One Datapoint is Not Enough: Disentangling Grouped Data with Confounding

Dan Andrei Iliescu [1] **Damon J Wischik** [1]

Abstract

Group-instance disentanglement is the problem of learning distinct representations for within-group and across-group variation. We introduce the Context-Aware Variational Autoencoder (CxVAE) – a method that can perform group-instance disentanglement on group-confounded problems (i.e., datasets where a single observation is insufficient for inferring instance variable accurately). We construct a synthetic dataset on which current state-of-the-art methods fail to disentangle. We propose a novel method whose instance encoder is conditioned on the group variable. Our model achieves considerable gains in both disentanglement quality and performance on the downstream task of multiple imputation. Finally, we show how the performance gap widens between CxVAE and the current state-of-the-art as we increase the strength of the confounding effect in our dataset.

1. Introduction

Imagine a dataset of observations organized into N groups $\mathbf{x}_i = \{x_{i1}, ..., x_{iK_i}\}$, $i \in 1 : N$. These could be pictures grouped by content, clinical outcomes grouped by the patient, or film ratings grouped by user.

Group-instance disentanglement (GID) is the goal of training a representation network $r(\mathbf{x})$ to produce a *group* code that captures only the variation across groups and a set of *instance* codes that capture only the variation within groups.

In the literature, this class of problems comes under different names: style-content disentanglement (Tenenbaum & Freeman, 2000), content-transformation disentanglement (Hosoya, 2019), and disentanglement with group supervision (Shu et al., 2020), to name a few.

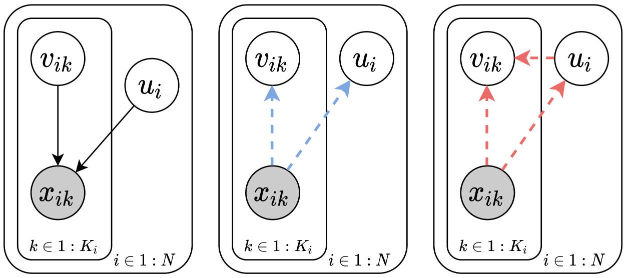


Figure 1. Group-instance generative model (**left**). Variational latent posterior of current methods (**middle**). Our conditional variational latent posterior (**right**).

The current state-of-the-art in GID uses a hierarchical process to model the data: an unobserved group variable u and an independent unobserved instance variable v_k come together to generate an observation x_k (Bouchacourt et al., 2018; Hosoya, 2019; Németh, 2020). This is the group-instance generative model (Figure 1-left). Part of the standard setup is to train this model using the Variational Autoencoder framework (Kingma & Welling, 2014; Rezende et al., 2014) by optimising a lower-bound to the data likelihood expressed in terms of a variational latent posterior distribution $q(u, \mathbf{v}|\mathbf{x})$.

When defining the variational latent posterior, existing works (Bouchacourt et al., 2018; Hosoya, 2019; Németh, 2020; Chen & Batmanghelich, 2020) make the simplifying assumption that the group and instance variables are conditionally independent given the observations (Figure 1-middle). We refer to this property as the *independence assumption*:

$$q(u, \mathbf{v}|\mathbf{x}) = q(u|\mathbf{x}) \sum_k q(v_k|x_k) \tag{1}$$

Our study shows empirically that this assumption hinders disentanglement on a yet unexplored class of downstream problems for GID: these problems have the property that the conditional distribution between the instance variable and observation $p(v|x)$ changes from group to group, thus making it impractical and even unfeasible to infer the instance variable from a single observation. We call these *group-confounded problems* because the group variable acts

[*] Equal contribution [1] Department of Computer Science, University of Cambridge, UK. Correspondence to: Dan Andrei Iliescu <dai24@cam.ac.uk>.

Proceedings of the 39^{th} *International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

as a confounder for the relationship between x and v.

Our strategy to learn representations under group-confounding is deceptively simple: condition the inference variable on the previously inferred group variable $q(v_k|x_k, u)$ (Figure 1-right). In statistics, conditioning on the confounding variable is the standard approach to control its effect (Greenland et al., 1999).

