parameter λ for all latent variables are set to be trainable across all layers. Each digit has its trainable vector, $\lambda \in \mathbb{R}^d$ that is used across all layers. To show the disentanglement of learned latent representation, we first obtain the root latent variables of a set of images through forward message passing. Each latent variable's values are changed increasingly within a range centered at the value of the latent variable obtained from last step. This perturbation is performed for each image in the set. Figure 12 shows the change of images by increasing one latent variable from a small value to a larger one. The figure presents some of the latent variables that have obvious effects on images, and most of the $d = 196$ variables do not impact the generation significantly. Latent variables $i = 6$ and $i = 60$ control the digit width.

Variable $i = 19$ affects the brightness. $i = 92, i = 157$ and some of the variables not displayed here control the style of the generated digits.



(a) i=6, Width (b) i=19, Brightness (c) i=60, Upper width

(d) i=92 (e) i=119, Angle (f) i=157

Figure 12. MNIST: Increasing each latent variable from a small value to a larger one.