MIXTURE REPRESENTATION LEARNING WITH COUPLED AUTOENCODING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Jointly identifying a mixture of discrete and continuous factors of variability can help unravel complex phenomena. We study this problem by proposing an unsupervised framework called coupled mixture VAE (cpl-mixVAE), which utilizes multiple interacting autoencoding agents. The individual agents operate on augmented copies of training samples to learn mixture representations, while being encouraged to reach consensus on the categorical assignments. We provide theoretical justification to motivate the use of a multi-agent framework, and formulate it as a variational inference problem. We benchmark our approach on MNIST and dSprites, achieving state-of-the-art categorical assignments while preserving interpretability of the continuous factors. We then demonstrate the utility of this approach in jointly identifying cell types and type-specific, activity-regulated genes for a single-cell gene expression dataset profiling over 100 cortical neuron types.

1 INTRODUCTION

Complex phenomena can be attributed to a mixture of discrete and continuous factors of variability. Such complexity is crucial to understand in a variety of different contexts, from learning models for image datasets to identifying factors underlying neuronal identity. A common approach to study these phenomena is clustering, which can produce representations that jointly capture the dependence on discrete and continuous factors. Generative models can learn such representations, which has recently received attention from the deep learning community. Deep Gaussian mixture models are among the first deep generative models to jointly represent discrete and continuous factors, in which a continuous representation is decomposed into discrete clusters (Johnson et al., 2016; Dilokthanakul et al., 2016; Jiang et al., 2017). However, such models have mainly focused on clustering without regard to interpretability. Adversarial and variational methods have been proposed to learn mixture representations that can identify interpretable continuous factors. While adversarial learning, e.g. InfoGAN (Chen et al., 2016) is susceptible to stability issues (Kim & Mnih, 2018; Dupont, 2018; Jeong & Song, 2019), variational approaches, e.g. JointVAE and CascadeVAE have produced promising and more stable results (Dupont, 2018; Jeong & Song, 2019). However, such variational methods utilizing a single autoencoding agent rely either on a heuristic data-dependent embedding capacity, or on solving a separate optimization problem for the discrete variable. Thus, learning interpretable and stable mixture representations remains challenging.

We introduce a multi-agent variational framework to jointly infer discrete and continuous factors through collective decision making, while sidestepping heuristic approaches used by single-agent frameworks. Coupling of autoencoding agents has been previously studied in the context of multimodal recordings, where each agent learns a continuous latent representation for one of the data modalities (Feng et al., 2014; Gala et al., 2019). Here, we propose pairwise-coupled autoencoders to learn a mixture representation for a single data modality in an unsupervised fashion. Each autoencoder agent receives an augmented copy of the given sample with the same class label. To achieve this, we design a novel type-preserving augmentation that generates noisy copies of the data using within-class variabilities, while preserving its class identity. Coupling across the agents is achieved by encouraging categorical variables to be invariant under the augmentation, which regularizes the agents to learn interpretable representations. We demonstrate that such a coupled multi-agent architecture can increase inference accuracy and robustness by exploiting within-cluster variabilities, without requiring a prior distribution on the relative abundances of categories.

Our contributions can be summarized as follows: (i) We first provide theoretical justification to motivate the advantage of collective decision making for more accurate categorical assignments,

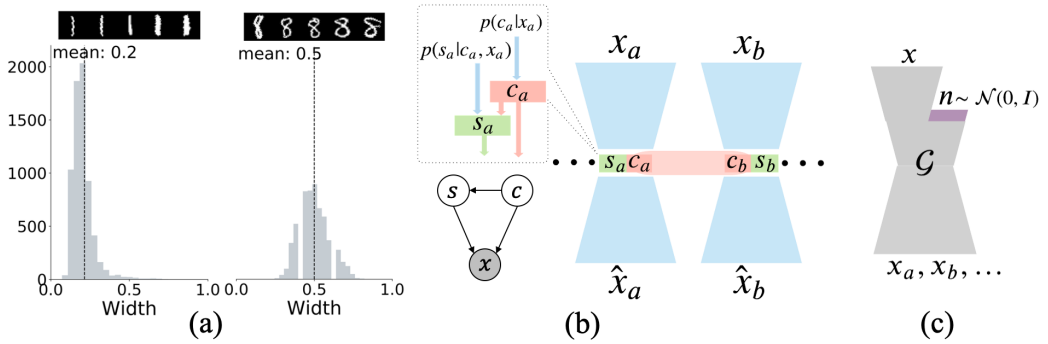

Figure 1: (a) Empirical distributions for the continuous state variable representing stroke width are digit-dependent, illustrating dependence of style on type. (b) Each agent in the multi-agent cpl-mixVAE model captures the type dependence of the state variable according to the graphical model. Agents are coupled by imposing a penalty on mismatches in the categorical assignments $\mathbf{c}_a$ and $\mathbf{c}_b$ for augmented copies $\mathbf{x}_a$ and $\mathbf{x}_b$ of a given sample $\mathbf{x}$. (c) Type-preserving augmentation network $\mathcal{G}$ generates augmented samples by first projecting $\mathbf{x}$ to a lower dimension, then concatenating it with Gaussian noise $\mathbf{n}$, and finally projecting it back to the data space. This network is trained with an adversarial loss to ensure that type is conserved for all generated samples.

utilizing noisy copies of the same sample. To obtain such samples, we propose an unsupervised type-preserving augmentation method. (ii) We formulate collective decision making as a variational inference problem with multiple agents. In this formulation, we introduce an approximation of Aitchison distance in the simplex to compare categorical assignments of the agents, which avoids mode collapse. (iii) We benchmark our method and display its superiority over comparable approaches using the MNIST and dSprites datasets. (iv) Finally, we apply the method to a challenging single cell gene expression dataset for a population of neurons. We demonstrate that our method can be used to discover discrete categories referred to as neuronal types and type-specific genes regulating the continuous within-type variability (e.g., metabolic states, disease states).