To our knowledge, we are the first to apply the paradigm of confounding to the group variable in a latent inference problem. GID typically follows the philosophy of fair representations (Louizos et al., 2016; Achille & Soatto, 2018) where the grouping is interpreted as *nuisance variable* whose effect on the inferred instance variable must be minimised using additional regularisation objectives (Németh, 2020). As we will show, controlling the effect of the group variable by conditioning on it (confounding approach) rather than training agains it (invariance approach) encourages marginal independence between the latent variables whilst simultaneously allowing for more precision in the estimation of the instance variables.

1.1. Motivating Problem

Consider the problem of multiple imputation in standardized test scores (Gelman & Hill, 2006). We are given a dataset containing the scores of different students over several standardised tests. Some students did not take some tests, so the corresponding scores are missing. The task is to impute the missing scores in order to clean the dataset for further analysis.

Group-instance disentanglement provides a straightforward way to perform multiple imputations on this dataset. First, we establish that the group variable represents the student, and the instance variable represents the test. To impute the score that student i would get on test j, we train the GID model on a dataset comprising test scores grouped by students. Second, we infer the instance variables for scores that other students have obtained on test j. Finally, we generate a distribution of scores by combining the group variable for student i with the instance variables for test j. We can report several statistics of this empirical distribution (Rubin, 1996) but for our evaluation we take the mean.

On this problem, GID has a considerable advantage over using categorical variables to model tests (Pinheiro & Bates, 2001; Gelman & Hill, 2006): GID can easily generalise to new students and tests without training again (Tenenbaum & Freeman, 2000).

The limitation of the current GID methods is the independence assumption Equation (1): because the instance variables are inferred separately for each observation the instance encoder discards useful information. Because different students will have different scores for the same test, inferring the test effect without conditioning on the student will produce widely varying estimates across groups (Pinheiro & Bates, 2001).

In our work, we use this imputation problem as a proof of concept by comparing the existing state-of-the-art in group-instance disentanglement (Bouchacourt et al., 2018; Hosoya, 2019; Németh, 2020) with our proposed model, the Context-Aware Variational Autoencoder (CxVAE), that uses a conditional instance encoder. The same conclusions apply to many other similar imputation problems with group-confounding, such as collaborative filtering using matrix factorization models (Koren et al., 2009) or imputation in longitudinal studies (Spratt et al., 2010).

Our contribution is threefold:

1. We show empirically that conditioning the instance encoder on the group variable leads to a considerable increase in performance on the multiple imputation task of standardised test scores (Gelman & Hill, 2006).
2. Using the Mutual Information Gap (Chen et al., 2018), we show that the learned representations of the conditional model are more disentangled than the representations of the unconditional models.
3. Finally, we investigate the performance gap between our conditional model and the current best unconditional model. We show that this gap can be fully explained by the confounding effect of the group variable, both in terms of multiple imputations error and disentanglement quality.

2. Background

The Group-Instance Generative Model (Bouchacourt et al., 2018; Hosoya, 2019) is a multi-level model that uses two latent variables to generate grouped data: the instance variable $v_{ik} \sim \mathcal{N}(0, 1)$ controls the variation within groups, and the group variable $u_i \sim \mathcal{N}(0, 1)$ controls the variation across groups (Figure 1-left). The likelihood of a group $\mathbf{x}$ is:

$$p(\mathbf{x}) = \mathbb{E}_{p(u)} \prod_{k=1}^{K} \mathbb{E}_{p(v_k)} \left[p(x_k|u, v_k) \right] \tag{2}$$

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

2.1. Variational Inference

Because the exact likelihood is intractable, the standard approach to train the GID model is with a Variational Autoencoder (Kingma & Welling, 2014; Rezende et al., 2014) which performs optimisation by introducing a variational latent posterior $q(u, \mathbf{v}|\mathbf{x})$ and maximizing the Evidence Lower Bound (Jordan et al., 2004):

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(u,\mathbf{v}|\mathbf{x})}[\log p(\mathbf{x}|u, \mathbf{v})] - \mathrm{KL}[q(u, \mathbf{v}|\mathbf{x})||p(u, \mathbf{v})] \tag{3}$$

Existing methods use a class of variational distributions that assume conditional independence between the latent variables.

$$q(u, \mathbf{v}|\mathbf{x}) = q(u|\mathbf{x}) \prod_{k=1}^{K} q(v_k|x_k) \tag{4}$$

3. Context-Aware Variational Autoencoder

We propose a new model which can perform well on group-confounded problems. We call our model the Context-Aware Variational Autoencoder (CxVAE).

