One Datapoint is Not Enough: Disentangling Grouped Data with Confounding

Anonymous Authors[1]

Abstract

Group-instance disentanglement is the problem of learning separate representations for within-group and across-group variation. We introduce the Context-Aware Variational Autoencoder (CxVAE), a method which can perform group-instance disentanglement on datasets with confounding (i.e. where a single observation is not sufficient to accurately infer the group and instance variables). First, we generate a dataset with confounding that cannot be disentangled by the current state-of-the-art methods. Next, we improve upon these methods by proposing 3 modifications: 1) conditioning the instance variable on the group variable, 2) a more expressive group encoder, 3) a regularization objective that encourages independence between the instance variable and the grouping. Our method shows considerable gains in performance measured by several disentanglement metrics: holdout reconstruction error, unsupervised translation error, and latent code probing. Finally, we explore how adjusting the parameters of the data-generating process affects the performance gap between CxVAE and the state-of-the-art.

1. Introduction

Group-instance disentanglement is the task of learning a representation network $r(\mathbf{x})$ which transforms a group of observations $\mathbf{x} = \{x_1, ..., x_K\}$ (e.g. images grouped by author, temperature readings grouped by weather station, etc) into a *group* code u and a set of *instance* codes $\mathbf{v} = \{v_1, ..., v_K\}$. The goal is that the group code should capture all the information that distinguishes one group from another, while the instance code should capture all the information that distinguishes the observations within a group, but none of the group information.

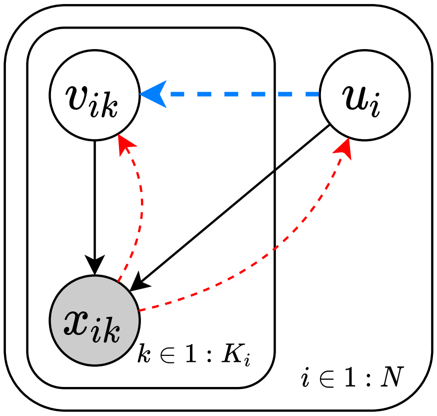


Figure 1. **Group-Instance Generative Model.** For each group i, a group variableu_i and a set of instance variables $\mathbf{v}$ generates the set of observations $\mathbf{x}$. The conditional $q(v_{ik}|u_i)$ (**blue arrow**) is absent from the existing literature but present in our proposed variational model. This enables our model to accurately infer the instance variable even when the groups are confounded.

The standard model for this task is the *Group-Instance Generative Model* (GIGM) (Bouchacourt et al., 2018; Hosoya, 2019), depicted in Figure 1. The model treats group and instance as independent latent variables involved in the generation of the data. We, in common with existing methods, train this model following the Variational Auteoncoder paradigm (Kingma & Welling, 2014; Rezende et al., 2014). In this framework, disentanglement is a property of the variational latent posterior $q(u, \mathbf{v}|\mathbf{x})$ which is implemented as a representation network.

Implementing the variational latent posterior so that it learns disentangled representations is not trivial and there exist several approaches. Challenges include how to accumulate evidence from multiple observations in order to infer the group variable (Bouchacourt et al., 2018; Hosoya, 2019) and how to design a regularization objective that encourages disentanglement in practice (Németh, 2020).

Many datasets require that the inference model be able to capture the joint distribution between the group and instance variables conditioned on the observations. These are datasets where two identical observations could have

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

come from different groups, and thus have different instance factors. We call these datasets *with confounding* because in the data-generating process u is a confounding variable when estimating v from x.

Such datasets are very common. For example, the problem of soybean plant growth (Davidian & Giltinan, 2003) has a dataset with confounding. It is impossible to deduce the age (instance) of a plant based solely on its leaf weight (observation) without also knowing its genotype (group).

We show empirically that the state-of-the-art methods fail to learn disentangled representations on datasets with confounding.We generate a simple dataset of student exam scores grouped by school. The instance variable captures the score of a student in relation to the distribution of scores at their school. Instead of disentangling, the existing methods split the representation of the group factor across both group and instance variables.

In order to tackle this challenge, we propose the Context-Aware Variational Autoencoder (CxVAE), another approach to the GIGM model with several novel modifications. The main novelty is an instance encoder conditioned on the inferred group variable, which enables the variational latent posterior to match the full generative latent posterior. Additionally, we use an expressive Deep Set network (Zaheer et al., 2017) to infer the group variable more accurately, and also design a novel regularization objective that encourages independence between the inferred instance variable and the ground-truth group factor.

We present the following contributions:

1. We formalize the notion of datasets with confounding and generate a simple dataset on which the current group-instance methods fail to learn disentangled representations.