2 Preliminaries

For an observation $\mathbf{x} \in \mathbb{R}^D$, a variational autoencoder (VAE) learns a generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ and a variational distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, where $\mathbf{z} \in \mathbb{R}^M$ for $M \ll D$ is a latent variable with a parameterized distribution $p(\mathbf{z})$ (Kingma & Welling, 2013). *Disentangling* different sources of variability into different dimensions of $\mathbf{z}$ enables an interpretable selection of latent factors (Higgins et al., 2017; Locatello et al., 2018). However, in many real-world applications the inherent mixture distribution of continuous and discrete variations is often overlooked. This problem can be addressed within the VAE framework in an unsupervised fashion by introducing a categorical latent variable $\mathbf{c} \in \mathcal{S}^K$, denoting the class label defined in a K-simplex, alongside the continuous latent variable $\mathbf{s} \in \mathbb{R}^M$. Here, we refer to the continuous variable $\mathbf{s}$ as the *state* or *style* variable interchangeably. Assuming $\mathbf{s}$ and $\mathbf{c}$ are independent random variables, the evidence lower bound (ELBO) (Blei et al., 2017) for a single autoencoding agent with the distributions parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is given by,

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{s},\mathbf{c}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s},\mathbf{c})\right] - D_{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x}) \| p(\mathbf{s})\right) - D_{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{c}|\mathbf{x}) \| p(\mathbf{c})\right). \quad (1)$$

Maximizing ELBO in Eq. 1 to jointly learn $q(\mathbf{s}|\mathbf{x})$ and $q(\mathbf{c}|\mathbf{x})$ is challenging due to the mode collapse problem, where the network ignores a subset of latent variables. Akin to β-VAE (Higgins et al., 2017; Burgess et al., 2018), JointVAE assigns controlled information capacities to both continuous and categorical factors to prevent mode collapse (Dupont, 2018). A drawback of this method is that the capacities are dataset-dependent, and need to be tuned empirically over training iterations. As an alternative, CascadeVAE (Jeong & Song, 2019) maximizes the ELBO by iterating over two separate optimizations for the continuous and categorical variables after a warm-up period, instead of a fully gradient-based optimization. Although the computational cost for the suggested optimization for the categorical variable has an approximately linear dependence on the number of categories and batch size, it can still be a deterrent for problems with numerous categories and unbalanced datasets requiring larger batch sizes. Thus, single-agent VAEs fall short of efficiently learning interpretable mixture representations, either due to their reliance on a heuristic embedding capacity, or lacking a fully variational approach.

In addition to the issues discussed above, the performance and interpretability of those VAE approaches are further limited by the common assumption that the continuous variable representing

the style of the data is independent of the class label. In practice, style often depends on class label. For instance, even for the well-studied MNIST dataset, the histograms of common digit styles, e.g. "width", markedly vary for different digits (Fig. 1a). Moreover, in the discussed approaches, e.g. JointVAE, unsupervised clustering of only the continuous variable achieves a relatively high classification accuracy ($\sim 66\%$, see supplementary F and G), underscoring that the independence assumption is not valid.

3 Coupled mixture VAE framework

The key intuition behind multi-agent networks is cooperation for decision making to improve probabilistic estimation. Collective decision making has been studied under different contexts and a popular name referring to its advantages is the "wisdom of the crowd" (Surowiecki, 2005). When unanimous decisions made by a crowd (multiple agents) form a probability distribution, multiple estimates from the same sample increase the expected probability of true assignment. This is theoretically justified by Proposition 1 in the context of categorical decision making.

Definition 1. (Confidence) *Suppose the sample* $\mathbf{x}$ *is generated by a multivariate mixture distribution, so that* $p(\mathbf{x}) = \sum_{\phi} p(\phi)p(\mathbf{x}|\phi)$*, where* $p(\phi)$ *denotes an arbitrary prior for categories abundance.*

$$\mathcal{C}(k) = \mathbb{E}_{p(\mathbf{x})}\left[\log p(\phi = k|\mathbf{x})\right]. \tag{2}$$

Proposition 1. *Consider the problem of mixture representation learning in a multi-agent framework, where the accuracy of categorical assignment for a single agent is imperfect. There exists an* A*, number of the agents such that the confidence of correct categorical assignment for* A*-agent is higher than that of one agent.* (Proof in supplementary Section A)

While in an unsupervised framework, defining required number of agents in the absence of the categorical prior and category-dependent information remains as a challenge, in the particular case of uniform prior distribution over categories, we have the following Corollary.