The defining feature of our model is a variational latent posterior whose instance variable is conditioned on the group variable:

$$q(u, \mathbf{v}|\mathbf{x}) = q(u|\mathbf{x}) \prod_{k=1}^{K} q(v_k|x_k, u) \tag{5}$$

Thus, the instance encoder is implemented as a network which takes as input the concatenation of the observation x_k and previously sampled group code $\hat{u}$.

$$q(v_k|x_k, u) = \mathcal{N}(\mu, \sigma),\ (\mu, \sigma) = f(x_k, u), \tag{6}$$

This form of variational latent posterior is likely to learn a disentangled representation because it has the potential to learn the true generative latent posterior, which is disentangled by definition. The generative latent posterior can be factorized as:

$$p(u, \mathbf{v}|\mathbf{x}) = p(u|\mathbf{x}) \prod_{k=1}^{K} p(v_k|x_k, u) \tag{7}$$

Existing methods use $q(v_k|x_k)$ as a proxy for $p(v_k|x_k, u)$, which makes their encoders vulnerable to the confounding effect of the group factor.

We are the first to apply the idea of conditioning the instance variable on the group variable to GID. This idea has been previously used in the related problem of sequence disentanglement (Hsu et al., 2017; Li & Mandt, 2018), another example of a group-confounded problem: the variable controlling the long time-scale confounds the inference of the variable controlling the short time scale.

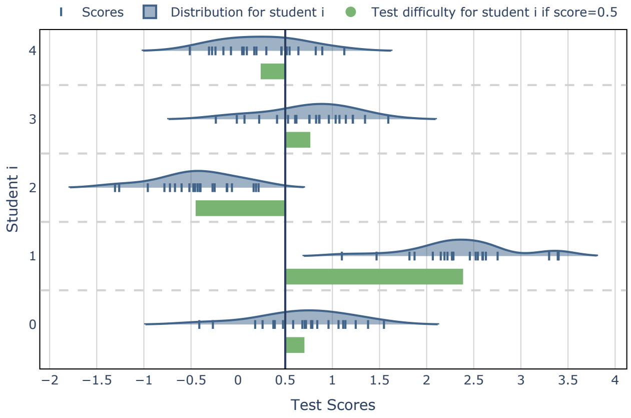


Figure 2. **Dataset with confounding: test scores.** The score of student i on test j is a function of the student's aptitude profile α_i, β_i (the group variable) and the test difficulty δ_j (the instance variable). The green bar shows "what difficulty test must have student i taken in order to get the score 0.5". The test difficulty has markedly different values for different students.

4. Evaluation

We show that by making the instance encoder conditional on the inferred group variable, we obtain a considerable gain in multiple-imputation accuracy and a marked improvement in disentanglement.

4.1. Dataset

We generate a synthetic dataset with the "varying intercept, varying slope" mixed model used in Gelman & Hill (2006) to model student scores on multiple standardised tests. The score of student i on test j is generated as a function of the student's aptitude profile α_i, β_i (the group variablce) and the test difficulty δ_j (the instance variable). We assume all factors to be normally distributed:

$$s_{ij} = 2\alpha_i - (\beta_i^2 + 1)\delta_j + \epsilon_{ij} \tag{8}$$

$$\alpha_i, \beta_i, \delta_j \sim \mathcal{N}(0, 1) \tag{9}$$

$$\epsilon_{ij} \sim \mathcal{N}(0, 0.1) \tag{10}$$

A quick look at the dataset generated by this process reveals that the group is a confounding variable (Figure 2): estimating the instance variable for a particular value of s_{ij} while ignoring the group variable yields large residuals predictive of the group effect.

For the evaluation procedure, we use the above model to generate $N = 32,768$ values for α_i, β_i and $M = 128$ values for δ_j. We then randomly select half of the students and half of the tests to generate a training dataset with $2,097,152$ scores split across $16,384$ groups. We take the other half of students and tests to generate the holdout dataset, so that every testing student and group is unseen at training time.

Then, we simulate a missing-completely-at-random pattern (Rubin, 1975) in the training data by removing each score in the dataset with a probability of 0.5. We can now build our grouped dataset by re-indexing the remaining training data such that $x_{ik} = s_{ij}$. The number of non-missing scores for student i gives the value of K_i.