2. We propose a new model, called the Context-Aware Variational Autoencoder (CxVAE) which can learn disentangled representations on emergent groups.

3. We perform extensive evaluation showing that our model outperforms the existing methods in holdout-data log-likelihood, unsupervised translation, and ground-truth factor prediction.

4. We complete an ablation study showing the relative performance gain brought by each modification of our model.

5. By varying the amount of confounding in the data-generating process, we show that confounding explains the performance gain of the CxVAE over existing methods. We also show how different magnitude ratios between the group and instance factors affect disentanglement.

2. Generative Group-Instance

The Group Variational Autoencoder (Bouchacourt et al., 2018; Hosoya, 2019) is a family of models that use two latent variables to represent grouped data: one that captures the variation within groups, and one for the variation across groups.

Assume a dataset of the form $\{x_{ik}\}_{i\in 1:N,\ k\in 1:K_i}$ where N is the number of groups and K_i is the number of observations in group i. GVAE defines a generative model that maps a $\mathcal{N}(0,1)$ group latent variable u_i and a $\mathcal{N}(0,1)$ instance latent variable v_{ik} to a given data observation x_{ik}. In other words, the likelihood of a group is:

$$p(\mathbf{x}) = \mathbb{E}_{p(u)} \prod_{k=1}^{K} \mathbb{E}_{p(v_k)} \left[p(x_k|u, v_k) \right]$$

We omit the index of the group i for notational simplicity, since the groups are independent and identically distributed.

2.1. Variational Inference

Because the exact likelihood is intractable, the Variational Autodencoder (Kingma & Welling, 2014; Rezende et al., 2014) performs optimization by introducting a variational latent posterior $q(u, \mathbf{v}|\mathbf{x})$ and maximizing the Evidence Lower Bound (Jordan et al., 2004):

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(u,\mathbf{v}|\mathbf{x})}[\log p(\mathbf{x}|u,\mathbf{v})] - \mathrm{KL}[q(u,\mathbf{v}|\mathbf{x})||p(u,\mathbf{v})]$$

Existing methods use a class of variational distributions that assume independence between the latent variables in a group when conditioned on the data.

$$q(u,\mathbf{v}|\mathbf{x}) = q(u|\mathbf{x}) \prod_{k=1}^{K} q(v_k|x_k)$$

In our work, we show that this assumption hinders disentanglement when the generative model is entangled.

2.2. Group Encoder

We, in common with existing methods, implement the variational group posterior with normal density with (μ, Σ) computed with an encoder network. The way to implement the encoder network is not obvious, since the number of inputs K varies across groups. Hosoya (2019) encode separately each observation in the group using the same encoder E_u to produce μ_k, Σ_k and then averages the outputs.

$$\mu = \frac{1}{N}\sum_{k=1}^{K} \mu_k,\ \Sigma = \frac{1}{N}\sum_{k=1}^{K} \Sigma_k$$

Bouchacourt et al. (2018) also encode each observation individually and then accumulate the evidence through a product of normal densities, computed using the following equations:

$$\Sigma^{-1} = \sum_{k=1}^{K} \Sigma_k^{-1},\ \mu^T \Sigma^{-1} = \sum_{k=1}^{K} \mu_k^T \Sigma_k^{-1}$$

They justify that such a product of normals produces a valid evidence accumulation using the following result:

$$q(u|\mathbf{x}) \propto \prod_{k=1}^{K} q(u|x_k)$$

However, the above is not a universal property, since

$$\begin{aligned} q(u|\mathbf{x}) =& \frac{\prod_{k=1}^{K} q(x_k)}{q(\mathbf{x})} \frac{1}{q(u)^{K-1}} \prod_{k=1}^{K} q(u|x_k) \\ \propto& \frac{1}{q(u)^{K-1}} \prod_{k=1}^{K} q(u|x_k) \end{aligned} \tag{1}$$

In fact, by using a product of normals to accumulate evidence, the authors implicitly assume that the marginal distribution of the inferred group variable is a uniform. This has the effect of sampling u values which are less representative of the current group and more skewed towards the marginal distribution of u.

In our work, we propose a more general approach to encoding u by using a Deep Sets network (Zaheer et al., 2017) to encode the whole set of observations instead of encoding each observation separately.

2.3. Regularization

In certain cases, the model might learn to encode both kinds of variation (within- and across-group) in the instance variable, effectively turning the model into a standard VAE. In this eventuality, the group variable becomes irrelevant and disentanglement is not achieved.

Such behaviour has been identified by Hosoya (2019) and Németh (2020) to occur when the instance code too high-dimensional, the instance encoder too expressive, or group sizes too small. One solution is to limit the dimensionality of the instance code (Hosoya, 2019), with the downside of hindering the overall model performance.