Corollary 1. *For a uniform prior on the discrete factors, one pair of coupled agents (*$A = 2$*) is sufficient to increase the confidence of correct categorical assignment.* (see supplementary Section A)

3.1 Multi-agent VAE

Using the insight obtained from Proposition 1, we formulate collective decision making for an A-agent VAE network (Fig. 1b) as the following constrained optimization.

$$\begin{aligned} \max \quad & \mathcal{L}_{\mathbf{s}_1|\mathbf{c}_1}(\boldsymbol{\phi}_1, \boldsymbol{\theta}_1) + \cdots + \mathcal{L}_{\mathbf{s}_A|\mathbf{c}_A}(\boldsymbol{\phi}_A, \boldsymbol{\theta}_A) \\ & \text{s.t. } \mathbf{c}_1 = \cdots = \mathbf{c}_A \end{aligned} \tag{3}$$

Here, $\mathcal{L}_{\mathbf{s}_a|\mathbf{c}_a}(\boldsymbol{\phi}_a, \boldsymbol{\theta}_a)$ is the variational loss for agent a as follows,

$$\begin{aligned} \mathcal{L}_{\mathbf{s}_a|\mathbf{c}_a}(\boldsymbol{\phi}_a, \boldsymbol{\theta}_a) \quad = \quad & \mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x})}\left[\log p(\mathbf{x}_a|\mathbf{s}_a, \mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right] \\ & -\mathbb{E}_{q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{c}_a|\mathbf{x}_a)\|p(\mathbf{c}_a)\right)\right]. \end{aligned} \tag{4}$$

In Eq. (4), for each agent, we use the graphical model in Fig. 1b and modify the loss function in Eq. (1) by conditioning state on the categorical variable (derivation in supplementary Section B). Not only is it challenging to solve the maximization in Eq. 3 due to the equality constraint, but the objective remains a function of the prior $p(\mathbf{c})$ which is unknown, and typically non-uniform. To overcome this, we introduce an equivalent formulation for Eq. 3 based on the pairwise coupling paradigm as follows (derivation in supplementary Section C).

$$\begin{aligned} \max \quad & \sum_{a=1}^{A} (A-1)\left(\mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a, \mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right]\right) \\ & -\sum_{a<b} \mathbb{E}_{q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)}\mathbb{E}_{q(\mathbf{s}_b|\mathbf{c}_b,\mathbf{x}_b)}\left[D_{KL}\left(q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b)\|p(\mathbf{c}_a, \mathbf{c}_b)\right)\right] \\ & \text{s.t. } \mathbf{c}_a = \mathbf{c}_b \;\; \forall a, b \in [1, A],\; a < b \end{aligned} \tag{5}$$

Here, in the last term, the KL divergence across coupled agents is a function of the joint distribution $p(\mathbf{c}_a, \mathbf{c}_b)$, rather than $p(\mathbf{c})$. We relax Eq. 5 into an unconstrained problem by assuming a differentiable form for $p(\mathbf{c}_a, \mathbf{c}_b)$ (full derivation in supplementary Section D).

$$\begin{aligned} \max \quad & \sum_{a=1}^{A} (A-1)\left(\mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a, \mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a, \mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right]\right) \\ & +\sum_{a<b} H(\mathbf{c}_a|\mathbf{x}_a) + H(\mathbf{c}_b|\mathbf{x}_b) - \lambda\mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\mathbb{E}_{q(\mathbf{c}_b|\mathbf{x}_b)}\left[d^2(\mathbf{c}_a, \mathbf{c}_b)\right] \end{aligned} \tag{6}$$

According to the final expression in Eq. 6, the agents try to achieve identical categorical assignments while independently learning their own style variables. For each pair of agents, there are two entropy based confidence penalty terms, which are mode collapse regularizers (Pereyra et al., 2017). There is also distance $d(\mathbf{c}_a, \mathbf{c}_b)$ between a pair of categorical variables, which encourages the consensus on the categorical assignment controlled by *coupling hyperparameter* λ. The distance is defined as $d(\mathbf{c}_a, \mathbf{c}_b) = \|clr(\mathbf{c}_a) - clr(\mathbf{c}_b)\|_2, \forall \mathbf{c}_a, \mathbf{c}_b \in \mathcal{S}^K$, where $clr(.)$ denotes the isometric centered-log-ratio transformation and therefore d satisfies the conditions of a mathematical metric according to Aitchison geometry (Aitchison, 1982; Egozcue et al., 2003). To sample from $q(\mathbf{c}_a|\mathbf{x}_a)$ in a gradient descent framework, we use the Gumbel-softmax distribution (Jang et al., 2016; Maddison et al., 2014) with a temperature parameter $0 < \tau \leq 1$.

In the rest of this study, we refer to the model in Eq. 6 as *cpl-mixVAE* (Fig. 1b). Note that this formulation can be easily extended to include an additional hyperparameter to encourage disentanglement of continuous variables as in β-VAE Higgins et al. (2017). To train this model in an unsupervised fashion according to Eq. 6, we require augmented samples $\mathbf{x_a}$ for any given sample $\mathbf{x}$.

3.2 Type-preserving augmentation

Augmentation can be considered as a generative process (Antoniou et al., 2017). We seek a generative model that not only learns the data distribution, but also transformations that represent within-class variations in an unsupervised manner. Learning such transformations is generally not straightforward, and requires prior knowledge about the underlying invariances. While conventional transformations such as rotation, scaling, or translation can serve as type-preserving augmentations for many image datasets, they may not capture the richness of the underlying process. Moreover, such augmentation strategies cannot be used when within-class invariance are unknown. Suggested alternatives to conventional augmentations either rely on class label, or are specific to image data (Hauberg et al., 2016; Jaiswal et al., 2018; Antoniou et al., 2017).