We use the same exact data-generating model as in Equation (8) as a probabilistic decoder function for the autoencoder models that we evaluate.

We have chosen to evaluate on a synthetic dataset because it allows for fine control over the parameters of the data-generating process (especially relevant in Section 5) and it also enables us to measure the quality of disentanglement using the principled Mutual Information Gap (Section 4.2.2)

4.2. Metrics

We quantitatively evaluate the group-instance disentanglement methods with respect to both multiple imputation performance and quality of disentanglement.

As a general metric, we report the reconstruction error (MSE) on the holdout data for every experiment, commonly used as a proxy for the likelihood of the holdout set.

4.2.1. Multiple Imputation

We measure multiple imputation error using the following algorithm:

1. We split the holdout dataset into 128 batches of 128 students. We pick one student in each batch and randomly remove their scores with probability 0.5 (we will use these to measure the error). We then use the other 127 students in each batch to predict the missing scores for this student.

2. *How do we predict the missing scores* $\mathbf{j}$ *of student* a *with the scores of student* b*?* For student b, we infer a set of instance variables from their scores on the same tests as the missing scores of student a by sampling $q(\mathbf{v}_{b,\mathbf{j}}|\mathbf{x}_b)$. For student a, we infer the group variable using the non-missing scores by sampling $q(u_a|\mathbf{x}_{a,-\mathbf{j}})$. We then generate the set of missing scores with the group variable of student a and each instance variable of student b by sampling $p(x_{aj}|u_a, v_{aj})$ for every $j \in \mathbf{j}$ missing test.

3. We now have, for each missing score in the holdout dataset, 127 generated scores. We take the squared error between each generated score and the ground-truth missing score, and then average over all generated scores. This is our measure of error.

Multiple imputation in this setting is equivalent to the problem of unsupervised translation in group-instance disentanglement (Tenenbaum & Freeman, 2000). In our case, we are translating the scores of student a to the missing scores of student b. We can use translation as an additional qualitative comparison between the conditional and unconditional model, which can be seen in Figure 3.

4.2.2. Mutual Information Gap

We use the Mutual Information Gap (Chen et al., 2018) to measure the quality of the disentanglement. We measure empirically the amount of mutual information between the inferred latent variables $u, \mathbf{v}$ and the student effect α, β which represents the ground-truth group factor. Consequently, the goal is to have maximum mutual information between the group variable u and the ground-truth, and minimum mutual information between the instance variables $\mathbf{v}$ and the ground-truth. The gap between the two (normalized with the entropy of the ground-truth factors) is the metric of disentanglement:

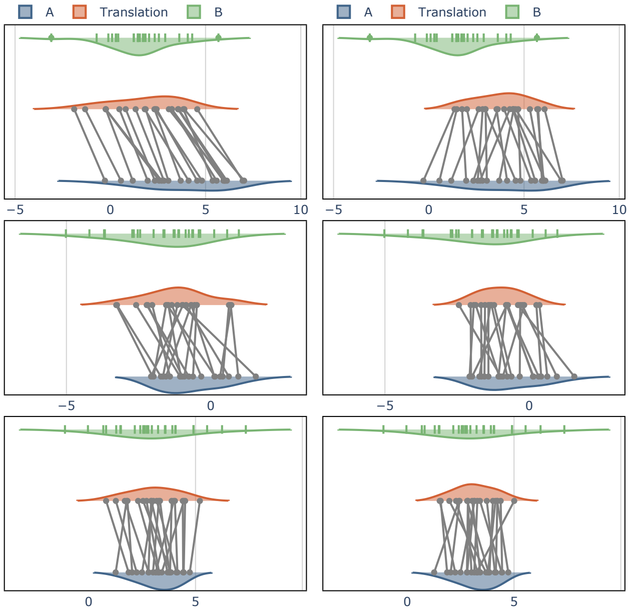
$$\text{MIG} = \frac{1}{H(\alpha, \beta)}(I(u; \alpha, \beta) - I(\mathbf{v}; \alpha, \beta)) \tag{11}$$

Since the data-generating process is known, the mutual information between the inferred group variable and the ground-truth group variable $I(u; \alpha, \beta)$ is straightforward to implement by following the process of Chen et al. (2018).

