Learning Structured Latent Factors from Dependent Data: A Generative Model Framework from Information-Theoretic Perspective

Ruixiang Zhang [1 2] **Masanori Koyama** [3] **Katsuhiko Ishiguro** [3]

Abstract

Learning controllable and generalizable representation of multivariate data with desired structural properties remains a fundamental problem in machine learning. In this paper, we present a novel framework for learning generative models with various underlying structures in the latent space. We represent the inductive bias in the form of mask variables to model the dependency structure in the graphical model and extend the theory of multivariate information bottleneck (Friedman et al., 2001) to enforce it. Our model provides a principled approach to learn a set of semantically meaningful latent factors that reflect various types of desired structures like capturing correlation or encoding invariance, while also offering the flexibility to automatically estimate the dependency structure from data. We show that our framework unifies many existing generative models and can be applied to a variety of tasks, including multi-modal data modeling, algorithmic fairness, and out-of-distribution generalization.

1. Introduction

Learning structured latent representation of multivariate data is a fundamental problem in machine learning. Many latent variable generative models have been proposed to date based on different inductive biases that reflect the model's assumptions or people's domain knowledge. For instance, the objectives of the family of β-VAEs (Higgins et al., 2016; Chen et al., 2018; Kim & Mnih, 2018) try to enforce a coordinate-wise independent structure among latent variables to discover disentangled factors of variations.

While these methods have been proven useful in the field of applications on which they were evaluated, most of them are built-in rather heuristic ways to encode the desired structure. One usually needs to construct an entirely different model whenever the domain of application changes. In general, the type of inductive bias differs significantly across different applications. It is a burden to craft a different architecture for each application, and there have not been many studies done for the general and unified way of explicitly representing an inductive bias to be enforced in generative models.

In this paper, we propose a framework of generative model that can represent various types of inductive biases in the form of Bayesian networks. Our method can not only unify many existing generative models in previous studies, but it also can lead to new insights about establishing connections between different models across different domains and extending them to new applications.

We summarize our contributions in this work as: (i) We propose a novel general framework of probabilistic generative model with explicit dependency structure representation to learn structured latent representation of multivariate data. (ii) We propose an information-theoretic training objective by generalizing the multivariate information bottleneck theory to encode prior knowledge or impose inductive bias. (Sec. 3.3) (iii) We propose a flexible and tractable inference model with linear number of inference networks coupled with super-exponential number of possible dependency structures to model exponential number of inference distributions. (Sec. 3.4) (iv) We show that our proposed framework unifies many existing models and demonstrate its effectiveness in different application tasks, including multi-modal data generative modeling, algorithmic fairness, and out-of-distribution generalization.

2. Background

2.1. Notations

We use capital letters (i.e. $\mathrm{X} \equiv \mathrm{X}_{1:N}$) to denote a vector of N random variables, and lower case letters (i.e. $\mathbf{x}$) for the values. We use $P(\mathrm{X})$ to denote the probability distribution and corresponding density with $p(\mathbf{x})$. Given a set $\mathbb{S} \subseteq \{1, 2, \dots, N\}$ of indexes, we use $\mathrm{X}^{\mathbb{S}} \equiv [\mathrm{X}_i]_{i\in\mathbb{S}}$ to represent corresponding subset of random variables. Similar notation is used for binary indicator vector $\mathbf{b}$ that $\mathrm{X}^{\mathbf{b}} \equiv [\mathrm{X}_i]_{\mathbf{b}_i=1}$.

---

[1]Mila [2]Université de Montréal, Montréal, Canada [3]Preferred Networks Inc., Tokyo, Japan. Correspondence to: Ruixiang Zhang <ruixiang.zhang@umontreal.ca>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

2.2. Probability and information theory

A **Bayesian network** $\mathcal{G} \equiv (\mathcal{V}, \mathcal{E})$ defined over random variables X is a directed acyclic graph, consisting of a set of nodes $\mathcal{V} \equiv \{\mathrm{X}_i\}_{i=1}^N$ and a set of directed edges $\mathcal{E} \subseteq \mathcal{V}^2$. A node $\mathbf{u}$ is called a *parent* of $\mathbf{v}$ if $(\mathbf{v}, \mathbf{u}) \in \mathcal{E}$, and for each random variable X_i, the set of parents of X_i is denoted by $Pa^{\mathcal{G}}_{\mathrm{X}_i}$. We use $\mathcal{G}^{\emptyset}$ to denote an empty Bayesian network $\mathcal{G}^{\emptyset} \equiv (\mathcal{V}, \emptyset)$. If a distribution $P(\mathrm{X})$ is consistent with a Bayesian network $\mathcal{G}$, then it can be factorized as $p(\mathbf{x}) = \prod_i p(\mathbf{x}_i \mid Pa^{\mathcal{G}}_{\mathbf{x}_i})$, denoted by $p \models \mathcal{G}$.

We then briefly introduce the information theory concepts used in this paper here. The Shannon **Entropy** is defined as $\mathcal{H}(\mathrm{X}) = -\mathbb{E}_{p(\mathbf{x})} \log p(\mathbf{x})$ to measure the average number of bits needed to encode values of $\mathrm{X} \sim P(\mathrm{X})$. The **Kullback–Leibler Divergence** (KLD) is one of the most fundamental distance between probability distributions defined as $D_{\mathrm{KL}}(P \parallel Q) = \mathbb{E}_p \log \frac{p}{q}$. **Mutual Information** $\mathcal{I}(\mathrm{X};\mathrm{Y}) = \mathbb{E}_{p(\mathbf{x},\mathbf{y})} \log \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}$ quantifies the mutual dependence between two random variables X and Y. The mutual information is zero if and only if X and Y are independent. **Multi-Information** is one of multivariate mutual information defined as $\mathcal{I}(\mathrm{X}_1, \ldots, \mathrm{X}_N) = D_{\mathrm{KL}}(p(\mathbf{x}_{1:N}) \parallel \prod_{i=1}^N p(\mathbf{x}_i))$, which generalizes the mutual information concept to quantify the multivariate statistical independence for arbitrary number of random variables. (Lin, 1991) proposed a **generalized Jensen-Shannon divergence** defined as $D^{\boldsymbol{\pi}}_{\mathrm{JS}} = \mathcal{H}\left(\sum_{i=1}^N \pi_i P_i\right) - \sum_{i=1}^N \pi_i \mathcal{H}(P_i)$, where $P_1, \ldots, P_N$ are N distributions with weights $\pi_1, \ldots, \pi_N$. Commonly used Jensen-Shannon divergence (JSD) can be seen as a special case when $N = 2$ and $\pi_1 = \pi_2 = \frac{1}{2}$. (Nielsen, 2019) further generalized the arithmetic mean $\sum_{i=1}^N \pi_i P_i$ to other abstract means and proposed closed-form results of geometric mean of exponential family distributions and the divergence among them.

As shown in (Friedman et al., 2001), if a distribution $P(\mathrm{X}_{1:N})$ is consistent with a Bayesian network $\mathcal{G}$, the multi-information $\mathcal{I}(\mathrm{X})$ can be expressed as a sum of all local mutual information terms: $\mathcal{I}(\mathrm{X}) = \sum_{i=1}^N \mathcal{I}\left(\mathrm{X}_i; Pa^{\mathcal{G}}_{\mathrm{X}_i}\right)$. Then the multi-information in $P(\mathrm{X})$ with respect to an arbitrary valid Bayesian network $\mathcal{G}$ can be defined [1] as $\mathcal{I}^{\mathcal{G}}_p(\mathrm{X}) = \sum_{i=1}^N \mathcal{I}^{\mathcal{G}}_p\left(\mathrm{X}_i; Pa^{\mathcal{G}}_{\mathrm{X}_i}\right)$. The **M-projection** (Koller & Friedman, 2009; Friedman et al., 2001) of distribution $P(\mathrm{X})$ to the set of distribution that is consistent with a Bayesian network $\mathcal{G}$ is defined as $\mathbb{D}(p \parallel \mathcal{G}) = \min_{q \models \mathcal{G}} D_{\mathrm{KL}}(p \parallel q)$. Then the following results was introduced in (Friedman et al., 2001)

$$\mathbb{D}(p \parallel \mathcal{G}) = \min_{q \models \mathcal{G}} D_{\mathrm{KL}}(p \parallel q) = \mathcal{I}_p(\mathrm{X}) - \mathcal{I}^{\mathcal{G}}_p(\mathrm{X}) \tag{1}$$

where we use subscript to denote the distribution that the mutual information term is evaluated with, and we use superscript to denote the graphical structure that the indicates set of parent nodes used in $\mathcal{I}^{\mathcal{G}}_p\left(\mathrm{X}_i; Pa^{\mathcal{G}}_{\mathrm{X}_i}\right)$.