As a more targeted solution, Hosoya (2019) propose an adversarial loss minimizing the mutual information between an observation and the instance variable inferred from the other observations in the group:

$$I_r(x, v) = \mathrm{KL}[r(x, v)||r(x)r(v)]$$

where $r(x, v) = r(v|x)r(x)$ is the joint distribution of an observation and the instance variable inferred from any of the other observations in the group and

$$r(v|x_k) = \frac{1}{K-1} \sum_{l=1,\ l\neq k}^{K} q(v|x_l)$$

The mutual information is approximated empirically using the results of Belghazi et al. (2018).

$$\begin{aligned} I_r(x, v) \approx \max_T\ & \mathbb{E}_{r(x,v)}[T(x, v)] \\ & - \log \mathbb{E}_{r(x)r(v)}[\exp T(x, v)] \end{aligned} \tag{2}$$

T is a neural network and the expectation terms are computed by sampling.

- To sample $r(x, v)$, first choose a group i, then choose two instances from that group $k, l \in K_i$. x will be the observation x_{ik} and v will be sampled from $q(v|x_{il})$.
- To sample $r(x)r(v)$, choose two groups i, j and two instances in each group $k \in K_i, l \in K_j$. Take x_{ik} for x and sample $q(v|x_{jl})$.

In our view, this method has the following limitation: Even when the instance variable does contain group information, the value of $I_r(x, v)$ might still be small, because it might be difficult to ascertain the group based on one single observation x. In our work, we propose a modification to this regularization term such that the network T takes as input all the observations in the group instead of only one.

3. Entangled Group and Instance Variables

We call the group and instance variables *entangled* when they are not independent conditioned on the data $p(u, v|x) \neq p(u|x)p(v|x)$. A useful heuristic for establishing whether the variables are entangled is to ask “Does knowing the group variable for an observation influence my belief about its instance variable?”

This property of the generative model is present in many machine learning tasks, such as collaborative filtering, 3D novel view synthesis. For example, in the context of the Netflix Challenge, where the task was to predict what score a user would give to a new film, one cannot infer what film is associated with a given score without also knowing the user.

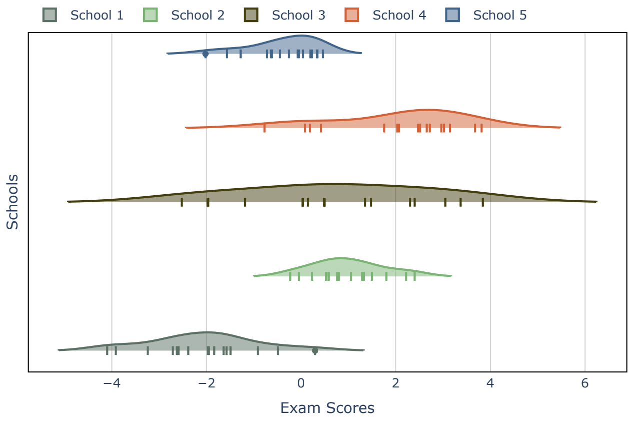


Figure 2. **Example of our exam-scores dataset.** Each school is a different group of student scores. The relative quality of a score cannot be inferred without knowing the distribution of scores in the school.

Strictly speaking, most real-world models are entangled. However, in many cases, the mutual information between the group and instance variable, conditioned on the observation, is negligible. For example, in handwritten digit recognition, one can infer the digit value depicted in an image without knowing the author.

In this paper, we claim that the current methods in the GVAE family do not perform well in tasks where the group and instance variables are entangled.

3.1. Exam-Scores Problem

Suppose we wanted to model the exam scores of students from different schools. Our model must separate the school-level effect (the group factor) from the student-level effect (the instance factor). We define the following generative model:

$$x_{ik} = 2\,\mu_i + (\sigma_i^2 + 1)v_{ik},\ i \in 1:N,\ k \in 1:K_i$$

where x_{ij} is the student score, $u_i = (\mu_i, \sigma_i)$ is the school-level effect, v_{ij} is the student-level effect. We assume a $\mathcal{N}(0,1)$ prior distribution for the latent variables.

We first sample the model to generate a dataset ($N = 32,768,\ K_i \sim \text{Poisson}(16) + 8$) and then use the same model as the generative model in our Variational Autoencoder, instead of a neural network. The figure below shows what the data looks like.

Looking at the data, it is easy to see that this model is entangled, because the relative performance v of a student within their own school, given their absolute score x, depends on the distribution of scores within the school u.