We measure only the mutual information between the ground-truth group factor and the latent variables because, as pointed out by Németh (2020), the common failure case we are trying to guard against in group-instance disentanglement is that the instance variables $\mathbf{v}$ might learn information belonging to the ground-truth group factors α, β.

4.3. Conditional vs Unconditional Model

We compare our conditional 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 implement each encoder as an MLP with 1 hidden layer of 32 activations. The decoder will be the hard-coded generative model from Equation (8). The group variable will have 2 dimensions (one for α and one for β) and the instance variable will have 1 dimension (for δ).



(a) CxVAE (b) GVAE

Figure 3. **Unsupervised translation with the unconditional GVAE vs our conditional CxVAE**. The task is to generate a set of test scores (**middle**) by translating the scores of the source student (**bottom**) onto the distribution of scores of the target student (**top**). Visual inspection suggests that our model better captures the distribution of the target group.

For all experiments, our CxVAE will be a modified GVAE such that the group variable u_i is concatenated with the observation x_{ik} and fed into the instance encoder in order to compute the instance variable v_{ik}.

We train each model for 64 epochs, and use the last 10 epochs for evaluation. Additionally, we run the experiment for 100 different random seeds initialisations, both for the data generating process and the networks. We use the same 100 seeds in each model. This gives us 1000 measurements to plot in Figure 4.

4.3.1. Results

Our CxVAE produces considerable improvements over the state-of-the-art both in terms of multiple imputation accuracy and disentanglement quality (Figure 4). While the scores of the existing methods cluster together, the gap between them and the CxVAE is larger than the 95% confidence interval. of any method. Note also that CxVAE produces a smaller reconstruction error on the holdout dataset, suggesting that disentangled representations generalise better to unseen groups.

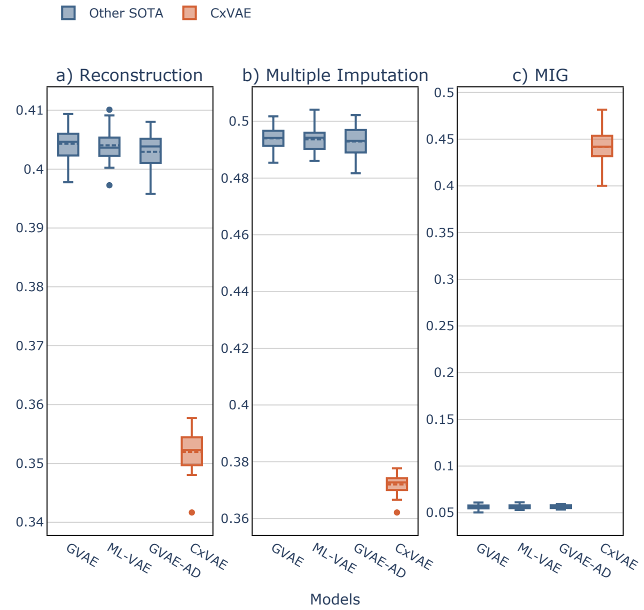


Figure 4. Our CxVAE produces considerable improvements over the state-of-the-art in every metric considered: holdout reconstruction error (**lower is better**), multiple imputation error (**lower is better**), MIG disentanglement score (**higher is better**).

5. Strength of Confounding Effect

Is the increased performance increase of the our CxVAE due to the fact that the data-generating process is confounded?

To answer this, we modify the data-generating process from Equation (8) by incorporating a hyper-parameter γ to control the strength of the confounding effect that the group variable has on the conditional distribution between the instance variable and the data observation $p(v|x)$. For $\gamma = 1$, the confounding will be the same as before. For $\gamma = 0$, there will be no confounding. Our modified data-generating process is the following:

$$s_{ij} = \begin{bmatrix} \gamma 2\alpha_i - (\beta_i^2 + 1)^\gamma \delta_j + \epsilon_{ij} \\ 2\alpha_i - \gamma(\beta_i^2 + 1)\delta_j + \epsilon_{ij} \end{bmatrix} \tag{12}$$

$$\alpha_i, \beta_i, \delta_j \sim \mathcal{N}(0, 1) \tag{13}$$

$$\epsilon_{ij} \sim \mathcal{N}(0, 0.1) \tag{14}$$

Firstly, our data is now 2-dimensional. The models for each component is very similar to the model in Equation (8) and both models share the same generative factors $\alpha_i, \beta_i, \delta_j$. The only difference between components is that in one component, γ controls the magnitude of the student effect, while in the other, γ controls the magnitude of the test effect.