2.3. Variational autoencoder

Variational autoencoder (VAE) (Kingma & Welling, 2014) is a probabilistic latent variable generative model $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$, where $p_{\boldsymbol{\theta}}(\mathbf{z})$ is the prior of latent variables Z and $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$ is the likelihood distribution for observed variable X. The generative model is often optimized together with a tractable distribution $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ that approximates the posterior distribution. The distributions are usually parametrized by neural networks with parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. The inference model and generation model are jointly optimized by a lower-bound of the KLD between $q_{\boldsymbol{\phi}}$ and $p_{\boldsymbol{\theta}}$ in the augmented space (X, Z), namely *ELBO*:

$$\mathbb{E}_{q_{\boldsymbol{\phi}}} \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x})} D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})) \equiv \mathcal{L}_{\mathrm{ELBO}} \tag{2}$$

Note $-\mathcal{L}_{\mathrm{ELBO}} \geq D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{x}) q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}))$ where $q_{\boldsymbol{\phi}}(\mathbf{x}) = p_{\mathrm{data}}(\mathbf{x})$ denotes the empirical data distribution. The above objective can be optimized efficiently with the re-parametrization trick (Kingma & Welling, 2014; 2019).

2.4. Multivariate information bottleneck

Multivariate Information Bottleneck (MIB) theory proposed by (Friedman et al., 2001; Slonim et al., 2006) extends the information bottleneck theory (Tishby et al., 2000) to multivariate setting. Given a set of observed variable X, MIB framework introduced a Bayesian network $\mathcal{G}^{\mathrm{in}}$ to define the solution space of latent variables Z as $q(\mathrm{X}, \mathrm{Z}) \models \mathcal{G}^{\mathrm{in}}$. Another Bayesian network $\mathcal{G}^{\mathrm{out}}$ is introduced to specify the relevant information to be preserved in Z. Then the MIB functional objective is defined as $\mathcal{L}^1_{MIB}(q) = \mathcal{I}^{\mathcal{G}^{\mathrm{in}}}_q(\mathrm{X}) - \beta \mathcal{I}^{\mathcal{G}^{\mathrm{out}}}_q(\mathrm{X})$. An alternative structural MIB functional objective is defined as $\mathcal{L}^2_{MIB}(q) = \mathcal{I}^{\mathcal{G}^{\mathrm{in}}}_q(\mathrm{X}) + \gamma \mathbb{D}(q(\mathbf{x}, \mathbf{z}) \parallel \mathcal{G}^{\mathrm{out}})$, and further relaxed by (Elidan & Friedman, 2005) as $\mathcal{L}^2_{MIB}(q, p) = \mathcal{I}^{\mathcal{G}^{\mathrm{in}}}_q(\mathrm{X}) + \gamma D_{\mathrm{KL}}(q(\mathbf{x}, \mathbf{z}) \parallel p(\mathbf{x}, \mathbf{z}))$. We refer to (Friedman et al., 2001; Elidan & Friedman, 2005) for more details of MIB theory.

3. Framework

3.1. Preliminaries

Given a dataset $\mathcal{D} = \left\{\mathbf{x}^d\right\}_{d=1}^{|\mathcal{D}|}$, we assume that observations are generated from some random process governed by a set of latent factors, which could be categorized into two types: private latent factors $\mathrm{U} \equiv \mathrm{U}_{1:N} \equiv [\mathrm{U}_1, \mathrm{U}_2, \ldots, \mathrm{U}_N]$ and common latent factors $\mathrm{Z} \equiv \mathrm{Z}_{1:M} \equiv [\mathrm{Z}_1, \mathrm{Z}_2, \ldots, \mathrm{Z}_M]$. We use U_i to denote the latent factors that are exclusive

[1]Note that $P(\mathrm{X})$ is not necessarily consistent with $\mathcal{G}$ here

to the variable $\mathbf{x}_i$ and assume a jointly independent prior distribution $P(\mathrm{U})$. We use Z to denote the latent factors that are possibly shared by some subset of observed variables and assume a prior distribution $P(\mathrm{Z})$. The dimension of each U_i and Z_j is arbitrary.

3.2. Generative model with explicit dependency structure representation

Generation model We explicitly model the dependency structure from Z to X in the random generation process with a binary matrix variable $\mathrm{M}^p \equiv \left[\mathrm{M}^p_{ij}\right] \in \{0,1\}_{N\times M}$. $\mathrm{M}^p_{ij} = 1$ when the latent factor Z_j contributes to the random generation process of X_i, or otherwise $\mathrm{M}^p_{ij} = 0$. Let $\mathrm{M}^p_i = [\mathrm{M}^p_{i1}, \mathrm{M}^p_{i2}, \ldots, \mathrm{M}^p_{iM}]$ denote the i-th row of M. we can define our generative model $p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z},\mathbf{u})$ as

$$p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z},\mathbf{u}) = p_{\boldsymbol{\theta}}(\mathbf{z})\prod_{i=1}^{N} p_{\boldsymbol{\theta}}(\mathbf{u}_i)\prod_{i=1}^{N} p_{\boldsymbol{\theta}}(\mathbf{x}_i \mid \mathbf{z}^{\mathbf{m}^p_i}, \mathbf{u}_i) \quad (3)$$

where $\boldsymbol{\theta}$ is the parameter for parameterizing the generation model distribution. The structure of the generation model is illustrated by Bayesian network $\mathcal{G}^p_{\text{full}}$ in Figure 2, where the structural variable M^p is depicted as the dashed arrows.

Inference We introduce an inference model to approximate the true posterior distributions. We introduce another binary matrix variable $\mathrm{M}^q \equiv \left[\mathrm{M}^q_{ij}\right] \in \{0,1\}_{N\times M}$. $\mathrm{M}^q_{ij} = 1$ when the observed variable X_i contributes to the inference process of Z_j, or otherwise $\mathrm{M}^q_{ij} = 0$. Let $\mathrm{M}^q_j = \left[\mathrm{M}^q_{1j}, \mathrm{M}^q_{2j}, \ldots, \mathrm{M}^q_{Nj}\right]$ denote the j-th column of M^q. We assume that latent variables are conditional jointly independent given observed variables. Then we could define our inference model $q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z},\mathbf{u})$ as:

$$q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z},\mathbf{u}) = q_{\boldsymbol{\phi}}(\mathbf{x})\prod_{i=1}^{N} q_{\boldsymbol{\phi}}(\mathbf{u}_i \mid \mathbf{x}_i)\prod_{j=1}^{M} q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}^{\mathbf{m}^q_j}) \quad (4)$$

where $\boldsymbol{\phi}$ is the parameter for parameterizing the inference distribution. The structure of the inference model is illustrated by Bayesian network $\mathcal{G}^q_{\text{full}}$ in Figure 2, where the structural variable M^q is depicted as the dashed arrows.

