Autoregressive Variational Auto-Encoder for Causal Modeling

Tom Marty [1 2]

Abstract

Causal modeling is an active area of research that consist in inferring the functional components of a causal model from multi-intervention data. Previous work usually assume that the correspondence between samples and interventional regimes is known, which is not always a realistic setting. Without this information, modeling these causal mechanisms poses significant challenges due to the inherent confounding bias that arises from the different possible interventions in the data. In this project, we propose to use a Variational autoencoder with discrete latent in order to recover the unseen interventional regimes required to deconfound the estimation of the independent causal mechanisms.

Disclaimer This report is an original research direction that the author wanted to explore, and is not directly based on any existing work. The results presented in this report are preliminary, and will be further developed in the next iteration.

1. Background

We will start by introducing some of the basic concepts that will be used throughout the report. Firstly, We will introduce the concept of Structural Causal Models (SCM), the notion of intervention on these structure, then quickly present the usual assumptions made in the causal modeling literature. If the reader is already familiar with these concepts, we recommend skipping directly to Section 2.

1.1. Structural Causal Models

A Structural Causal Model (SCM) mathematically formalize the cause-effect relationships between the different random variables in a system. Precisely, an SCM is a triplet $\mathcal{S}(\mathcal{G}, \mathbb{P}_{\boldsymbol{\epsilon}}, \mathcal{F})$ that defines the data-generating process of N endogenous variables $\mathcal{X} = \{X_1, X_2, \cdots, X_N\}$ and N independent exogenous noise terms $\boldsymbol{\epsilon} = \{\epsilon_1, \epsilon_2, \cdots, \epsilon_N\}$. It is composed of three terms :

- A Directed Acyclic Graph [1] $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that represents the causal structure of the model. Each node X_i in the graph represent a random variable and each edge $X_i \to X_j$ represent a direct causal relationship between the two variables.
- A distribution over N independent exogenous noise terms $\boldsymbol{\epsilon} = \{\epsilon_1, \epsilon_2, \cdots, \epsilon_N\}$ that represent the unobserved factors influencing the endogenous variables.
- A set of N mechanism $\mathcal{F} = \{f_1, f_2, \cdots, f_N\}$ that represent the functional form of the causal relationships between the variables in the model. Each variable X_i is a function of its parents $\text{Pa}(X_i)$ defined by the DAG and some exogenous noise ϵ_i sampled from $\mathbb{P}_{\boldsymbol{\epsilon_i}}$:

$$X_i = f_i(\text{Pa}(X_i), \epsilon_i) \quad \epsilon_i \sim \mathbb{P}_{\epsilon_i} \tag{1}$$

This formulation induces a set of conditional probability distributions $p(X_i|\text{Pa}(X_i))$ such that the joint distribution over the variables X can be expressed as a product of these conditional distributions:

$$p(X) = \prod_{i=1}^{N} p(X_i|\text{Pa}(X_i)).$$

For the rest of this report, we will consider the variables X_i and ϵ_i to be univariate.

1.2. interventions and targets

An important concept in the field of Causality is the notion of **intervention**. An intervention $\mathcal{I}$ is a modification of the data-generating process of a causal model $\mathcal{S}$ that consist in modifying a set of mechanisms $f_k \in \mathcal{F}_{\mathcal{I}}$:

$$f_k(\text{Pa}(X_i), \epsilon_i) \to f_k^{\mathcal{I}}(\text{Pa}(X_i), \epsilon_i) \quad \forall k \in \mathcal{F}_{\mathcal{I}} \tag{2}$$

As defined, an intervention can possibly affect multiple mechanisms in the system. For a given intervention, the **targets** refers to the set of variables in the causal model

[1] Université de Montréal, Québec, Canada [2] Mila, Quebec AI Institute. Correspondence to: Tom Marty <tom.marty@mila.quebec>.