This model is unconfounded when $\gamma = 0$ because each

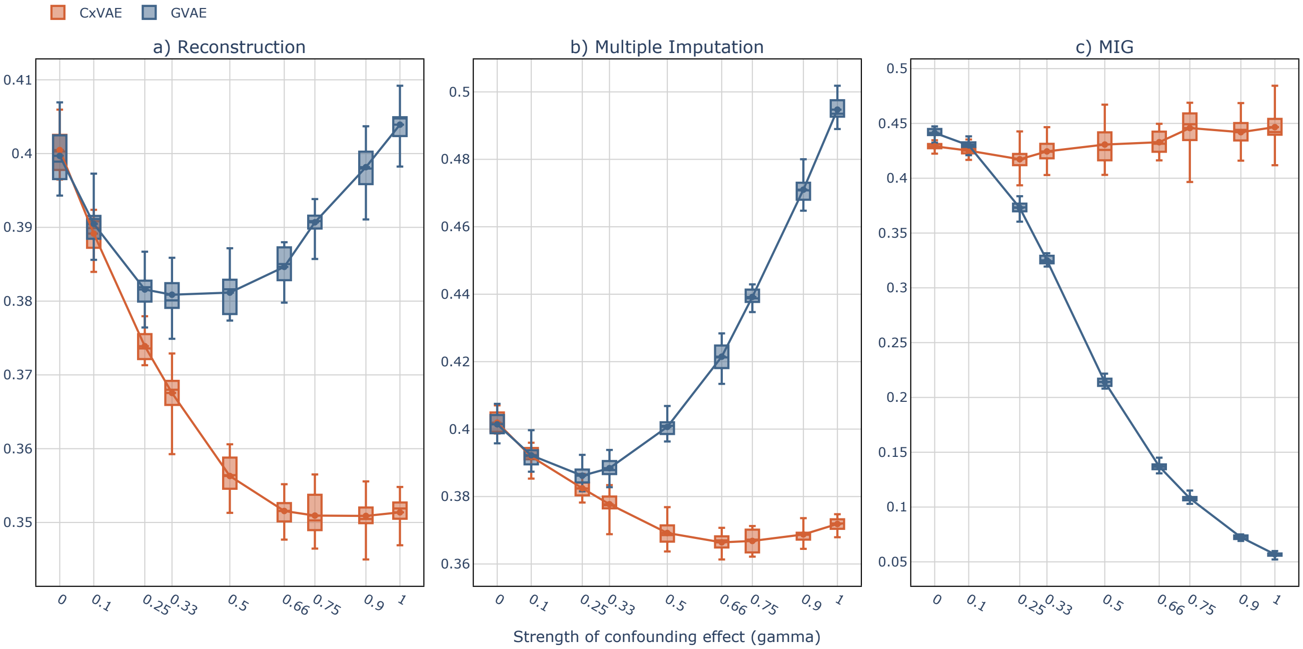


Figure 5. **The performance gain of our CxVAE increases with more confounding.** We show performance on datasets generated with different values of the γ hyper-parameter, controlling the strength of the confounding effect. For low values of γ the instance factor is not confounded and so the CxVAE and GVAE perform equally well. However, as γ increases, GVAE performs more poorly.

ground-truth factor controls a separate component of the data. Inferring the test effect requires only the test component and can ignore the student component. When $\gamma = 1$, the dimensions of the data are duplicates, so the problem is exactly the same as previously.

As an intuition, this 2-dimensional model corresponds to a setting where students take a similar practice test before each test, and we suspect that the impact the student-effect has on the score compared to the test-effect changes. Therefore, we assign one dimension of the data to the practice test and one to the actual test, and model the trade-off between effects with the hyper-parameter γ.

If our hypothesis is correct, that the confounding effect of the group variable causes the performance gap between the conditional and unconditional models, then the gap should decrease as γ approaches 0.