3.3. Learning from information-theoretic perspective

We motivate our learning objective based on the MIB (Friedman et al., 2001) theory. We can define a Bayesian network $\mathcal{G}^q \equiv (\mathcal{V}^q, \mathcal{E}^q)$ that is consistent with the inference model distribution $q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z},\mathbf{u}) \models \mathcal{G}^q$ according to M^q. A directed edge from X_i to U_i is added for each $i \in \{1,2,\ldots,N\}$ and an edge from X_i to Z_j is added if and only if $\mathrm{m}^q_{ij} = 1$. Note that we could omit all edges between observed variables in $\mathcal{G}^q$ as shown in (Friedman et al., 2001; Elidan & Friedman, 2005). A Bayesian network $\mathcal{G}^p \equiv (\mathcal{V}^p, \mathcal{E}^p)$ can be constructed according to M^p in a similar way. As introduced in Eq. 2.4, we have the following structural variational objective from the MIB theory:

$$\min_{p_{\boldsymbol{\theta}}\models\mathcal{G}^p, q_{\boldsymbol{\phi}}\models\mathcal{G}^q} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi}) = \mathcal{I}^{\mathcal{G}^q}_q + \gamma D_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \parallel p_{\boldsymbol{\theta}}) \quad (5)$$

The above objective provides a principled way to trade-off between (i)the compactness of learned latent representation measured by $\mathcal{I}^{\mathcal{G}^q}_q$, and (ii)the consistency between $q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z},\mathbf{u})$ and $p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z},\mathbf{u})$ measured by the KLD, through $\gamma > 0$.

We further generalize this objective to enable encoding a broader class of prior knowledge or desired structures into the latent space. We prescribe the dependency structure and conditional independence rules that the learned joint distribution of $(\mathbf{x},\mathbf{z},\mathbf{u})$ should follow, in the form of a set of Bayesian networks $\left\{\mathcal{G}^k \equiv \left(\mathcal{V}^k, \mathcal{E}^k\right)\right\}, k = 1,\ldots,K$. We optimize over the inference distributions $q_{\boldsymbol{\phi}}$ to make it as consistent with $\mathcal{G}^k$ as possible, measured by its M-projection to $\mathcal{G}^k$. Formally we have the following constrained optimization objective:

$$\begin{aligned}&\min_{p_{\boldsymbol{\theta}}\models\mathcal{G}^p, q_{\boldsymbol{\phi}}\models\mathcal{G}^q} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi}) = D_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \parallel p_{\boldsymbol{\theta}})\\ &\text{s.t.} \quad \mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^k) = 0 \quad k = 1,2,\ldots,K\end{aligned} \quad (6)$$

In this way, we impose the preferences over the structure of learned distributions as explicit constraints. We relax the above constrained optimization objective with generalized Lagrangian

$$\max_{\boldsymbol{\beta}\geq\mathbf{0}} \min_{p_{\boldsymbol{\theta}}\models\mathcal{G}^p, q_{\boldsymbol{\phi}}\models\mathcal{G}^q} \mathcal{L} = D_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \parallel p_{\boldsymbol{\theta}}) + \sum_{k=1}^{K}\beta_k \mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^k) \quad (7)$$

where $\boldsymbol{\beta} \equiv [\beta_1, \beta_2, \ldots, \beta_K]$ is the vector of Lagrangian multipliers. In this work we fix $\boldsymbol{\beta}$ as constant hyper-parameters, governing the trade-off between structural regularization and distribution consistency matching. Following the idea proposed in (Zhao et al., 2018), we could also generalize the distribution matching loss by using a vector of T *cost functions* $\boldsymbol{C} \equiv [\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_T]$ and a vector of Lagrangian multipliers $\boldsymbol{\alpha} \equiv [\alpha_1, \alpha_2, \ldots, \alpha_T]$. Each $\mathbb{C}_i$ can be any probability distribution divergence between $q_{\boldsymbol{\phi}}$ and $p_{\boldsymbol{\theta}}$, or any measurable cost function defined over corresponding samples. Thus we could decompose the overall objective as

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{str\_reg}}\\ \mathcal{L}_{\text{dist}} &= \sum_{t=1}^{T}\alpha_t \mathbb{C}_t(q_{\boldsymbol{\phi}} \parallel p_{\boldsymbol{\theta}}), \quad \boldsymbol{\alpha} \geq \mathbf{0}\\ \mathcal{L}_{\text{str\_reg}} &= \sum_{k=1}^{K}\beta_k \mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^k), \quad \boldsymbol{\beta} \geq \mathbf{0}\,.\end{aligned} \quad (8)$$

By setting $\mathbb{C}_1 = D_{\mathrm{KL}}(q_{\boldsymbol{\phi}} \parallel p_{\boldsymbol{\theta}})$ and $\mathcal{G}^1 = \mathcal{G}^{\emptyset}$, we can obtain that the original MIB structural variational objective in Eq. 5 as a special case. We include the detailed proof in Appendix. A.1.

3.4. Tractable inference and generation

Though we have our generation and inference model defined in Sec. 3.2, it's not clear yet how we practically parametrize $q_{\boldsymbol{\phi}}$ and $p_{\boldsymbol{\theta}}$ in a tractable and flexible way, to handle super-exponential number of possible structures $\mathrm{M}^p, \mathrm{M}^q$ and efficient inference and optimization.

Inference model We identify the key desiderata of our inference model defined in Eq. 4 as (i)being compatible with any valid structure variable M^q and (ii)being able to handle missing observed variables in $q(\mathbf{z}_j \mid \mathbf{x}^{\mathbf{m}^q_j})$, in an unified and principled way. Building upon the assumption of our generation model distribution $p_{\boldsymbol{\theta}}$ in Eq. 3 that all observed variables X are conditional jointly independent given Z, we have following factorized formulation in the true posterior distribution $p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$ by applying Bayes' rule:

$$\begin{aligned} p_{\boldsymbol{\theta}}\left(\mathbf{z} \mid \mathbf{x}^{\mathbb{S}}\right) &= \frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathbb{S}} \mid \mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathbb{S}})} = \frac{p_{\boldsymbol{\theta}}(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathbb{S}})} \prod_{i \in \mathbb{S}} p_{\boldsymbol{\theta}}\left(\mathbf{x}_i \mid \mathbf{z}\right) \\ &= \frac{p_{\boldsymbol{\theta}}(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x}^{\mathbb{S}})} \prod_{i \in \mathbb{S}} \frac{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}_i) p_{\boldsymbol{\theta}}(\mathbf{x}_i)}{p_{\boldsymbol{\theta}}(\mathbf{z})} \propto p_{\boldsymbol{\theta}}(\mathbf{z}) \prod_{i \in \mathbb{S}} \frac{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}_i)}{p_{\boldsymbol{\theta}}(\mathbf{z})} \end{aligned} \tag{9}$$

where $\mathbb{S} \subseteq \{1, 2, \ldots, N\}$. In this way, we established the relationship between the joint posterior distribution $p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$ and the individual posterior distribution $p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}_i)$. We adopt the same formulation in our inference model distribution as $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}^{\mathbb{S}}) \propto p_{\boldsymbol{\theta}}(\mathbf{z}) \prod_{i \in \mathbb{S}} \frac{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)}{p_{\boldsymbol{\theta}}(\mathbf{z})}$, using N individual approximate posterior distributions $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)$. In this work, we assume that $p_{\boldsymbol{\theta}}(\mathbf{z})$ and $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)$ are all following factorized Gaussian distributions. And each individual posterior $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)$ can be represented as:

$$q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i) = \prod_{j=1}^{M} q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)^{\mathbf{m}^q_{ij}} p_{\boldsymbol{\theta}}(\mathbf{z}_j)^{1-\mathbf{m}^q_{ij}} \tag{10}$$

where each $q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)$ is a multiplicative mixture between the approximated posterior $q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)$ and the prior $p_{\boldsymbol{\theta}}(\mathbf{z}_j)$, weighted by $\mathbf{m}^q_{ij}$. Since the quotient of two Gaussian distributions is also a Gaussian under well-defined conditions, we could parametrize the quotient $\frac{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)}{p_{\boldsymbol{\theta}}(\mathbf{z})}$ using a Gaussian distribution parametrized by $\tilde{q}_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)$. In this case