4. Context-Aware Variational Autoencoder (CxVAE)

We propose a new model which can perform well on datasets generated from entangled group and instance variable. We call our model the Context-Aware Variational Autoencoder. Our model comprises the following changes with respect to the standard GVAE:

1. The group encoder is implemented as a Deep Sets network (Zaheer et al., 2017). The encoder has the following form:

   $$\mu, \Sigma = E_u(\mathbf{x}) = E_u^B\left(\sum_{k=1}^{K} E_u^A(x_k)\right)$$

   where E_u^A, E_u^B are two neural networks.

2. The variational instance posterior is dependent on the inferred group variable:

   $$q(u, \mathbf{v}|\mathbf{x}) = q(u|\mathbf{x}) \prod_{k=1}^{K} q(v_k|x_k, u)$$

   In practice, our instance encoder takes as input a vector concatenating x_k and u. This idea is not new, and has been used previously in sequence disentanglement (Li & Mandt, 2018). This allows the instance encoder to differentiate the between observations with similar values but which come from different groups.

3. We propose a regularization objective similar to the one in Németh (2020), but whereby we minimize *the mutual information between one inferred instance variable and all the other observations in the group*. More precisely, our objective is to minimize $I_r(\mathbf{x}_{-k}, v_k)$ where $r(v_k|\mathbf{x}_{-k}) = q(v_k|x_k, u)$. Following the same approximation as in Németh (2020), the objective takes the following form:

   $$\begin{aligned} I_r(\mathbf{x}_{-k}, v_k) \approx \max_T\ & \mathbb{E}_{r(\mathbf{x}_{-k}, v_k)}[T(\mathbf{x}_{-k}, v_k)] \\ & - \log\mathbb{E}_{r(\mathbf{x}_{-k})r(v_k)}[\exp T(\mathbf{x}_{-k}, v_k)] \end{aligned} \tag{3}$$

   T is implemented as a Deep Sets neural network for the observations, with the instance code concatenated in the middle:

   $$T(\mathbf{x}_{-k}, v_k) = T^\beta\left(v_k, \frac{1}{K}\sum_{l=1,\ l\neq k}^{K} T^\alpha(x_l)\right)$$

   Again, the expectations are computed by sampling:

- To sample $r(x, v)$, first choose a group i, then choose one instance from that group $k \in K_i$. $\mathbf{x}_{i,-k}$ will be all the observations in the group apart from k, and v_k will be sampled from $q(v|x_{ik})$.
- To sample $r(\mathbf{x}_{-k})r(v_k)$, choose two groups i, j and two instances in each group $k \in K_i,\ l \in K_j$. Take $\mathbf{x}_{i,-k}$ for $\mathbf{x}_{-k}$ and sample $q(v|x_{jl})$.

This regularization objective also minimizes the mutual information between the inferred instance variables and the true data generating group variable, but uses as a proxy for the latter all but one of the observations in the group, instead of just one observation.

5. Measuring Disentanglement

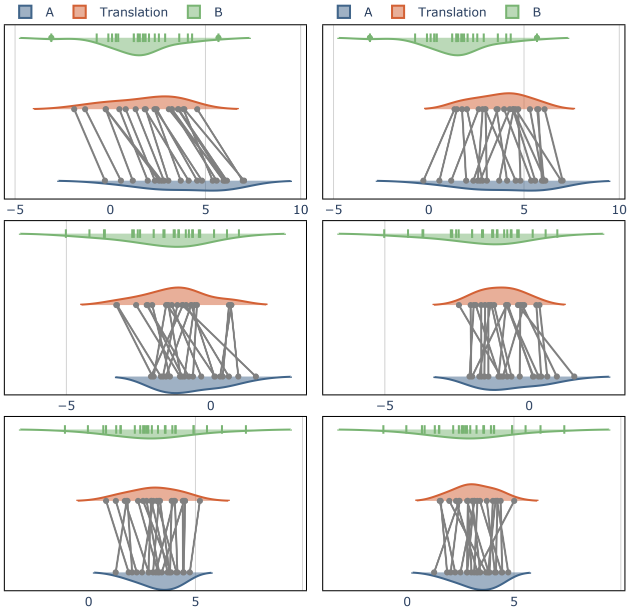
5.1. Probing u and $\mathbf{v}$

Our first evaluation metric is a direct application of the definition of disentanglement; that a given representation should capture as much information as possible about the corresponding data-generating factor, and as little as possible about the other factors (Locatello et al., 2020). In the case of group-instance, the inferred group variable $\hat{u}$, for example, should be maximally predictive of the data-generating group factor u, and minimally predictive of the set of instance factors $\mathbf{v}$

However, not all combinations of inferred latents and data-generating factors are worth analysing. As pointed out by Németh (2020), the failure case we are trying to prevent in group-instance disentanglement is that the inferred instance variables $\hat{\mathbf{v}}$ might learn information pertaining to the data-generating group factor u. Other cases, such that the inferred group variable might learn to represent instance information, are not possible because of the structure of the variational model. The group encoder is an order-agnostic accumulation of input observations and the instance encoder is applied separately to each observation (so no way for information from observation x_l to bleed into instance variable $\hat{v}_k$).