5.0.1. Results

The measurements displayed in Figure 5 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. Related Work

Group-Instance Disentanglement Recent work (Shu et al., 2020; Locatello et al., 2020) has identified GID as a subproblem of weakly-supervised disentanglement, where disentangled representations are learned with the help of non-datapoint supervision (e.g. grouping, ranking, restricted labelling). Early work in this area focused on separating between visual concepts (Kulkarni et al., 2015; Reed et al., 2015). This area has received renewed interest after the theoretical impossibility result of Locatello et al. (2019) and the identifiability proofs of Khemakhem et al. (2020) and Mita et al. (2021). However, a key aspect of recent weakly-supervised models is the interpretation of the grouping as a signal of similarity between datapoints (Chen & Batmanghelich, 2020).

In our case, we interpret the grouping as a structural factor which enables our model to accumulate evidence from multiple observations even at test-time. This view is clearly present in the related problem of sequence disentanglement (Hsu et al., 2017; Denton & Birodkar, 2017; Li & Mandt, 2018), where the instance variable (short timescale) is always inferred conditionally on the group variable (long timescale).

Confounding This concept has been studied since the dawn of statistics and is most rigorously defined in the causal inference literature: confounding is present if an intervention (treatment) on the source variable x changes the conditional distribution of a target variable $p(y|\mathrm{do}(x)) \neq p(y|x)$ (Greenland et al., 1999). Usually the confounding variable u is a parent of both source x and target y, and the solution is to control for u by conditioning on it or stratifying the experiment.

The notion of a grouping variable acting as a confounder when learning a predictor also has a long history in the field of supervised learning (Morabia, 2010). When the confounder manifests as a distribution shift rather than as discrete groups, this phenomenon also comes under the name of population drift (Kelly et al., 1999), a defining feature of federated learning (Kairouz et al., 2021).

7. Conclusions

In this work, we investigate the problem of group-instance disentanglement in settings where the inference of the instance variable is confounded by the group variable. We show empirically that conditioning the instance encoder on the group variable produces better disentangled representations than the unconditional models that comprise the current state-of-the-art. We also show that the strength of the confounding effect in the dataset determines the performance gap between the conditional and unconditional models. Our evaluation is run on the downstream task of multiple imputation on grouped data, a problem on which group-instance models have not been applied before.

References

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9, 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.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018.

Denton, E. L. and Birodkar, V. Unsupervised learning of disentangled representations from video. *ArXiv*, abs/1705.10915, 2017.

Gelman, A. and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

Greenland, S., Robins, J. M., and Pearl, J. Confounding and collapsibility in causal inference. *Statistical Science*, 14: 29–46, 1999.

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

Hsu, W.-N., Zhang, Y., and Glass, J. R. Unsupervised learning of disentangled and interpretable representations from sequential data. In *NIPS*, 2017.

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.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z. B., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. Y. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, O., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D. X., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14:1–210, 2021.

Kelly, M. G., Hand, D. J., and Adams, N. M. The impact of changing populations on classifier performance. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 367–371. Association for Computing Machinery, 1999. ISBN 1581131437.

Khemakhem, I., Kingma, D. P., and Hyvärinen, A. Variational autoencoders and nonlinear ica: A unifying framework. *ArXiv*, abs/1907.04809, 2020.

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

Koren, Y., Bell, R. M., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer*, 42, 2009.

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. B. Deep convolutional inverse graphics network. In *NIPS*, 2015.

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

Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *Proceedings of the 36th International Conference on Machine Learning*, 2019.

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.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. S. The variational fair autoencoder. *CoRR*, abs/1511.00830, 2016.

Mita, G., Filippone, M., and Michiardi, P. An identifiable double vae for disentangled representations. In *ICML*, 2021.

Morabia, A. History of the modern epidemiological concept of confounding. *Journal of Epidemiology & Community Health*, 65:297 – 300, 2010.

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.

Reed, S. E., Zhang, Y., Zhang, Y., and Lee, H. Deep visual analogy-making. In *NIPS*, 2015.

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

Rubin, D. B. Inference and missing data. *Psychometrika*, 1975:19, 1975.

Rubin, D. B. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489, 1996.

Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020.

Spratt, M., Carpenter, J. R., Sterne, J. A. C., Carlin, J. B., Heron, J., Henderson, J. A., and Tilling, K. Strategies for multiple imputation in longitudinal studies. *American journal of epidemiology*, 172 4:478–87, 2010.

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