[1] *commonly called a DAG

whose mechanisms are altered by the intervention. There are different types of interventions that can be considered, we usually consider *single* intervention where the set $\mathcal{F}_\mathcal{I}$ is reduced to a single altered mechanism, we refer the reader to Pearl (2009) for a more detailed discussion on interventions.

interventional regime Another important notion is the notion of interventional regime, which caracterise a set of samples $D_{\mathcal{I}_k}$ that were generated under the same intervention $\mathcal{I}_k$. Additionally, when no intervention is applied, the data is said to be generated under the observational regime $\mathcal{O}$.

interventional Faithfulness First introduced by Chevalley et al. (2024), this assumption guarantees that any valid intervention $\mathcal{I}_k$ should significantly alter the distribution of the downstream nodes of the intervened mechanism such that inversely inferring the possible intervention from the downstream nodes is possible.

2. Motivation and Setting

In causal modeling, given a dataset $\mathcal{D}$ generated from a partially-known SCM $\mathcal{S} = (\mathcal{G}, \mathbb{P}_\epsilon, \mathcal{F})$, the goal is usually to retrieve full knowledge about $\mathcal{S}$. On the one hand, in *Causal Discovery*, the challenge is to infer the graph $\mathcal{G}$, which is a particularly arduous task because of the inherent difficulties arising from discrete optimization (a graph a is discrete structure) under constraint (e.g. acyclicity). On the other hand, in *Causal Modelling*, the goal is to learn the mechanisms $\mathcal{F}$. Depending on what is to be learned, different assumptions have to be made on the data-generation process in order to identify the causal model. These assumptions can be summarized as follows:

1. unknown/known interventional regime
2. unknown/known single/multiple interventional targets
3. unknown/known graph structure
4. restricting assumptions on the mechanisms (e.g. Additive noise, linear mechanisms, gaussian noise...)

In this project, we will consider a slightly unusual setting : We consider a dataset $\mathcal{D} = \{D_\mathcal{O}, D_{I_1}, \cdots, D_{I_N}\}$ composed of samples generated under $N+1$ regimes containing the observational regime $D_\mathcal{O}$ and N interventional regimes D_{I_k}. Each regime is obtained by performing a soft intervention on a single mechanism. Each mechanism is intervened only once, meaning that there are N mechanisms and N corresponding interventions. We discuss the relevance of such a specific setting in Section 2.1. Finally, we consider **that the graph $\mathcal{G}$ is known**. This assumption is not absurd since the graph structure can often be obtained beforehand using expert domain knowledge or by running a causal discovery algorithm. Similar to (Faria et al., 2022), the subtlety of our setting is that the data is completely shuffled such that for each sample $X \in \mathcal{D}$ neither the interventional regime it belongs to nor the interventional target of this regime is known.

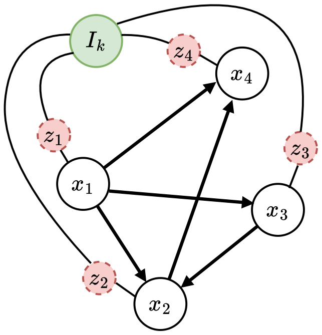

Figure 1. Graphical model of the data-generation process augmented with the unobserved latent variables. I_k represent the interventional regime and the z_i $i \in [1, N]$ are boolean variable that encode the state of each mechanism $p(X_i|\text{Pa}(X_i))$

.

Goal The objective of the problem is two-fold :

- Cluster the samples $X_i \in \mathcal{D}$ according to the different regimes and identify which one correspond to the observational regime $D_\mathcal{O}$.
- Model the mechanisms $f(X_i|\text{Pa}(X_i))$ under each of the interventional regime D_{I_k}.

We treat this problem as a latent-variable model where the latent variables encode the interventional regime for the input sample. Given that we usually don't know the functionnal form of the mechanisms of the true causal model, the posterior distribution over the latent variables is most of the time intractable. For this reason, we decided to use a variational Auto-Encoder (VAE) to approximate the posterior distribution over an arbitrary variational family. We refer the reader to Section 3 for a more detailed explanation of the method. In Figure 1, we illustrate the data-generation process.