Thus, the metric is computed as follows: at training time, we learn two regressors, called probes, to predict the data-generating group factor u: one probe takes as input the inferred group variable $\hat{u}$ and the other takes as input the set of inferred instance variables $\hat{\mathbf{v}}$. The prediction error rate of these regressors on the holdout set gives the disentanglement metric; the more disentangled the representation is, the lower the error of the u-probe and the higher the error of the v-probe. Since we are using a synthetic dataset for evaluation, we can measure the mean squared error directly between the prediction and the ground-truth group factor.

Training predictors on latent representations is a universal approach for evaluating disentanglement. In particular, Bouchacourt et al. (2018); Németh (2020) have a very similar setup to ours, the only difference being that they use a classifier instead of a regressor. This is because they evaluate on real data whose ground-truth factors are unknown, so they use human lables as a proxy.



(a) CxVAE (b) ML-VAE

Figure 3. **Unsupervised translation with our CxVAE and the ML-VAE**. The task is to generate a set of observations with the instance variables or the source group and the group variable of the target group. Visual inspection suggests that our model better captures the mean and variance of the target group.

5.2. Unsupervised Translation

Unsupervised translation is the process of transforming an observation by changing its group code while keeping its instance code fixed. This is a primary downstream task for disentangled representations because it requires a clean separation between the group and instance representations (Tenenbaum & Freeman, 2000). Therefore, it can be used to quantify the quality of disentanglement.

Formally, let i, j be the indices of the source and target group, respectively. Translation involves sampling a group of instance codes from the source group $q(\mathbf{v}_i|\mathbf{x}_i)$ and a group code from the target group $q(u_j|\mathbf{x}_j)$. Then, we generate the translated observations by combining each instance code with the group code $p(x'_k|\hat{u}_j, \hat{\mathbf{v}}_j)$. An example of translation comparing our CxVAE with the ML-VAE (Bouchacourt et al., 2018) can be seen in Figure 3.

We measure the translation quality by taking the mean squared error between the translation performed with each

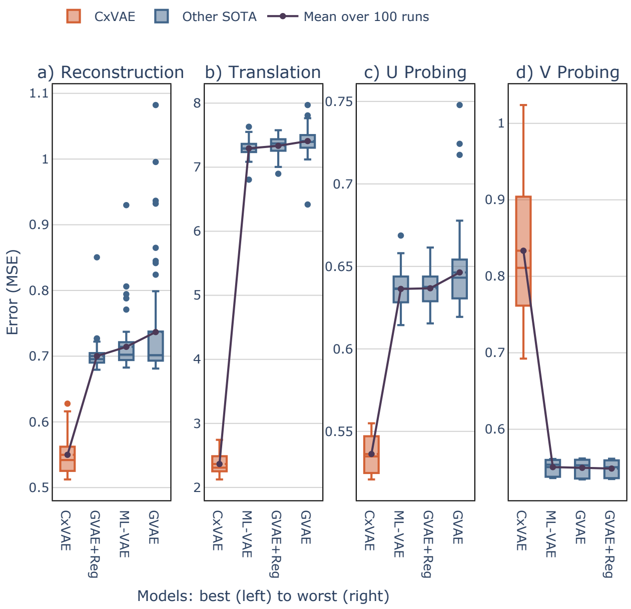


Figure 4. Our CxVAE produces considerable improvements over the state-of-the-art in every disentanglement metric considered: holdout reconstruction error (**lower is better**), translation error (**lower is better**), u-probing (**lower is better**) and $\mathbf{v}$-probing (**higher is better**). We show box-plots for the distribution of scores over 100 random initialisations.