To this end, inspired by DAGAN (Antoniou et al., 2017), we propose an unsupervised type-preserving augmentation using a VAE-GAN (Larsen et al., 2016)-like architecture, Fig. 1.c. We seek a network $\mathcal{G}$ such that a noisy copy, $\mathbf{x_a}$ can be obtained as a variation of the given sample, $\mathbf{x}$, based on its low dimensional representation that is concatenated with Gaussian noise $\mathbf{n}$. To prevent the network from disregarding the noise, we formulate the training procedure as the following minmax optimization which uses a discriminator network $\mathcal{D}$ as a regularizer.

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathcal{V}(\mathcal{D}, \mathcal{G}) - \mathcal{R}(\mathcal{G}) + \mathcal{T}_\alpha(\mathcal{G}) + \gamma d(\mathcal{G}) \tag{7}$$

While training, $\mathcal{G}$ generates two samples: $\mathbf{x_n}$ and $\mathbf{x_{\not n}}$. The former denotes $\mathbf{x_a}$, and the latter is a sample generated in the absence of noise. In Eq. 7, $\mathcal{V}(\mathcal{D}, \mathcal{G}) = \mathbb{E}_{\mathbf{x}}[\log \mathcal{D}(\mathbf{x})] + \mathbb{E}_{\mathbf{x}}[\log(1 - \mathcal{D}(\mathbf{x_{\not n}}))] + \mathbb{E}_{\mathbf{x},\mathbf{n}}[\log(1 - \mathcal{D}(\mathbf{x_n}))]$ is the value function for the joint training of the discriminator and generator; $\mathcal{R}(\mathcal{G}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})]$ is the reconstruction loss, which operates only over $\hat{\mathbf{x}}$; $\mathcal{T}_\alpha(\mathcal{G}) = \max(\|\mathbf{x} - \mathbf{x_{\not n}}\|_2 - \|\mathbf{x} - \mathbf{x_n}\|_2 + \alpha, 0)$ is the triplet loss that prevents network $\mathcal{G}$ from disregarding noise and generating identical samples; and $d(\mathcal{G}) = D_{KL}(q(\mathbf{z}|\mathbf{x}) \| q(\mathbf{z}|\mathbf{x}, \mathbf{n}))$ is the distance between the latent variables in the absence and presence of noise. $d(\mathcal{G})$ is a regularizer to encourage original and noisy samples to be located close to one another in the latent space and is controlled by hyperparameter $\gamma \ll 1$.

3.3 Stabilizing the training by mini-batch variance

The solution to the maximization problem in Eq. 6, which includes minimization of $d(\mathbf{c}_a, \mathbf{c}_b)$, has trivial local optima that result in the mode collapse issue (Lucas et al., 2019), in which during learning the network ignores a subset of the discrete latent space. For instance, in the extreme case, the network learns $\mathbf{c}_{a_n} = \mathbf{c}_{b_n} = \mathbf{c}$, $\forall\, n$, where n denotes the sample index. In this scenario, the state variable is compelled to act as the latent variable of a classical variational autoencoder, while the model fails to deliver an interpretable mixture representation despite achieving an overall low loss value. We regularize the Aitchison distance $d(\mathbf{c}_{a_n}, \mathbf{c}_{b_n})$,between the categorical assignments of the n-th samples of agents a and b by using mini-batch statistics to avoid mode collapse: $d^2_\sigma(\mathbf{c}_{a_n}, \mathbf{c}_{b_n}) = \sum_k \left(\sigma^{-1}_{a_k} \log c_{a_{n_k}} - \sigma^{-1}_{b_k} \log c_{b_{n_k}}\right)^2$, where $\sigma^2_{a_k}$ indicates the variance of the k-th category of agent a. In the following proposition we show that d^2_σ is an approximation of the Aitchison distance in the probability simplex.

Proposition 2. *Suppose* $\mathbf{c}_{a_n}, \mathbf{c}_{b_n} \in \mathcal{S}^K$*, where* $\mathcal{S}^K$ *is a simplex of* K *parts and* n *is the sample index. If* $d\left(\mathbf{c}_{a_n}, \mathbf{c}_{b_n}\right)$ *denotes the Aitchison distance, then*

$$d_\sigma^2\left(\mathbf{c}_{a_n}, \mathbf{c}_{b_n}\right) - d^2\left(\mathbf{c}_{a_n}, \mathbf{c}_{b_n}\right) \leq \frac{1}{K}\left(\left(\tau_{\mathbf{c}} + \tau_\sigma\right)^2 + K^2\tau_\sigma^2 - \Delta_\sigma^2\right)$$

where $\tau_{\mathbf{c}} = \max_k\{\log c_{a_{n_k}} - \log c_{b_{n_k}}\}$, $\tau_\sigma = \max_k\{(\sigma_{a_k}^{-1} - 1)\log c_{a_{n_k}} - (\sigma_{b_k}^{-1} - 1)\log c_{b_{n_k}}\}$*, and* $\Delta_\sigma = \sum_k (\sigma_{a_k}^{-1} - 1)\log c_{a_{n_k}} - (\sigma_{b_k}^{-1} - 1)\log c_{b_{n_k}}$. (Proof in supplementary Section E)

Accordingly, as the Gumbel-softmax approximations of the categorical variable of the agents move closer to each other on the simplex, d_σ converges to d.

4 EXPERIMENTS

We assess the performance of cpl-mixVAE for three different datasets. To facilitate comparisons with other methods, first we conducted experiments on two benchmark datasets: MNIST and dSprites. Additionally, we used a single cell RNA-sequencing dataset (scRNA-seq) (Tasic et al., 2018), to evaluate the utility of our approach in identifying neuronal cell types and type-specific genes regulating the continuous within-type variability.