As we mentioned previously, usually the challenge in causal modelling lies in the inference of the causal structure (Brouillard et al., 2020; Lorch et al., 2022). Even with a known graph in this case, we argue that inferring the mechanisms remains challenging because of the unobserved confounding regime that governs these mechanisms. Indeed, straightforward maximization of data-loglikelihood will yield probabilistic models $\hat{p_{obs}}(X_i|\text{Pa}(X_i))$ that will be biaised by samples for which the true mechanism

$p_{obs}(X_i|\text{Pa}(X_i))$ was in fact not in its observational state : $p_{obs} \to p_{int}$. In order to unbiaise this estimate, we therefore need to account for these possible interventions by also modeling and estimating the state of each mechanism for the sample X_i using some categorical variable. Recalling that data-points sampled under the same interventional regime were obtained using the same set of mechanisms, we can instead aim to encode the regime itself in the latent variable, in an attempt to encoder this information more compactly.

2.1. Necessary assumption for identification

One of our goal is to *identify the observational regime*, for which no internvention was applied. In order to do so, we argue that we need to make additional assumptions on the data-generation process, otherwise the observational regime will be indistinguishable from the other regimes. In Figure 2, we discuss the necessity of these assumptions trought a simple example with two variable and one mechanism. In this example, we have two variables X_1 and X_2. We have access to infinite data comming from possibly different regimes. Finally we assume that the intervention applied is strong enough to significantly alter the data distribution, such that each sub-distribution that generated D_{I_k} is well separated from the others. Trought this example, we argue that the necessary conditions for the identification of the observational regime are **single-target internventional regime** and **unique internvention per mechanism**. Formal proof of this claim is left for future work.

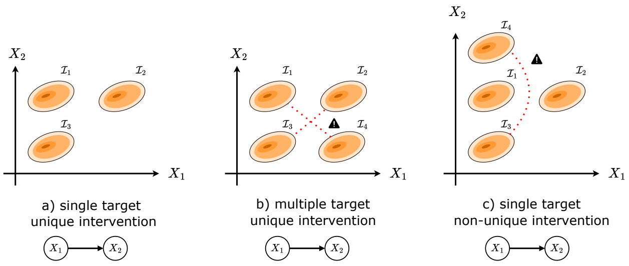

Figure 2. Illustration of the necessity of the assumption considered on the data-generation process. a). **Single Target / Uniq. Intervention**: The regime $\mathcal{I}_1$ can't be the observational regime $\mathcal{O}$ because this would break the interventional-Faithfulness assumption ($\mathcal{I}_2$ would then correspond to a intervention made on marginal $p(x_1)$ which should also alter the marginal of its descendent: $p(x_2)$, which is not what we observe). $\mathcal{I}_2$ can't be $\mathcal{O}$ because this would break the single target assumption (In $\mathcal{I}_1$, while the marginal $p(x_1)$ has changed, $p(x_2)$ remain not affected. For this to be possible is to have performed an intervention on $p(x_1)$ and one on $p(x_2|x_1)$ to cancel out the effect of the first one on $p(x_2)$). Therefore, the only possible observational regime is $\mathcal{I}_3$. b) **Multiple Target / Uniq. Intervention**: Here we add another regime $\mathcal{I}_4$ targeting both $p(x_1)$ and $p(x_2|x_1)$. In this case, the regimes become completely symmetric and the observational regime can't be identified. c) **Single Target / Non-Uniq. intervention**: Here we add another regime $\mathcal{I}_4$. In this case, $\mathcal{I}_1$ and $\mathcal{I}_4$ have symmetric role and the observational regime once again can't be identified.

3. Method

In this problem we are trying to learn a generative model of the data under unobserved interventions, for this purpose we will use a variational Auto-Encoder (Kingma & Welling, 2022) with discrete latent variables, the decoding will take profit of the known graph structure to reconstruct the variables X_i from their sampled parents $\text{Pa}(X_i)$, in an autoregressive fashion. The integration of known causal structure into VAEs is not a novel idea, Komanduri et al. (2022) used the same kind of structural prior to enforce disentangled representations that align with the causal structure. In our setting the latent variables are the interventional target $z_i \quad i \in [1, N]$ of each sample x. In order to faithfully model the data-generation process, A VAE is a general framework for solving statistical modeling problem with unobserved latent variables. Compared to the standard EM for which a closed form update does not always exist, variational methods allows to describe a wider variety of posterior distributions, allowing to tackle more complex and realistic problems. Given a data point x, the Variational Auto-encoder (VAE) is tasked to learn an approximate posterior distribution over these latent variables $q_\phi(z|x)$. Basically, the algorithm consist in maximizing the **evidence lower bound (ELBO)**, a lower bound of the log-likelihood of the data defined as follows :

$$\log p_\theta(x) \geq \underbrace{-\text{KL}\left(q_\phi(z|x) \,||\, p_\theta(z)\right) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]}_{\text{ELBO}}$$