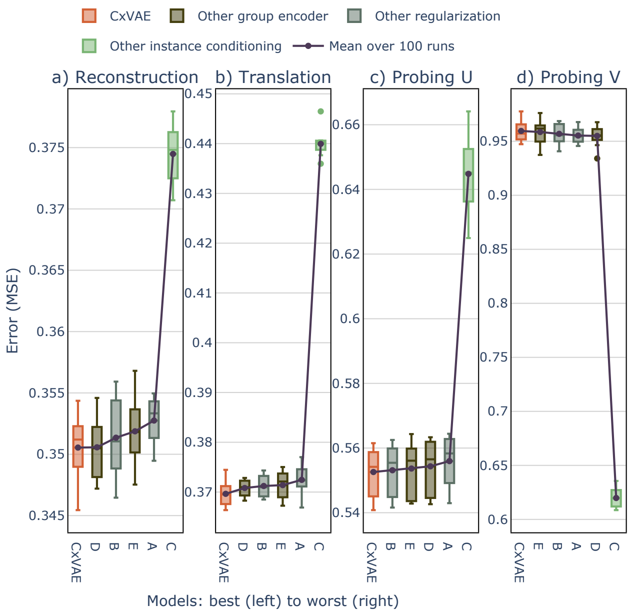


Figure 5. **Ablation study.** Replacing any novel component of our CxVAE with an existing alternative leads to worse performance. Note that the largest reduction in performance is brought by making the instance encoder unconditional. The next largest reduction is caused by changing the regularization, followed by using a less expressive group encoder.

model and the ground-truth translation computed using the true data-generating process.

6. Experiments

In the first experiment, we compare our CxVAE with the state-of-the-art in group-instance disentanglement, namely the Multi-Level VAE (ML-VAE) (Bouchacourt et al., 2018), the Group VAE (GVAE) (Hosoya, 2019) and the Adversarial GVAE (GVAE-AD) (Németh, 2020).

We generate a dataset with $N = 65,536$ groups. The size of each group is distributed according to a Poisson distribution $K_i \sim 8 + \text{Poisson}(16)$. We train each model for 64 epochs, and use the last 10 epochs for evaluation. Additionally, we run the experiment for 10 different random seeds initialisations, both for the data generating process and the networks. We use the same 10 seeds to compare. This gives us 100 measurements to plot in Figure 4.

Results Our CxVAE produces considerable improvements over the state-of-the-art in every disentanglement metric considered (Figure 4). While the scores of the existing methods cluster together, the gap between them and the CxVAE is larger than any 95% confidence interval. It is also interesting to note how CxVAE produces a smaller reconstruction error on the holdout dataset, suggesting that disentangled representations generalise better to unseen groups.

6.1. Ablation Study

In order to quantify the contribution of each of the modifications that comprise our proposed model, we perform an ablation study whereby we measure the decrease in performance resulting from replacing a proposed component with an existing alternative. For example, when considering the group encoder, we include in the comparison two hybrid models based on the CxVAE: one of them has a product of normals encoder as in the ML-VAE, while the other has an averaging encoder as in the GVAE. For the instance encoder, the alternative is to have no conditioning on the group variable. For the regularization, we compare with the adversarial loss of GVAE+AD and with no regularization at all. The results of the experiment are displayed in Figure 5.

Results The full CxVAE has a better performance than any alternative configuration. In particular, the choice of a conditional instance encoder brings the largest improvement in performance, followed by the regularization objective, then the group encoder.