MNIST: According to the approximately uniform distribution of handwritten digits in the dataset, we used a 2-agent cpl-mixVAE with shared categorical variable to learn an interpretable representation. Each agent learned a mixture generative model including a 10-dimensional categorical variable representing digits (type), and a 10-dimensional continuous random variable representing the writing style (state). To generate noisy samples, the augmenter was trained on the MNIST dataset ahead of training the cpl-mixVAE model. Fig. 2a displays example noisy samples generated by the type-preserving augmentation for MNIST. During training of cpl-mixVAE, each agent received an augmented copy of the original image. To interpret the role of the continuous factor, we fix the discrete latent variable and change the state variable according to the conditional state distribution learned for each category. Fig. 2(b-e) illustrates these continuous latent traversal results for four dimensions of the state variable obtained cpl-mixVAE. Each row corresponds to a different dimension of the categorical variable, and the state variable monotonically changes across the columns. Panels (b), (c), and (d) represent commonly-identified continuous factors with global attributes, while panel (e) represents roundness, all in a digit-dependent manner.

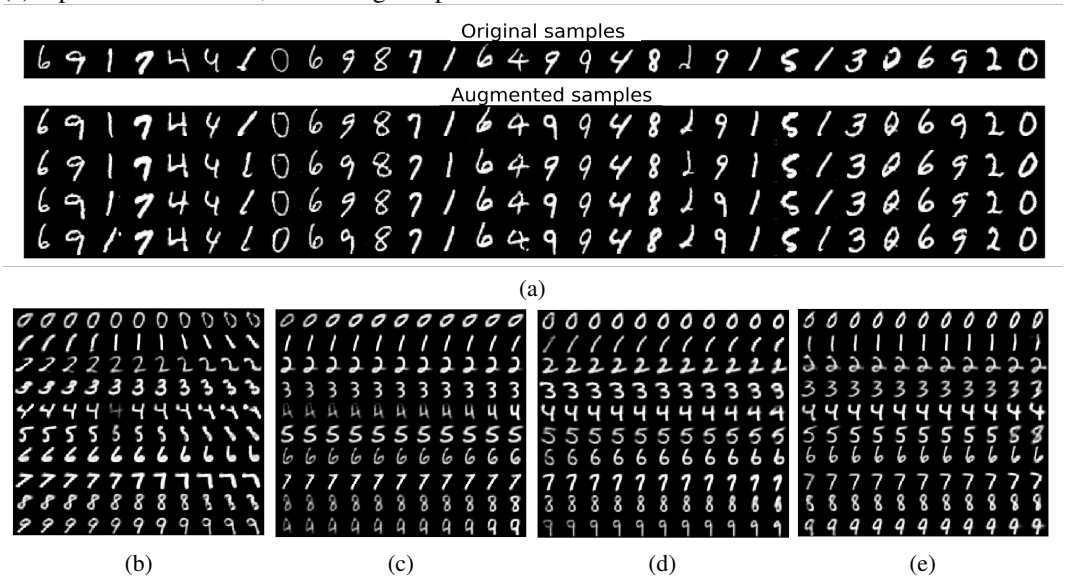

Figure 2: (a) Augmented samples for the MNIST dataset generated by the type-preserving augmentation conserve type of the original sample. (b-e) Continuous latent traversals of cpl-mixVAE with 10-dimensional continuous and 10-dimensional categorical variables. Examples of (b) rotation angle, (c) stroke thickness, (d) character width and (e) roundness of looped features are presented. The discrete variable $\mathbf{c}$ is constant for all reconstructions in the same row.

Method	$\mathcal{L}_{\text{rec}} \downarrow$	ACC $\uparrow$ (mean $\pm$ s.d.)
InfoGAN	169.4	77.87 $\pm$ 21.68
CascadeVAE	-	81.41 $\pm$ 09.54
JointVAE	122.0	68.99 $\pm$ 11.76
JointVAE†	126.7	62.19 $\pm$ 05.73
JointVAE‡	130.5	68.21 $\pm$ 09.58
cpl-mixVAE($\mathbf{s} \not| \mathbf{c}$)	113.9	79.63 $\pm$ 08.32
cpl-mixVAE^{a}($\mathbf{s} \mid \mathbf{c}$)	**105.9**	82.92 $\pm$ 04.64
cpl-mixVAE($\mathbf{s} \mid \mathbf{c}$)	113.5	**84.56** $\pm$ **06.47**

Table 1: Clustering results for the MNIST dataset, over 10 runs with 15,000 training iterations. For InfoGAN, we used the same network and the same parameters reported in the original paper by Chen et al. (2016). For CascadeVAE, all JointVAEs, and cpl-mixVAE models, we used $|\mathbf{c}| = 10$, $|\mathbf{s}| = 10$, and $\tau = 0.67$. For cpl-mixVAE models, the coupling factor is set to $\lambda = 1$.