In this optimization problem, we want to find the parameters θ and ϕ that maximize the ELBO. $q_\phi(z|x)$ is the approximate posterior distribution over the latent variables and ϕ defines the variational family of distributions used to approximate the posterior. $p_\theta(z)$ is the prior distribution over the latent variables and $p_\theta(x|z)$ is the likelihood of the data given the latent variables. We refer to figure 3 for a representation of the full architecture of the model.

Regime Learning ($q_\phi(z|x)$) Given that for each mechanism f_i, we have a set of samples $D_{\mathcal{I}_k}$ that were generated under the same intervention $\mathcal{I}_k$, we can leverage this shared characteristic to better encode the interventions. We therefore propose to use a boolean latent variable z_i that encodes the state of each mechanism f_i (intervened or not intervened). We use the following formulation : $\tilde{z} \sim \mathcal{B}(\pi_\phi(x))$ and $z = \text{one-hot}(\tilde{z})$, where $\mathcal{B}$ is a multinoulli distribution parameterized by $\pi_\phi(x)$, the output of a neural-network that predict the probability of each mechanism to be intervened.

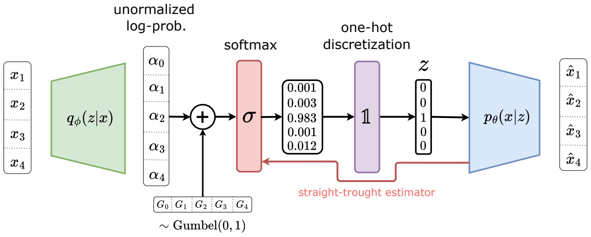

Figure 3. Architecture of the variational Auto-Encoder. An encoder takes as input the data x and outputs the parameters of the approximate posterior categorical distribution $q_\phi(z|x)$. The sampled discrete latent variables z are then fed to the decoder that outputs the parameters of the likelihood distribution $p_\theta(x|z)$. During backpropagation, the discrete latents are approximated using the Gumbel-Softmax re-parametrization trick. For more details on the architecture of the likelyhood model $p_\theta(x|z)$, we refer to Figure 4.

Prior distribution ($p_\theta(z)$) Without any additional information on the proportion of each regime, we will use a uniform prior over the latent variables z which simplifies the derivation of the Kullback-Leibler divergence term of the ELBO :

$$\begin{aligned}
-\text{KL}\left(q_\phi(z|x)\,||\,p_\theta(z)\right) &= -\sum_i \pi_\phi(x)_i \log \frac{\pi_\phi(x)_i}{p_\theta(z_i)} \\
&= -\sum_i \pi_\phi(x)_i \log \frac{\pi_\phi(x)_i}{1/N} \\
&= -\sum_i \pi_\phi(x)_i \log \pi_\phi(x)_i - \log N \\
&= \text{H}(\pi_\phi(x)) - \log N
\end{aligned}$$

Mechanism Learning ($p_\theta(x|z)$) Given that the graph $\mathcal{G}$ is known, we can use the following factorization of the likelihood :

$$\begin{aligned}
p_\theta(x|z) &= \prod_{i=1}^{N} p_\theta(x_i|\text{Pa}(X_i), z_i) \\
&= \prod_{i=1}^{N} p_\theta(x_i|\text{Pa}(X_i), z_i)
\end{aligned}$$

We assume that $p_\theta(x_i|\text{Pa}(X_i), z_i) = \mathcal{N}(\mu_\theta(x_i, \text{Pa}(X_i), z_i), 1)$, where $\mu_\theta(x_i, \text{Pa}(X_i), z_i)$ is the output of a neural-network that takes as input the parents of X_i and the inferred state of the mechanism z_i. We model the likelihood as a Gaussian distribution with unit variance to simplify the optimization of the ELBO. This design choice suppose that the data was sampled from an SCM with additive Gaussian (Normal) noise. In future work, we plan to extend this model to more complex mechanisms. The likelihood term of the ELBO can now be written as follows :