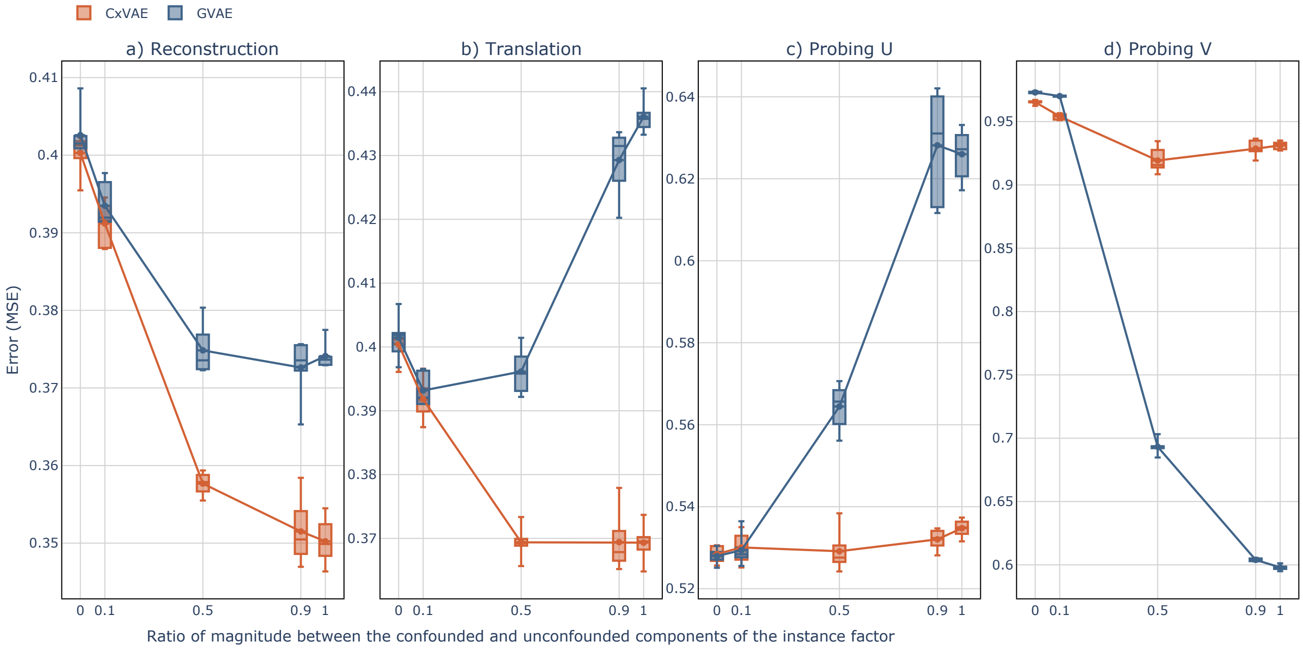


Figure 6. **The performance gain of our CxVAE increases with more confounding.** We show disentanglement on datasets generated with different values of the γ hyper-parameter, controlling the ratio of magnitudes between the confounded and unconfounded components of the instance factor. For low values of γ the instance factor is not confounded and so the CxVAE and GVAE perform equally well. However, as γ increases, GVAE performs more poorly.

6.2. Degree of Confounding

Is the increased performance of our CxVAE solely due to the confounded data-generating process, or is it simply a better all-round model? To answer this, we create different exam-scores datasets with different degrees of confounding. We do this by making the data 2-dimensional, with one dimension being confounded and the other being controlled only by the instance variable, and then trading off the magnitude of one to the other. Our new data generating process is the following:

$$x_{ik} = \begin{bmatrix} u_i + \gamma a_{ik} \\ (1-\gamma) b_{ik} \end{bmatrix}, \; v_{ik} = \begin{bmatrix} a_{ik} \\ b_{ik} \end{bmatrix} \tag{4}$$

where a_{ik} is the dimension of the instance factor confounded by the group factor, while b_{ik} is the unconfounded dimension. The trade-off is controlled using a hyper-parameter $\gamma \in [0, 1]$. For low values of γ, the confounded variable should explain a small amount of the variation within groups. In that case, our CxVAE should not perform much better than existing methods, if our hypothesis about the relationship between CxVAE and confounding is correct. Large values of γ correspond to the settings we have already tested, so the performance gap should be large.

Results The measurements displayed in Figure 6 confirm our expectations. For low values of γ the performance of our CxVAE is evenly matched to the GVAE. As γ increases, CxVAE scores remain relatively stable while GVAE scores decrease substantially. It is clear that the degree of confounding in the dataset explains the performance gain that we see in the CxVAE.

6.3. Stronger Group or Stronger Instance

It is interesting to observe how changing the relative strength of the group and instance ground-truth factors in the dataset impacts differentely the disentanglement in our CxVAE and GVAE. Since our data is generated by a simple mixed model, we can adjust the data-generating process as follows:

$$x_{ik} = \lambda u_i + (1-\lambda) v_{ik} \tag{5}$$

where $\lambda \in [0, 1]$ is a hyper-parameter controlling the strength of the group factor μ_i relative to the instance factor v_{ik}.

We expect that different values for λ would change the characteristics of the disentanglement problem. A low value $\lambda \approx 0$ would generate a dataset with weak grouping that is virtually i.i.d. and that a normal autoencoder could represent. A large value $\lambda \approx 1$ would generate a dataset with strong