Table 1 displays the classification performance of the discrete latent variable (as the predicted class label) for InfoGAN, different single-agent VAE methods including JointVAE and CascadeVAE, and cpl-mixVAE. We report the accuracy of the categorical assignments (ACC) and the reconstruction loss across 10 random initializations. For CascadeVAE, we used the numbers reported in (Jeong & Song, 2019). For InfoGAN and JointVAE, we used the publicly available implementation and training procedure reported in (Chen et al., 2016; Dupont, 2018). All reported numbers for cpl-mixVAE models are average accuracies calculated across both agents. We reported the performance of the proposed coupled VAE for three cases: (i) cpl-mixVAE($\mathbf{s} \not| \mathbf{c}$), in which the state variable is independent of the discrete variable, (ii) cpl-mixVAE($\mathbf{s} \mid \mathbf{c}$), which is the model in Eq. 6 using the proposed data augmentation, and (iii) cpl-mixVAE^{a}($\mathbf{s} \mid \mathbf{c}$), in which we used random rotation ($[-20, 20]$ degree) as an affine transformation for augmentation. Our results show that the cpl-mixVAE($\mathbf{s} \mid \mathbf{c}$) obtained the best categorical assignment among all models. Moreover, cpl-mixVAE^{a}($\mathbf{s} \mid \mathbf{c}$) also achieved the second highest performance, which demonstrates that even using a simple augmentation strategy can enhance the representation learning. For a fair comparison, Gumble-softmax temperature, τ and latent dimensionality are set to the same values as those for JointVAE and CascadeVAE. To understand whether architecture differences put JointVAE at a disadvantage, we report the results for JointVAE†, which uses the same architecture for the basic encoder/decoder networks as the one used in cpl-mixVAE. That is, JointVAE† uses the same learning procedure as JointVAE, but its convolutional layers are replaced by fully-connected layers (see supplementary Section H for implementation details). Comparison of the results obtained with JointVAE and JointVAE† suggests that the superiority of cpl-mixVAE is not due to the network architecture. Additionally, to separate the impact of augmentation in the training, we report the results for JointVAE‡, in which the JointVAE model has been trained with noisy copies of the original MNIST dataset generated by the proposed augmentation method. The reported clustering performance for JointVAE‡ suggests that including data augmentation by itself does not enhance the categorical assignment.

We also studied the performance of cpl-mixVAE($\mathbf{s}|\mathbf{c}$) for different cardinalities of the categorical variable, $\mathbf{c}$, to discover the true number of categories. Fig. 3a shows the performance of cpl-mixVAE in terms of ACC and AMP as a function of $|\mathbf{c}| = K \in [7, 13]$. Here, AMP denotes the average of maximum posterior of categories i.e., $1/K \sum_{k=1}^{K} \max p(c_k|\mathbf{x})$. Our results show that the best performance is obtained for $K = 10$. Fig. 3b demonstrates the probabilistic assignment of $\mathbf{c}$ for all digits. As illustrated, an inadequate number of categories (left panel) results in some dimensions being used for more than one digit, while extra c_k leaves some categories under-utilized (right panel).

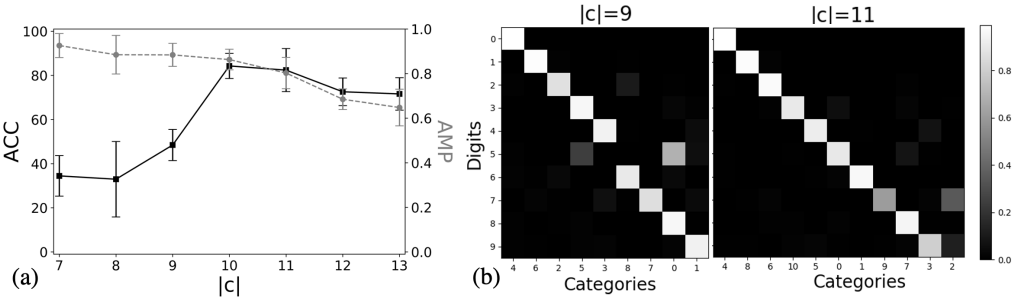

Figure 3: Clustering performance of cpl-mixVAE for MNIST, when the number of clusters ($|\mathbf{c}|$) is unknown. ACC is accuracy, AMP is average maximum posterior probability. Error bars indicate mean $\pm$ s.d. over 5 randomly initialized runs. Color bar indicates per-category accuracy.

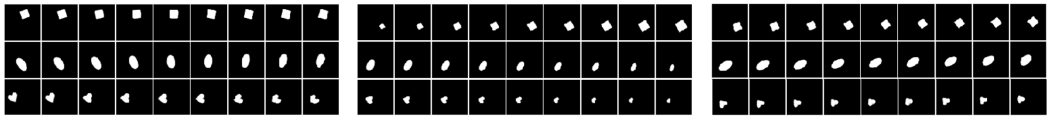

Figure 4: Interpretable continuous latent traversals of the trained cpl-mixVAE model with 6-dimensional continuous variable and 3-dimensional categorical variable for the dSprites dataset. Examples of (a) rotation, (b) scale, and (c) position are shown. The discrete variable is held as fixed for all reconstructions in the same row.

Table 2: Clustering results for the dSprites dataset, over 10 runs with $|\mathbf{c}| = 3$, $|\mathbf{s}| = 6$, $\tau = 0.67$, and $\lambda = 10$.