$$\begin{aligned} \frac{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)}{p_{\boldsymbol{\theta}}(\mathbf{z})} &= \prod_{j=1}^{M} \frac{q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)^{\mathbf{m}^q_{ij}} p_{\boldsymbol{\theta}}(\mathbf{z}_j)^{1-\mathbf{m}^q_{ij}}}{p_{\boldsymbol{\theta}}(\mathbf{z}_j)^{\mathbf{m}^q_{ij}+1-\mathbf{m}^q_{ij}}} \\ &= \prod_{j=1}^{M} \left(\frac{q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)}{p_{\boldsymbol{\theta}}(\mathbf{z}_j)} \right)^{\mathbf{m}^q_{ij}} = \prod_{j=1}^{M} \left(\tilde{q}_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)\right)^{\mathbf{m}^q_{ij}} \end{aligned} \tag{11}$$

where we use a inference network $\tilde{q}_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)$ to parametrize $q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)$ as $q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i) = \tilde{q}_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i) p_{\boldsymbol{\theta}}(\mathbf{z}_j)$. We show our full inference distribution $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ as:

$$q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}^{\mathbb{S}}) \propto \prod_{j=1}^{M} \left(p_{\boldsymbol{\theta}}(\mathbf{z}_j) \prod_{i \in \mathbb{S}} \left(\tilde{q}_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)\right)^{\mathbf{m}^q_{ij}} \right) \tag{12}$$

which is a weighted product-of-experts (Hinton, 2002) distribution for each latent variable Z_j. We include the detailed derivation in Appendix. A.1. The structure variable M^q_{ij} controls the weight of each multiplicative component $\tilde{q}_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)$ in the process of shaping the joint posterior distribution $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$. As a result of the Gaussian assumptions, the weighted product-of-experts distribution above has a closed-form solution. Suppose $p_{\boldsymbol{\theta}}(\mathbf{z}) \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \operatorname{diag}\left(\boldsymbol{\sigma}_0\right)\right)$, $\tilde{q}_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i) \sim \mathcal{N}\left(\boldsymbol{\mu}_i, \operatorname{diag}\left(\boldsymbol{\sigma}_i\right)\right)$ for $i = 1, 2, \ldots, N$. We introduce "dummy" variables in $\mathbf{m}^q$ that $\mathbf{m}^q_{0j} = 1$ for all j. Then we have

$$\begin{aligned} & q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}^{\mathbb{S}}) \sim \mathcal{N}\left(\boldsymbol{\mu}^q, \operatorname{diag}\left(\boldsymbol{\sigma}^q\right)\right) \\ & \frac{1}{\boldsymbol{\sigma}^q_j} = \sum_{i \in \mathbb{S} \cup \{0\}} \frac{\mathbf{m}^q_{ij}}{\boldsymbol{\sigma}_{ij}} \quad \boldsymbol{\mu}^q_j = \frac{1}{\boldsymbol{\sigma}^q_j} \sum_{i \in \mathbb{S} \cup \{0\}} \frac{\mathbf{m}^q_{ij}}{\boldsymbol{\sigma}_{ij}} \boldsymbol{\mu}_{ij} \,. \end{aligned} \tag{13}$$

With the derived inference model above, we are now able to model 2^N posterior inference distributions $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}^{\mathbb{S}})\ \forall \mathbb{S}$, coupled with $2^{N \times M}$ possible discrete structures M^q, with N inference networks $\tilde{q}_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i)$. Note that the introduced distribution $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ remains valid when we extend the value of structure variable M^q to continuous domain $\mathbb{R}^{N \times M}$, which paves the way to gradient-based structure learning.

Generation model We could parametrize our generation model $p_{\boldsymbol{\theta}}$ in a symmetric way using the weighted product-of-expert distributions using $p_{\boldsymbol{\theta}}(\mathbf{x}_i \mid \mathbf{z}_j)$ and M^p. In this work we adopt an alternative approach, due to the consideration that the Gaussian distribution assumption is inappropriate in complex raw data domain, like image pixels. We instead use M^p as a gating variable and parametrize $p_{\boldsymbol{\theta}}(\mathbf{x}_i \mid \mathbf{z}^{\mathbf{m}^p_i})$ in the form of $p_{\boldsymbol{\theta}}(\mathbf{x}_i \mid \mathbf{z}^{\mathbf{m}^p_i}) = p_{\boldsymbol{\theta}}(\mathbf{x}_i \mid \mathbf{z} \odot \mathbf{m}^p_i)$, where $\odot$ denotes element-wise multiplication. We can see that it's still tractable since the prior $p_{\boldsymbol{\theta}}(\mathbf{z})$ is known.

3.5. Tractable optimization

Structural regularization $\mathcal{L}_{\text{str\_reg}}$ Let's take a close look at the structural regularization term $\mathcal{L}_{\text{str\_reg}}$ in our training objective Eq. 8. As introduced in Sec.2.4, we have $\mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^k) = \sum_{\mathrm{v} \in \{\mathbf{x}, \mathbf{z}\}} \mathcal{I}_q(\mathrm{v}\ ;\ Pa^{\mathcal{G}^q}_{\mathrm{v}}) - \sum_{\mathrm{v} \in \{\mathbf{x}, \mathbf{z}\}} \mathcal{I}_q(\mathrm{v}\ ;\ Pa^{\mathcal{G}^k}_{\mathrm{v}})$. This objective poses new challenge to estimate and optimize mutual information. Note that any differentiable mutual information estimations and optimization methods can be applied here. In this paper, we propose to use tractable variational lower/upper-bounds of the intractable mutual information by re-using distributions $q_{\boldsymbol{\phi}}$ and $p_{\boldsymbol{\theta}}$. We refer to (Poole et al., 2019) for a detailed

Algorithm 1 Training with optional structure learning

Require: dataset $\mathcal{D} = \left\{\mathbf{x}^d\right\}_{d=1}^{|\mathcal{D}|}$
Require: parameters $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\rho^q}, \boldsymbol{\rho^p}$
Require: Bayesian Networks $\left\{\mathcal{G}^k \equiv \left(\mathcal{V}^k, \mathcal{E}^k\right)\right\}$
Require: hyper-parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$
Require: number of iterations to update distribution parameters $steps\_dist > 0$
Require: number of iterations to update structure parameters $steps\_str \geq 0$
Require: mini-batch size bs
Require: gradient-based optimizer opt
initialize all parameters $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\rho^q}, \boldsymbol{\rho^p}$
repeat
  for $step = 1$ **to** $steps\_dist$ **do**
    randomly sample a mini-batch $\mathcal{B}$ of size bs from dataset $\mathcal{D}$
    evaluate loss $\mathcal{L}_{\text{dist}}^{\mathcal{B}}$ using Eq. 8
    compute gradients $\nabla_{\boldsymbol{\phi}} \mathcal{L}_{\text{dist}}^{\mathcal{B}}, \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{dist}}^{\mathcal{B}}$
    $opt.optimize([\boldsymbol{\phi}, \boldsymbol{\theta}], [\nabla_{\boldsymbol{\phi}} \mathcal{L}_{\text{dist}}^{\mathcal{B}}, \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{dist}}^{\mathcal{B}}])$
  end for
  for $step = 1$ **to** $steps\_str$ **do**
    randomly sample a mini-batch $\mathcal{B}$ of size bs from dataset $\mathcal{D}$
    evaluate loss $\mathcal{L}_{\text{score}}^{\mathcal{B}}$ using Eq. 15
    compute gradients $\nabla_{\boldsymbol{\rho^q}} \mathcal{L}_{\text{score}}^{\mathcal{B}}, \nabla_{\boldsymbol{\rho^p}} \mathcal{L}_{\text{score}}^{\mathcal{B}}$
    $opt.optimize([\boldsymbol{\rho^q}, \boldsymbol{\rho^p}], [\nabla_{\boldsymbol{\rho^q}} \mathcal{L}_{\text{score}}^{\mathcal{B}}, \nabla_{\boldsymbol{\rho^p}} \mathcal{L}_{\text{score}}^{\mathcal{B}}])$
  end for
until converged

review and discussion of state-of-the-art tractable mutual information optimization methods.