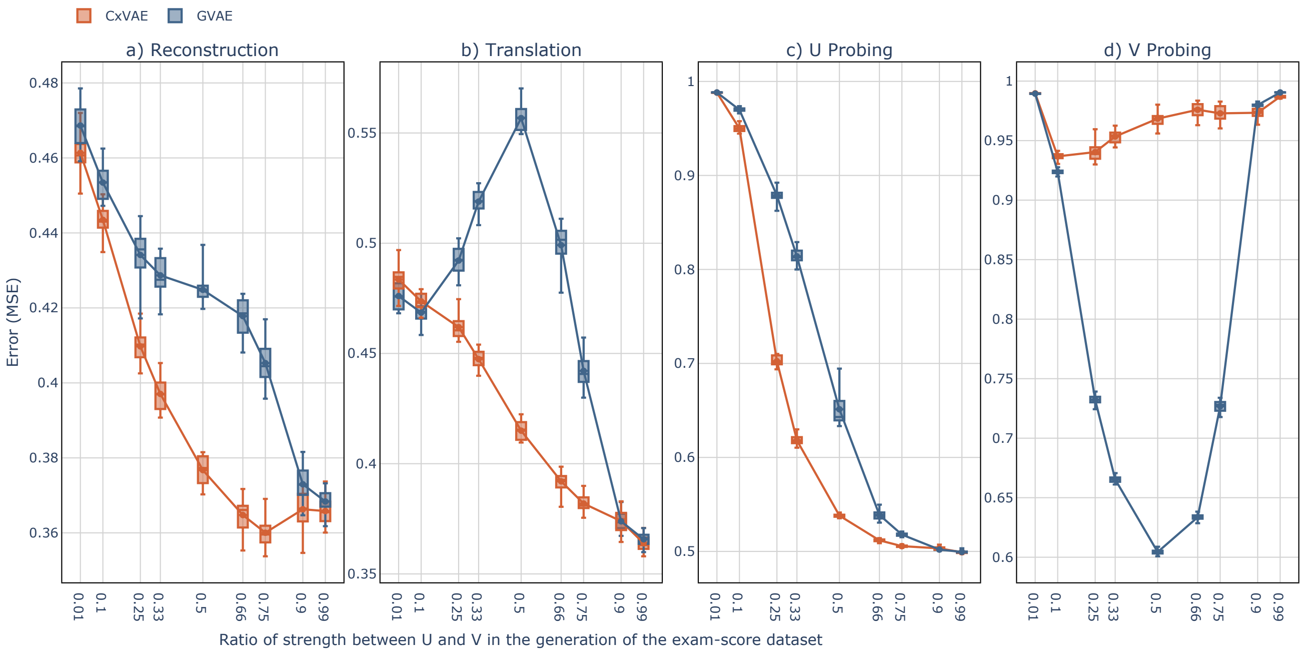


Figure 7. **Similar magnitudes of the group and instance factors produces a dip in the performance of the GVAE.** We show disentanglement on exam-score datasets generated with different values of the λ hyper-parameter, controlling the ratio of magnitudes between the group and instance factors. The greatest gap in performance between our CxVAE and the GVAE occurs when $\lambda = 0.5$ when the magnitude of the variation across groups is equal to the magnitude within groups. Too small (weak groups) or too large (strong groups) values of λ lead to a normal autoencoding problem.

groups where the observations within a group are noisy copies of one another, a setting that can also be solved by simple autoencoding. A middle value $\lambda \approx 0.5$ would evenly match the impact of the group and instance variables. Here we expect the performance gain of the CxVAE over competing methods to be greatest.

Results Looking at Figure 7 we note, as expected, that the largest performance gain of the CxVAE occurs at $\lambda = 0$. Specifically, while the performance of the CxVAE increases steadily as λ goes from 0 to 1, the scores of GVAE dip significantly around $\lambda = 0.5$. Translation and $\mathbf{v}$-probing are most sensitive to the decrease in performance of the GVAE. We also observe that the scores generally increase with λ even as the average magnitude of the data remains constant, possibly because for $\lambda \approx 0$ the group encoder cannot be used for learning, and thus expressiveness is lost.

7. Related Work

8. Conclusions

References

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Hjelm, R. D., and Courville, A. C. Mutual information neural estimation. In *ICML*, 2018.

Bouchacourt, D., Tomioka, R., and Nowozin, S. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *AAAI Conference on Artificial Intelligence*, 2018.

Chen, J. and Batmanghelich, K. Weakly supervised disentanglement by pairwise similarities. *AAAI Conference on Artificial Intelligence*, 2020.

Davidian, M. and Giltinan, D. M. Nonlinear models for repeated measurement data: An overview and update. Number 4, pp. 387–419, 2003.

Hosoya, H. Group-based learning of disentangled representations with generalizability for novel contents. In *IJCAI*, 2019.

Jordan, M. I., Ghahramani, Z., Jaakkola, T., and Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 2004.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

Li, Y. and Mandt, S. Disentangled sequential autoencoder. In *ICML*, 2018.

Liang, K. Y. and Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

Locatello, F., Poole, B., Rätsch, G., Scholkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Németh, J. Adversarial disentanglement with grouped observations. In *AAAI*, 2020.

Pinheiro, J. C. and Bates, D. M. Mixed-effects models in s and s-plus. *Technometrics*, 43:113 – 114, 2001.

Quinn, G. P. and Keough, M. Experimental design and data analysis for biologists. 2002.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

Schafer, J. L. and Graham, J. W. Missing data: our view of the state of the art. *Psychological methods*, 7 2:147–77, 2002.

Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgSwyBKvr.

Tenenbaum, J. B. and Freeman, W. T. Separating style and content with bilinear models. *Neural Computation*, 12:1247–1283, 2000.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Póczos, B., Salakhutdinov, R., and Smola, A. Deep sets. In *NIPS*, 2017.

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one, even using the one-column format.