We will use a to model each of the conditional distributions $p_\theta(x_i|\text{Pa}(X_i), z_i)$ for each mechanism f_i.

$$\begin{aligned}
&\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \\
&\simeq \underbrace{\frac{1}{L}\sum_{l=1}^{L}}_{\text{n latent samples}} \sum_{i=1}^{N} \log \mathcal{N}(x_i|\mu_\theta(x_i, \text{Pa}(X_i), z_i), 1) \\
&\simeq \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{N}\left(-\frac{1}{2}\left(x_i - \mu_\theta(x_i, \text{Pa}(X_i), z_i)\right)^2\right) + \text{const.}
\end{aligned}$$

Maximizing the expected log-likelihood of the data under the model is equivalent to minimizing the MSE loss for each mechanism of the causal model. In practice, when considering large enough mini-batch, we can use single-sample Monte-Carlo estimation $L = 1$ to limit the computational cost of the ELBO optimization.

In order to sample from the likelihood model, similarly to Yang et al. (2021) with their *Causal Layer*, we will use an autoregressive decoder to reconstruct the input sample x following the topological order of the graph $\mathcal{G}$. This is done by sampling each variable X_i given its parents $\text{Pa}(X_i)$ and the latent variable z_i encoding the state of the mechanism. We refer the reader to Figure 4 for a representation of the architecture of the likelihood model $p_\theta(x|z)$. Due to the stochastic nature of the sampling process, it is impossible for the model to reconstruct the input sample X_i perfectly. Instead, the model can only aim to approximate its conditional mean $\mathbb{E}_{p(X_i|\text{Pa}(X_i), z_i)}[X_i]$.

discrete re-parametrization trick The latent variable z lies in a discrete space, which makes the objective non-differentiable with respect to the sampling process $q_\phi(z|x)$. For this reason, we will use the Gumbel Softmax re-parameterization trick (Jang et al., 2017) to make the sampling process differentiable while keeping the discrete nature (i.e. boolean) of the latent variables.

4. Experiments

In this section, we discuss the experiments that we run to evaluate the performance of our model at inferring the right regimes and learning the mechanisms of the causal model. We will first describe the data-generation process, then we will present the evaluation metrics used to assess the performance of our model. The method is implemented in Pytorch Lightning and the code to reproduce precisely the experiments and plots is available at https://github.com/3rdCore/VariationalCausalModelling.

4.1. Data-generation process

To evaluate our model, we need a dataset that was generated from a known causal model that satisfy the assumptions

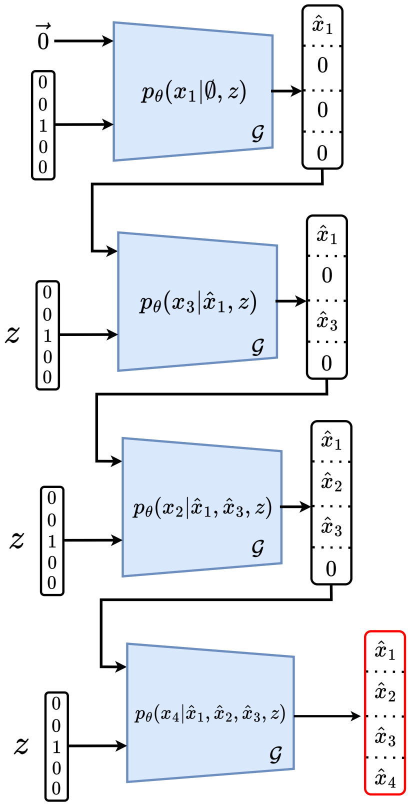
$\vec{0}$
0 0 1 0 0
$p_\theta(x_1|\emptyset, z)$
$\mathcal{G}$
$\hat{x}_1$ 0 0 0
z
0 0 1 0 0
$p_\theta(x_3|\hat{x}_1, z)$
$\mathcal{G}$
$\hat{x}_1$ 0 $\hat{x}_3$ 0
z
0 0 1 0 0
$p_\theta(x_2|\hat{x}_1, \hat{x}_3, z)$
$\mathcal{G}$
$\hat{x}_1$ $\hat{x}_2$ $\hat{x}_3$ 0
z
0 0 1 0 0
$p_\theta(x_4|\hat{x}_1, \hat{x}_2, \hat{x}_3, z)$
$\mathcal{G}$
$\hat{x}_1$ $\hat{x}_2$ $\hat{x}_3$ $\hat{x}_4$