	JointVAE	CascadeVAE	cpl-mixVAE($\mathbf{s} \mid \mathbf{c}$)
ACC (mean $\pm$ s.d.)	44.79 $\pm$ 03.88	78.84 $\pm$ 15.65	**96.30 $\pm$ 09.15**

dSprites: In this dataset, due to the uniform distribution of classes, again we used a 2-agent cpl-mixVAE model. Similar to JointVAE and CascadeVAE, we used a 3-dimensional categorical variable for learning the shape (type), and a 6-dimensional state variable representing the style of each shape. Fig. 4 illustrates these traversal results for the three dimensions of the state variable obtained from the cpl-mixVAE. Each row corresponds to a different dimension of the categorical variable, and the state variable monotonically changes across the columns. In addition, Table 2 shows the degree to which cpl-mixVAE outperforms the other two methods in clustering.

scRNA-seq: Compared to typical machine learning datasets, the scRNA-seq dataset is exceedingly high-dimensional, with over 10,000 genes. It includes 22,365 neurons, over 100 cell types with sizeable difference between the most and the least abundant clusters. Here, we excluded non-neuronal cells and used a subset of 5,000 most expressed genes based on their peak values. While more than 115 neuronal types are suggested for this scRNA-seq dataset (Tasic et al., 2018), a significant challenge of representation learning in this dataset is its substantial imbalance, where for the most- and the least-abundant types, there exist 140,4 and 16 samples, respectively. From the perspective of neuroscience, neurons as the basic building blocks of the brain, display both significant diversity and stereotypy in their shapes, gene expression, and response patterns. Individual cells inherently differ due to either their type or continuous within-type variations (Trapnell, 2015; Andrews & Hemberg, 2018).

We used a 115-dimensional discrete and a 2-dimensional continuous variable for discrete and continuous neuronal factors representation, respectively. Fig. 5a illustrates the performance of the 2-agent cpl-mixVAE model. The dendrograms on the y-axis displays the hierarchical relationship between neuron types according to Tasic et al. (2018). For many neuronal cells, whether the observed diversity corresponds to discrete variability or a continuum is an ongoing debate. While JointVAE failed to identify meaningful cell types, cpl-mixVAE successfully identified the majority of known types. The, accuracy for categorical assignment across the entire 115 types is 39% (chance level is $\sim 6\%$, based on the most abundant type).

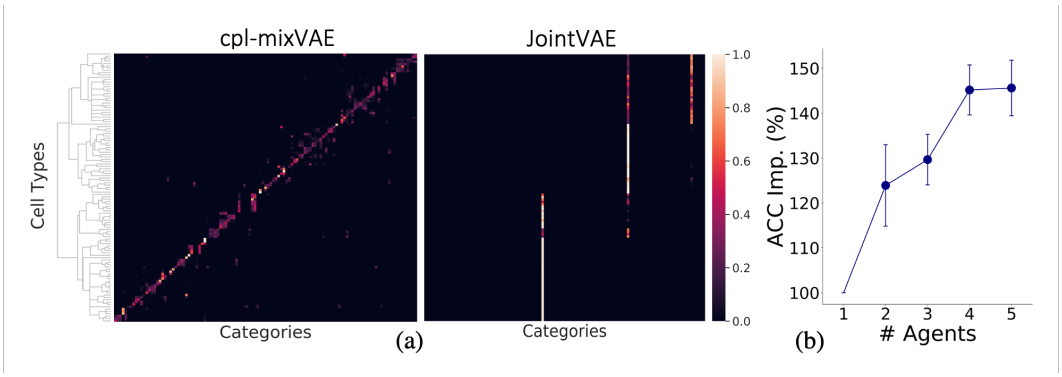

Figure 5: Categorical assignments for the scRNA-seq dataset. (a) Confusion matrices of cpl-mixVAE and JointVAE trained by $|\mathbf{c}| = 115$, $|\mathbf{s}| = 2$, $\tau = 1$, and $\lambda = 1$, over 45,000 iterations. The dendrogram on the y-axis shows marker-based hierarchical classification with 115 cell types. (b) Accuracy improvement by adding more agents in the cpl-mixVAE model, over 3 runs. The performance of A-agent for $A \geq 2$ is compared with the baseline single agent, JointVAE.

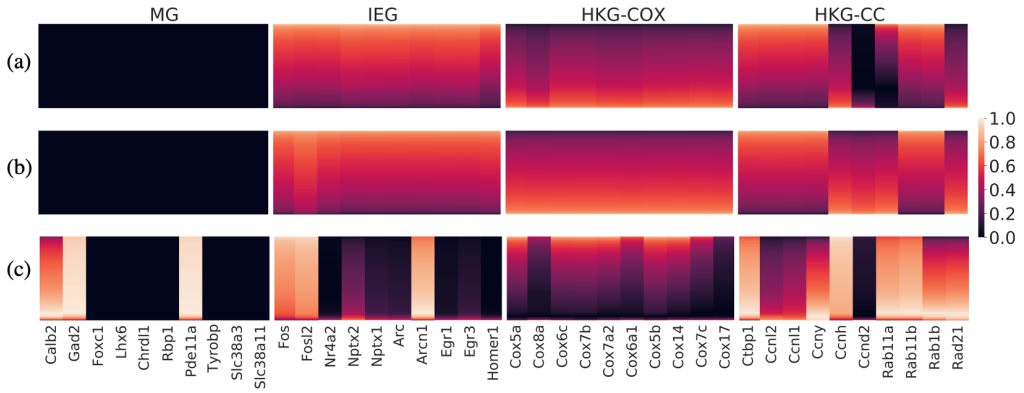

Figure 6: Continuous latent traversals for two excitatory cell types (a) "L5 NP ALM Trhr Nefl" and (b) "L6 CT Nxph2 Sla", and an inhibitory cell type (c) "Pvalb Akr1c18 Ntf3". For each type, the continuous latent traversal is color-mapped to a normalized reconstructed gene expression value (colorbar) as a function of the state variable for four gene subsets from left to right: marker genes (MG), immediate early genes (IEG), and two subgroups of house keeping genes, cytochrome c oxidase (HKG-COX) and cell cycle regulators (HKG-CC).