Distribution consistency $\mathcal{L}_{\text{dist}}$ We aim to achieve the consistency between the joint distribution $q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z}, \mathbf{u})$ and $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}, \mathbf{u})$ through T cost functions in $\mathcal{L}_{\text{dist}}$. With the proposed inference model in Sec. 3.4, we could decompose our $\mathcal{L}_{\text{dist}}$ into two primary components: (i) *Enforcing* $q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z}, \mathbf{u}) = p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}, \mathbf{u})$ Many previous works(Kingma & Welling, 2014; Tolstikhin et al., 2018; Dumoulin et al., 2017; Donahue et al., 2017) have been proposed to learn a latent variable generative model to model the joint distribution, any tractable objective can be utilized here, we adopt the *ELBO* as the default choice. (ii) *Enforcing* $q_{\boldsymbol{\phi}}(\mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{z})$ The reason that we explicitly include this objective in $\mathcal{L}_{\text{dist}}$ is due to our $p_{\boldsymbol{\theta}}$-dependent parametrization of $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \propto p_{\boldsymbol{\theta}}(\mathbf{z}) \prod_{i=1}^{N} \frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i)}{p_{\boldsymbol{\theta}}(\mathbf{z})}$. Thus we explicitly enforce the consistency between the induced marginal distribution $q_{\boldsymbol{\phi}}(\mathbf{z}) \equiv \mathbb{E}_{q_{\boldsymbol{\phi}}} q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ and $p_{\boldsymbol{\theta}}(\mathbf{z})$. Tractable divergence estimators for minimizing $C_T\left(q_{\boldsymbol{\phi}}(\mathbf{z}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})\right)$

Table 1. Distribution consistency objectives $\mathcal{L}_{\text{dist}}$

C	$definition$
$C_0(\mathbf{x}, \mathbf{z}, \mathbf{u})$	$D_{\text{KL}}(q_{\boldsymbol{\phi}} \parallel p_{\boldsymbol{\theta}})$
$C_1(\mathbf{x}, \mathbf{u})$	$-\mathcal{L}_{\text{ELBO}}(q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{u}), p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{u}))$
$C_2(\mathbf{x})$	$D_{\text{JS}}(q_{\boldsymbol{\phi}}(\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x}))$
$C_3(\mathbf{z})$	$D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z}))$
$C_4(\mathbf{x}_i, \mathbf{z})$	$D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{x}_i, \mathbf{z}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{z}))$

have been proposed and analyzed in previous works,

$$\mathcal{L}_{\text{dist}} = \sum_{t=1}^{T-1} \alpha_t C_t(q_{\boldsymbol{\phi}} \parallel p_{\boldsymbol{\theta}}) + \alpha_T C_T\left(q_{\boldsymbol{\phi}}(\mathbf{z}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})\right) . \tag{14}$$

With the distribution consistency objective and the compositional inference model introduced in Sec. 3.4, we could train the latent variable generative model in a weakly/semi-supervised manner in terms of (i) incomplete data where X is partially observed (e.g. missing attributes in feature vectors, or missing a modality in multi-modal dataset), and (ii) partial known dependency structure in M^q and M^p.

Structure learning In this work, we show that our proposed framework is capable of learning the structure of Bayesian network $\mathcal{G}^q$ and $\mathcal{G}^p$ based on many existing structure learning methods efficiently, with *gradient-based* optimization techniques, which avoids searching over the discrete super-exponential space. Specifically, we show that our proposed framework can (i) represent the assumptions made about the structure of the true data distribution in the form of a set of structural regularization in the form of Bayesian networks $\{\mathcal{G}^k\}$ as the *explicit inductive bias*. A score-based structure learning objective is then introduced where $\mathcal{L}_{\text{str\_reg}}$ plays a vital role in scoring each candidate structure; and (ii) utilize the non-stationary data from multiple environments (Hyvärinen et al., 2019; Arjovsky et al., 2019; Ke et al., 2019) as additional observed random variables. We show the score-based structure learning objective as below

$$\min_{\mathbf{m}^q, \mathbf{m}^p} \mathcal{L}_{\text{score}} = \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{str\_reg}} + \mathcal{L}_{\text{sparsity}} \,. \tag{15}$$

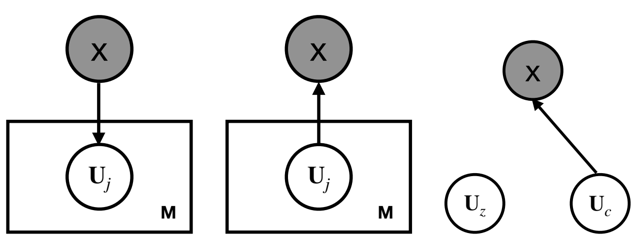
We assume a jointly factorized Bernoulli distribution prior for structure variable M^q and M^p, parametrized by $\boldsymbol{\rho^q}$ and $\boldsymbol{\rho^p}$. We use the gumbel-softmax trick proposed by (Jang et al., 2017; Maddison et al., 2017; Balog et al., 2017) as gradient estimators. Following the Bayesian Structural EM (Friedman, 1998; Elidan & Friedman, 2005) algorithm, we optimize the model alternatively between optimizing distributions $\mathcal{L}(q_{\boldsymbol{\phi}}, p_{\boldsymbol{\theta}})$ and structure variables $\mathcal{L}_{\text{score}}(\mathbf{m}^q, \mathbf{m}^p)$. We present the full algorithm to train the proposed generative model with optional structure learning procedure in Alg. 1.

Table 2. A unified view of {single/multi}-{modal/domain/view} models. C_i is referred to as the definition in Table. 1, $\mathcal{G}$ is referred to as the Bayesian networks in Figure. 1, 2. We use N to denote the number of views/domains/modals. We use ① to denote *shared/private latent space decomposition*, and use ② to denote *dependency structure learning*. Please see Appendix. A.1 for the full table.

MODELS	N	①	②	$\mathcal{G}^q$	$\mathcal{G}^p$	$\mathcal{L}_{\text{dist}}$	$\mathcal{L}_{\text{str\_reg}}$
VAE	1	×	×	$\left[\mathcal{G}^q_{\text{single}}\right]$	$\left[\mathcal{G}^p_{\text{single}}\right]$	$[1, C_1]$	$[]$
GAN	1	×	×	$[]$	$\mathcal{G}^p_{\text{single}}$	$[1, C_2]$	$[]$
INFOGAN	1	×	×	$[]$	$\mathcal{G}^p_{\text{single}}$	$[1, C_2]$	$[1, \mathcal{G}^{\text{InfoGAN}}]$
β-VAE	1	×	×	$\left[\mathcal{G}^q_{\text{single}}\right]$	$\left[\mathcal{G}^p_{\text{single}}\right]$	$[1, C_1], [\beta - 1, C_3]$	$[\beta - 1, \mathcal{G}^{\emptyset}]$
β-TCVAE	1	×	×	$\left[\mathcal{G}^q_{\text{single}}\right]$	$\left[\mathcal{G}^p_{\text{single}}\right]$	$[1, C_1], [\alpha_2, C_2]$	$[\beta, \mathcal{G}^p]$
JMVAE	2	×	×	$\left[\mathcal{G}^q_{\text{joint}}\right]$	$\left[\mathcal{G}^p_{\text{joint}}\right]$	$[1, C_1]$	$[\beta_i, \mathcal{G}^{\text{str}}_{\text{cross}}(\mathbf{x}_i)]$
MVAE	N	×	×	$\left[\mathcal{G}^q_{\text{joint}}, \mathcal{G}^q_{\text{marginal}}\right]$	$\left[\mathcal{G}^p_{\text{joint}}\right]$	$[1, C_1]$	$[\beta_i, \mathcal{G}^{\text{str}}_{\text{marginal}}(\mathbf{x}_i)]$
WYNER	2	✓	×	$\left[\mathcal{G}^q_{\text{joint}}, \mathcal{G}^q_{\text{marginal}}\right]$	$\left[\mathcal{G}^p_{\text{joint}}\right]$	$[1, C_1]$	$[\beta_i, \mathcal{G}^{\text{str}}_{\text{cross}}(\mathbf{x}_i)], [\beta_i, \mathcal{G}^{\text{str}}_{\text{private}}(\mathbf{x}_i)]$
OURS-MM	N	✓	✓	$[\mathcal{G}^q_{\text{full}}]$	$[\mathcal{G}^p_{\text{full}}]$	$[1, C_0]$	$[\beta_i, \mathcal{G}^{\text{str}}_{\text{cross}}(\{\mathbf{x}_i\})]$

4. Case study: Generative Data Modeling

In this section, we show various types of generative data modeling can be viewed as a structured latent space learning problem, which can be addressed by our proposed framework in a principled way.



$\mathcal{G}^q_{\text{single}}$ $\mathcal{G}^p_{\text{single}}$ $\mathcal{G}^{\text{str}}_{\text{info}}$

Figure 1. Bayesian networks for single-modal models

4.1. Single-modal generative model

Framework specification In single-modal data generative modeling setting, we have $N = 1$ observed variable $\mathrm{X} \equiv [\mathrm{X}_1]$ which could be in image, text or other modalities, and we only incorporate private latent variables U. We abuse the notation a little by assuming M latent variables $\mathrm{U} \equiv [\mathrm{U}_1, \mathrm{U}_2, \dots, \mathrm{U}_M]$[2].

A unified view We show that our proposed model unifies many existing generative models. We show that we can impose disentanglement as a special case of the structural regularization in latent space to obtain different existing disentangled representation learning methods. We summarize how existing generative models can be unified within our proposed information-theoretic framework in table 2. As an interesting example, we show that we can derive the β-vae objective with $\mathcal{L} = C_1 + (\beta - 1)C_3 + (\beta - 1)\mathcal{L}_{\text{str\_reg}}(\mathcal{G}^{\emptyset})$, where we impose the structural regularization $(\beta - 1)\mathbb{D}(q_{\boldsymbol{\phi}} \parallel \mathcal{G}^{\emptyset})$. In this way, we also established connections to the results in (Hoffman et al., 2017; Mathieu et al., 2019) that β-vae is optimizing *ELBO* with a $q_{\boldsymbol{\phi}}$-dependent implicit prior $r(\mathbf{u}) \propto q_{\boldsymbol{\phi}}(\mathbf{u})^{1-\beta} p_{\boldsymbol{\theta}}(\mathbf{u})^{\beta}$, we achieve this in a symmetric way by using a $p_{\boldsymbol{\theta}}$-dependent posterior $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \propto p_{\boldsymbol{\theta}}(\mathbf{z}) \prod_{i=1}^{N} \frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i)}{p_{\boldsymbol{\theta}}(\mathbf{z})}$. We further show that how we can unify other total-correlation based disentangled representation learning models (Chen et al., 2018; Esmaeili et al., 2019; Kim & Mnih, 2018) by explicitly imposing Bayesian structure $\mathcal{G}^p$ as structural regularization. We include detailed discussions and proofs in Appendix. A.2.

4.2. Multi-modal/domain/view generative model

Problem setup We represent the observed variables as $\mathrm{X}_{1:N} \equiv [\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_N]$, where we have N observed variables in different domains[3] and they might be statistically dependent. We thus aim to learn latent factors Z that explains the potential correlations among X. Meanwhile, we also learn latent factors U_i that explains the variations exclusive to one specific observed variable X_i. In this way, we could achieve explicit control over the domain-dependent and domain-invariant latent factors. For more details of the data generation process for this task and the model, please see Appendix. B.1.

[2]because we can define arbitrary dimension for U.

[3]We use the word domain to represent domain/modality/view.

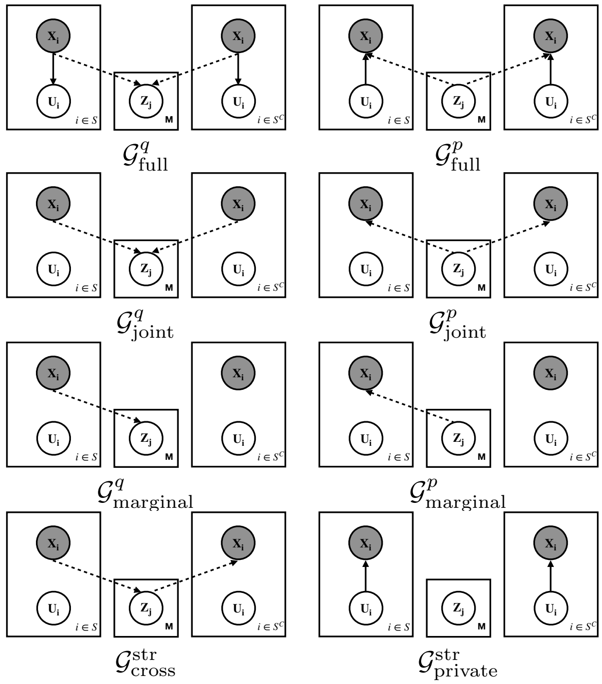
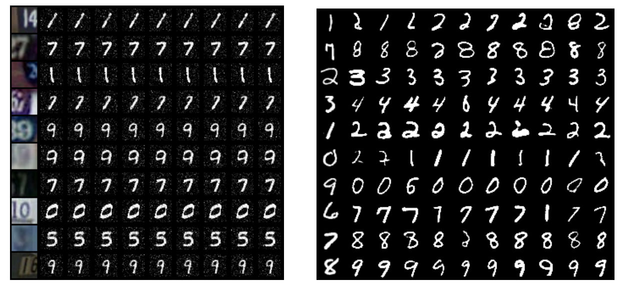


Figure 2. Bayesian networks of various inference, generation models and structural regularizations in multi-modal/domain/view setting.

A unified view We summarize the key results of unifying many existing multi-domain generative models in Table. 2. We prove and discuss some interesting connections to related works in more details in Appendix. A.3, including BiVCCA (Wang et al., 2016), JMVAE (Suzuki et al., 2017), TELBO (Vedantam et al., 2018), MVAE (Wu & Goodman, 2018), WynerVAE (Ryu et al., 2020), DIVA (Ilse et al., 2019) and CorEx (Steeg & Galstyan, 2014a;b; 2016; Gao et al., 2019).

Framework specification We present a specific implementation of our proposed framework for multi-domain generative modeling here. We show that it generalizes some heuristics used in previous models and demonstrate its effectiveness in several standard multi-modal datasets. We use $\mathcal{L}_{\text{dist}}$ in Table 1 to learn consistent inference model and joint, marginal, conditional generation model over $(\mathrm{X}, \mathrm{Z}, \mathrm{U})$. To embed multi-domain data into a shared latent space, we use the structural regularization that enforces Markov conditional independence structure $\mathrm{X}^{\mathbb{S}} \to \mathrm{Z} \to \mathrm{X}^{\mathbb{S}^{\mathsf{C}}}$. This structural regularization can be represented by $\mathcal{G}^{\text{str}}_{\text{cross}}$ in Figure 2, where $\mathrm{X} \equiv \left[\mathrm{X}^{\mathbb{S}}, \mathrm{X}^{\mathbb{S}^{\mathsf{C}}}\right]$ is a random bi-partition of X. Then we show that the objective can be upper-bounded by $\mathcal{L} = \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{str\_reg}} \leq \mathcal{L}_{\mathbf{x}} + \mathcal{L}_{\mathbf{u}} + \mathcal{L}_{\mathbf{z}}$, where $\mathcal{L}_{\mathbf{x}} = -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{u}|\mathbf{x})} \log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}, \mathbf{u})$, $\mathcal{L}_{\mathbf{u}} = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x})} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{u} \mid \mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{u}))$ and $\mathcal{L}_{\mathbf{z}} = \sum_{i=0}^{N} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x})} D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_i))$. We use $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}_0) \equiv p_{\boldsymbol{\theta}}(\mathbf{z})$ for the simplicity of notations. We further show that for each latent variable Z_j, $\mathcal{L}_{\mathbf{z}_j}$ term can be viewed as a generalized JS-divergence (Nielsen, 2019) among $q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_i)$ for $i \in \{1, \ldots, N\}$ using geometric-mean weighted by $\mathbf{m}^q_j$, which can be seen as a generalization of the implicit prior used in β-vae as discussed in 4.1. The detailed proof is presented in Appendix. A.3.

$$\mathcal{L}_{\mathbf{z}_j} = D^{\mathbf{m}^q_j}_{\text{JS}} \left(q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_0), q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_1), \ldots, q_{\boldsymbol{\phi}}(\mathbf{z}_j \mid \mathbf{x}_N)\right) \tag{16}$$



SVHN → MNIST MNIST → MNIST-Plus-1

Figure 3. Cross-domain generation samples. The leftmost column shows conditioned inputs.