Unlike the discussed benchmark datasets, the neuronal types are not uniformly distributed, Accordingly, as another experiment we applied more than two agents on the scRNA-seq dataset to investigate the accuracy improvement for categorical assignment. Fig. 5b illustrates the accuracy improvement with respect the a single agent model, i.e. JointVAE.

To examine the role of the continuous latent variable, which can profile activity-regulated genes, we applied a similar traversal analysis to that used for the MNIST and dSprites datasets. For a given cell sample and its discrete type, we changed every dimension of the continuous variable using conditional distribution, and inspected gene expression changes caused by continuous variable alterations. Fig. 6 shows the results of the state traversal experiment for two excitatory neurons belonging to the "L5 NP" (near-projecting) and "L6 CT" (corticothalamic) sub-classes, and an inhibitory neuron belonging to the "PV" (parvalbumin) class. In each sub-figure, the latent traversal is color-mapped to normalized reconstructed expression values, where the y-axis corresponds to one dimension of the continuous variable, and the x-axis corresponds to four gene subsets, namely (i) marker genes (MG) for the two excitatory types, (ii) immediate early genes (IEG), and two house keeping gene (HKG) subgroups (iii) cytochrome c oxidase (COX), and (iv) cell cycle (CC) regulators (Hrvatin et al., 2018; Tarasenko et al., 2017). MGs are normally expected to function as indicators for particular cell types whose normalized expression is unaffected by the regulatory activities of the cell. Indeed, the expression of the reported excitatory MGs remains constant for excitatory traversals but not necessarily for the inhibitory traversal (i.e., Calb2, Gad2, Pde11a in Fig. 6). In contrast, the expression of IEGs and HKGs should depend strongly on the metabolic and environmental conditions. Indeed, we find that the expression changes of IEGs and HKGs are for the most part monotonically linked to the continuous variable, reaffirming that the it captures relevant, interpretable continuous variability, as in the MNIST and dSprites examples. Lastly, the expression of IEGs and HKGs (activity-regulated genes) depends on the cell type e.g., not all IEGs are activated for all cell types. Notably, for the excitatory "L5 NP" and "L6 CT" cells that are proximate in the hierarchy (suggested by Tasic et al. (2018)), state traversal is quite similar. These results suggest that the continuous variable inferred by cpl-mixVAE provides insight when deciphering the molecular mechanisms shaping the landscape of biological states, e.g. metabolic, disease.

5 Conclusion

We have proposed cpl-mixVAE as a multi-agent framework using a type-preserving data augmentation to apply the power of collective decision making in unsupervised joint learning of discrete and continuous factors. This framework utilizes multiple pairwise-coupled autoencoding agents with a shared categorical variable, while independently learning the continuous variables. Our experimental results for all three datasets show that cpl-mixVAE outperforms comparable models. In addition, for a challenging gene expression dataset, we showed that the proposed framework can identify annotated neuronal types and differentiate between type-dependent and activity-regulated genes.

REFERENCES

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

Tallulah S Andrews and Martin Hemberg. Identifying cell populations with scrnaseq. *Molecular aspects of medicine*, 59:114–122, 2018.

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.

Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

Emilien Dupont. Learning disentangled joint continuous and discrete representations. In *Advances in Neural Information Processing Systems*, pp. 710–720, 2018.

Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3): 279–300, 2003.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 7–16, 2014.

Rohan Gala, Nathan Gouwens, Zizhen Yao, Agata Budzillo, Osnat Penn, Bosiljka Tasic, Gabe Murphy, Hongkui Zeng, and Uygar Sümbül. A coupled autoencoder approach for multi-modal analysis of cell types. In *Advances in Neural Information Processing Systems*, pp. 9263–9272, 2019.

Søren Hauberg, Oren Freifeld, Anders Boesen Lindbo Larsen, John Fisher, and Lars Hansen. Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In *Artificial Intelligence and Statistics*, pp. 342–350, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2(5):6, 2017.

Sinisa Hrvatin, Daniel R Hochbaum, M Aurel Nagy, Marcelo Cicconet, Keiramarie Robertson, Lucas Cheadle, Rapolas Zilionis, Alex Ratner, Rebeca Borges-Monroy, Allon M Klein, et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nature neuroscience*, 21(1):120–129, 2018.

Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Unsupervised adversarial invariance. In *Advances in Neural Information Processing Systems*, pp. 5092–5102, 2018.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Yeonwoo Jeong and Hyun Oh Song. Learning discrete and continuous factors of data via alternating disentanglement. *arXiv preprint arXiv:1905.09432*, 2019.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: an unsupervised and generative approach to clustering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 1965–1972, 2017.

Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pp. 2946–2954, 2016.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pp. 1558–1566. PMLR, 2016.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

James Lucas, George Tucker, Roger B Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. In *Advances in Neural Information Processing Systems*, pp. 9403–9413, 2019.

Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. In *Advances in Neural Information Processing Systems*, pp. 3086–3094, 2014.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

Tatyana N Tarasenko, Susan E Pacheco, Mary Kay Koenig, Julio Gomez-Rodriguez, Senta M Kapnick, Francisca Diaz, Patricia M Zerfas, Emanuele Barca, Jessica Sudderth, Ralph J DeBerardinis, et al. Cytochrome c oxidase activity is a metabolic checkpoint that regulates cell fate decisions during t cell activation and differentiation. *Cell metabolism*, 25(6):1254–1268, 2017.

Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, 2018.

Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome research*, 25(10): 1491–1498, 2015.