Experiment We validate the effectiveness of proposed model in multi-view/modal data modeling setting on bi-modal MNIST-Label, MNIST-SVHN and bi-view MNIST-MNIST-Plus-1 dataset. We show the generated samples in Figure 3. The left panel in the figure contains the examples of MNIST-style samples generated by the model trained on MNIST-SVHN dataset when conditioned on SVHN data examples. We can observe that the model is using the shared latent variable Z and private latent variables U to successfully generate the MNIST-style samples of same digit as the SVHN inputs. On the other hand, the right panel contains the examples of MNIST-style sample generated by the model trained on MNIST-Plus-1 dataset by conditioning on MNIST example. We can observe that the model is successfully generating $m + 1$ digit images when conditioned on m digit input. More detailed results are included in the Appendix. B.1.

5. Case study: Fair Representation Learning

In this section, we show that fair representation learning can be viewed as a structured latent space learning problem, where we aim to learn a latent subspace that is invariant to sensitive attributes while informative about target label.

Problem setup We use $[\mathrm{X}, \mathrm{A}, \mathrm{Y}]$ to represent the observed variables, where the variable X represents the multivariate raw observation like pixels of image sample, the variable A represents the sensitive attributes, and the variable Y represents the target label to be predicted. Following the

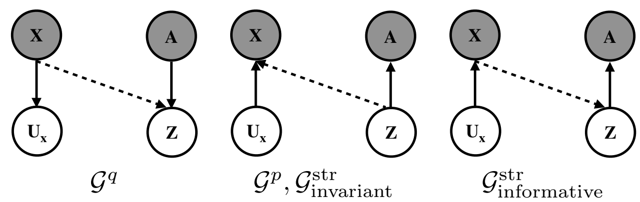


Figure 4. Bayesian networks for fair representation learning.

Table 3. Fair representation learning results on German and Adult datasets.

MODEL	ADULT		GERMAN	
	ACC	DEO	ACC	DEO
NAIVE SVM	0.80	0.09	0.74 ± 0.05	0.12 ± 0.05
SVM	0.79	0.08	0.74 ± 0.03	0.10 ± 0.06
NN	0.84	0.14	0.74 ± 0.04	0.47 ± 0.19
NN $+\chi^2$	0.83	0.03	0.73 ± 0.03	0.25 ± 0.14
FERM	0.77	0.01	0.73 ± 0.04	0.05 ± 0.03
OURS-MMD	0.83	0.02	0.72 ± 0.07	0.07 ± 0.09
OURS-TC	0.81	0.02	0.74 ± 0.08	0.08 ± 0.14
OURS-MINE	0.79	0.01	0.70 ± 0.11	0.05 ± 0.11

same setting in previous works (Song et al., 2019; Creager et al., 2019), the target label is not available during training phase. A linear classifier using learned representation is trained to predict the held-out label Y in testing time. We focus on the *Difference of Equal Opportunity* (DEO) notion in this work (Hardt et al., 2016). For the details of the data generation process, please see Appendix. A.4.

Framework specification we learn a joint distribution over $[\mathrm{X}, \mathrm{A}, \mathrm{Z}, \mathrm{U}]$ with the framework proposed. The shared latent variable Z aims to explain the hidden correlation among X and Z. We also enforce two structural regularizations, represented by two Bayesian networks $\mathcal{G}^{\text{str}}_{\text{invariant}}$ and $\mathcal{G}^{\text{str}}_{\text{informative}}$. The aim of $\mathcal{G}^{\text{str}}_{\text{invariant}}$ is to learn the private latent variables $\mathrm{U}_{\mathbf{x}}$ as the hidden factors that are invariant to the change of Z. Meanwhile, the aim of $\mathcal{G}^{\text{str}}_{\text{informative}}$ is to preserve as much information about X in Z. M^q and M^p are illustrated by $\mathcal{G}^q$ and $\mathcal{G}^p$ in Figure 4 correspondingly. Then we have the following learning objective

$$\mathcal{L} \leq -\mathbb{E}_{q_\phi} \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{a} \mid \mathbf{z}, \mathbf{u}) + \beta_2 \mathcal{I}_q(\mathbf{z}\ ;\ \mathbf{u}) + (1+\beta_1)\mathbb{E}_{q_\phi} D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}, \mathbf{a}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})) + const \tag{17}$$

Please refer to Appendix. A.4 for the detailed derivation and discussion.

Experiments We investigate the performance of our derived objective on the UCI German credit dataset and the UCI Adult dataset. For estimating and minimizing $\mathcal{I}_q(\mathbf{z}\ ;\ \mathbf{u})$, we adopted MMD (Gretton et al., 2006), total-correlation estimator in (Esmaeili et al., 2019) and MINE (Belghazi et al., 2018) and summarize all results in Table 3. We report the classification accuracy (ACC) and the aforementioned DEO in the table. The performances of all other baseline methods in the table are from (Mary et al., 2019; Donini et al., 2018). Please refer to Appendix. B.2 for more details.

6. Case study: Out-of-Distribution Generalization

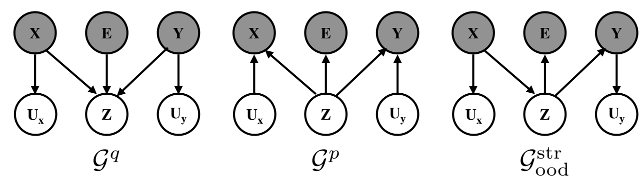


Figure 5. Bayesian networks for out-of-distribution generalization task. E in the diagram represents the index of the environmental factor, not the real value of E in the data generation process

Problem setup We show that discovering of true causation against spurious correlation through invariance can be viewed as a structured latent representation learning problem. Consider a set of environments $\mathcal{E}$ indexed by E, we have a data distribution $P^e(\mathrm{X}, \mathrm{Y})$ for each environment $\mathrm{E} = e$. We use $[\mathrm{X}, \mathrm{Y}, \mathrm{E}]$ to represent the observed variables, where X is data input, Y is label and E is the index of the corresponding environment index. The goal of this task is predict Y from X in a way that the performance of the predictor in the presence of the worst E is optimal. We derive an information-theoretic objective for out-of-distribution generalization task on Colored-MNIST dataset introduced in (Arjovsky et al., 2019). For more details of this experiment, please see the Appendix. B.3 as well as the original work (Peters et al., 2015; Arjovsky et al., 2019).

Framework specification As our structural regularization, we use the Bayesian network $\mathcal{G}^{\text{str}}_{\text{ood}}$ in Figure 5. The purpose of $\mathcal{G}^{\text{str}}_{\text{ood}}$ is to enforce that Z is sufficient statistic in making the prediction of Y and that $E \perp Y|Z$. We present the derived learning objective here

$$\mathcal{L}_{\text{info}} = \mathcal{L}_{\text{dist}} + \beta_1 D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}, \mathbf{e}, \mathbf{y}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})) + \beta_2 \mathcal{I}_q(\mathbf{x}, \mathbf{e}, \mathbf{y} \mid \mathbf{z}) \tag{18}$$

We further show that the idea in (Arjovsky et al., 2019) can be directly integrated into our proposed framework by imposing stable M^p structure as constraints across environments, measured by gradient-penalty, as discussed in Appendix. A.5.

Experiments We validate the proposed model on the Colored-MNIST classification task introduced in (Arjovsky et al., 2019). We also take the advantage of our proposed framework as a generative model that we could perform semi-supervised learning, where we use only 50% labeled data. We include more training setting details in Appendix. B.3. We compare our model against the baselines in

Table 4. Out-of-distribution generalization results on Colored-MNIST

MODEL	ACC. TRAIN ENVS.	ACC. TEST ENV.
RANDOM	50	50
OPTIMAL	75	75
ORACLE	73.5 ± 0.2	73.0 ± 0.4
ERM	87.4 ± 0.2	17.1 ± 0.6
IRM	70.8 ± 0.9	66.9 ± 2.5
OURS-FULL	67.8 ± 6.8	62.1 ± 6.1
OURS-SEMI	71.4 ± 6.1	58.7 ± 7.2

Table 4. We see that our proposed information-theoretic objective achieves comparable performance in both supervised and semi-supervised setting on the test-environment.

7. Conclusion

In this work, we propose a general information-theoretic framework for learning structured latent factors from multivariate data, by generalizing the multivariate information bottleneck theory. We show that the proposed framework can provide an unified view of many existing methods and insights on new models for many different challenging tasks like fair representation learning and out-of-distribution generalization.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019. URL http://arxiv.org/abs/1907.02893.

Balog, M., Tripuraneni, N., Ghahramani, Z., and Weller, A. Lost relatives of the gumbel trick. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 371–379. PMLR, 2017. URL http://proceedings.mlr.press/v70/balog17a.html.

Belghazi, I., Rajeswar, S., Baratin, A., Hjelm, R. D., and Courville, A. C. MINE: mutual information neural estimation. *CoRR*, abs/1801.04062, 2018. URL http://arxiv.org/abs/1801.04062.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pp. 2615–2625, 2018. URL https://bit.ly/37WQG7X.

Creager, E., Madras, D., Jacobsen, J., Weis, M. A., Swersky, K., Pitassi, T., and Zemel, R. S. Flexibly fair representation learning by disentanglement. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1436–1445. PMLR, 2019. URL http://proceedings.mlr.press/v97/creager19a.html.

Donahue, J., Krähenbühl, P., and Darrell, T. Adversarial feature learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BJtNZAFgg.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. Empirical risk minimization under fairness constraints. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pp. 2796–2806, 2018.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A. C. Adversarially learned inference. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1ElR4cgg.

Elidan, G. and Friedman, N. Learning hidden variable networks: The information bottleneck approach. *J. Mach. Learn. Res.*, 6:81–127, 2005. URL http://jmlr.org/papers/v6/elidan05a.html.

Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D. H., Dy, J. G., and van de Meent, J. Structured disentangled representations. In Chaudhuri, K. and Sugiyama, M. (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2525–2534. PMLR, 2019. URL http://proceedings.mlr.press/v89/esmaeili19a.html.

Friedman, N. The bayesian structural EM algorithm. In Cooper, G. F. and Moral, S. (eds.), *UAI '98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, University of Wisconsin Business School,* *Madison, Wisconsin, USA, July 24-26, 1998*, pp. 129–138. Morgan Kaufmann, 1998. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=241&proceeding_id=14.

Friedman, N., Mosenzon, O., Slonim, N., and Tishby, N. Multivariate information bottleneck. In Breese, J. S. and Koller, D. (eds.), *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, pp. 152–161. Morgan Kaufmann, 2001. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=95&proceeding_id=17.

Gao, S., Brekelmans, R., Steeg, G. V., and Galstyan, A. Auto-encoding total correlation explanation. In Chaudhuri, K. and Sugiyama, M. (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1157–1166. PMLR, 2019. URL http://proceedings.mlr.press/v89/gao19a.html.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In Schölkopf, B., Platt, J. C., and Hofmann, T. (eds.), *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pp. 513–520. MIT Press, 2006. URL https://bit.ly/2vIOzpV.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3315–3323, 2016.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *4th International Conference on Learning Representations, ICLR 2016*, 2016. URL https://openreview.net/forum?id=Sy2fzU9gl.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018. URL https://doi.org/10.1162/089976602760128018.

Hoffman, M. D., Riquelme, C., and Johnson, M. J. The β-vae' s implicit prior. In *Workshop on Bayesian Deep Learning, NIPS*, pp. 1–5, 2017.

Hyvärinen, A., Sasaki, H., and Turner, R. E. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In Chaudhuri, K. and Sugiyama, M. (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 2019. URL http://proceedings.mlr.press/v89/hyvarinen19a.html.

Ilse, M., Tomczak, J. M., Louizos, C., and Welling, M. DIVA: domain invariant variational autoencoder. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SkgkEL8FdV.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Pal, C., and Bengio, Y. Learning neural causal models from unknown interventions. *CoRR*, abs/1910.01075, 2019. URL http://arxiv.org/abs/1910.01075.

Kim, H. and Mnih, A. Disentangling by factorising. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2654–2663. PMLR, 2018. URL http://proceedings.mlr.press/v80/kim18b.html.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Kingma, D. P. and Welling, M. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Lin, J. Divergence measures based on the shannon entropy. *IEEE Trans. Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115. URL https://doi.org/10.1109/18.61115.

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=S1jE5L5gl.

Mary, J., Calauzènes, C., and Karoui, N. E. Fairness-aware learning for continuous attributes and treatments. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4382–4391. PMLR, 2019. URL http://proceedings.mlr.press/v97/mary19a.html.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. Disentangling disentanglement in variational autoencoders. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4402–4412. PMLR, 2019. URL http://proceedings.mlr.press/v97/mathieu19a.html.

Nielsen, F. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 78(A):1–42, 2015. ISSN 1369-7412.

Poole, B., Ozair, S., van den Oord, A., Alemi, A., and Tucker, G. On variational bounds of mutual information. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 2019. URL http://proceedings.mlr.press/v97/poole19a.html.

Ryu, J. J., Choi, Y., Kim, Y.-H., El-Khamy, M., and Lee, J. Wyner {vae}: A variational autoencoder with succinct common representation learning. *OpenReview*, 2020. URL https://openreview.net/forum?id=r1g1CAEKDH.

Slonim, N., Friedman, N., and Tishby, N. Multivariate information bottleneck. *Neural Computation*, 18:1739–1789, 2006.

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. Learning controllable fair representations. In Chaudhuri, K. and Sugiyama, M. (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2164–2173. PMLR, 2019. URL http://proceedings.mlr.press/v89/song19a.html.

Steeg, G. V. and Galstyan, A. Discovering structure in high-dimensional data through correlation explanation. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 577–585, 2014a. URL https://bit.ly/374PJJC.

Steeg, G. V. and Galstyan, A. Maximally informative hierarchical representations of high-dimensional data. *CoRR*, abs/1410.7404, 2014b. URL http://arxiv.org/abs/1410.7404.

Steeg, G. V. and Galstyan, A. The information sieve. In Balcan, M. and Weinberger, K. Q. (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 164–172. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/steeg16.html.

Suzuki, M., Nakayama, K., and Matsuo, Y. Joint multimodal learning with deep generative models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BkL7bONFe.

Tishby, N., Pereira, F. C. N., and Bialek, W. The information bottleneck method. *CoRR*, physics/0004057, 2000. URL http://arxiv.org/abs/physics/0004057.

Tolstikhin, I. O., Bousquet, O., Gelly, S., and Schölkopf, B. Wasserstein auto-encoders. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=HkL7n1-0b.

Vedantam, R., Fischer, I., Huang, J., and Murphy, K. Generative models of visually grounded imagination. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=HkCsm6lRb.

Wang, W., Lee, H., and Livescu, K. Deep variational canonical correlation analysis. *CoRR*, abs/1610.03454, 2016. URL http://arxiv.org/abs/1610.03454.

Wu, M. and Goodman, N. Multimodal generative models for scalable weakly-supervised learning. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pp. 5580–5590, 2018. URL https://bit.ly/31sV3W0.

Zhao, S., Song, J., and Ermon, S. A lagrangian perspective on latent variable generative models. In Globerson, A. and Silva, R. (eds.), *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pp. 1031–1041. AUAI Press, 2018. URL http://auai.org/uai2018/proceedings/papers/361.pdf.