Figure 4. Architecture of the likelihood model $p_\theta(x|z)$. Each mechanism f_i is modeled as a gaussian distribution with mean $\mu_\theta(x_i, \text{Pa}(X_i), z_i)$ and unit variance. The mean of the input sample is reconstructed autoregressively following the topological order of the graph $\mathcal{G}$.

we presented in Section 2. In this project, we consider that the SCM that generated the data is an additive noise model (ANM) with linear mechanisms[2]. The data is generated as follows : We have N variables $X = \{X_1, X_2, \cdots, X_N\}$. The DAG $\mathcal{G}$ is sampled from a Erdos-Renyi DAG distribution with a fixed probability of edge $p = 0.3$. Given that structure, the mechanisms f_i are defined as linear functions of their parents and the exogenous noise terms :

$$X_i = f_i(\text{Pa}(X_i), \epsilon_i) = \sum_{j \in \text{Pa}(X_i)} w_{ij} X_j + \epsilon_i \qquad (3)$$

In order to satisfy the faithfulness assumption, we sample the weights w_{ij} from a truncated uniform distribution $w_{ij} \sim \mathcal{U}(-1, 0.1) \cup \mathcal{U}(0.1, 1)$. In practice, even if this assumption is not verified, the model should still be able to ignore the unfaithful parents and learn the right mechanisms. The exogenous noise terms $\boldsymbol{\epsilon}$ are sampled from a standard normal distribution. In a future iteration, more complex causal model will be considered. The data is generated under N different interventions, each intervention consist in a shift $delta_i$ in the linear mechanisms f_i. Again, the shift is manually fixed to $delta_i = 5$, such that the data is well separated under each regime. Each sample has a 20% chance of being generated under the observational regime and a 80% chance of being generated under one of the N interventional regime. We generate a dataset of 10000 samples in total.

4.2. Model and Training Details

Encoder / Decoder Architecture It is implemented with a 3-layers MLP. Each layer in the MLP consists of 32 units and uses ReLU activation. We used a temperature of $T = 0.5$ for the Gumbel-Softmax re-parametrization trick.

For the decoder, each mechanism is structured as a 3-layers MLP, each with 32 units and ReLU activations. The final layer outputs the mean for the Gaussian distributions used to model each conditional probability distribution under intervention.

Training Procedure The model is trained using the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 64. Training is done over a maximum of 10 epochs. KL vanishing makes the training of VAEs with auto-regressive decoder difficult ,a phenomenon well-studied in the NLP literature. In order to mitigate this issue, following Fu et al. (2019), we use a cyclical annealing schedule for the KL term of the ELBO with $M = 5$ cycles and a proportion of 0.5 for the annealing fraction. Finally all the result are averaged over 10 different seeds.

[2] For ANM a closed-form update for EM might have existed

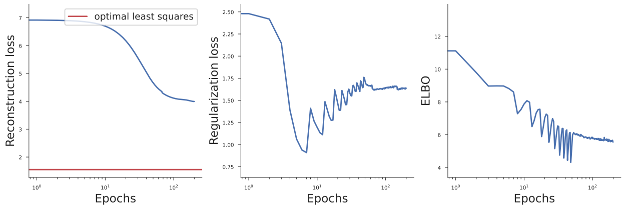

Figure 5. Training curves of the model. Horizontal line represent the optimal least-square error between the sample value and its conditional mean. The ELBO is decomposed into the reconstruction error and the KL term. The reconstruction error decreases steadily while the KL term oscillates following the cyclical annealing schedule.

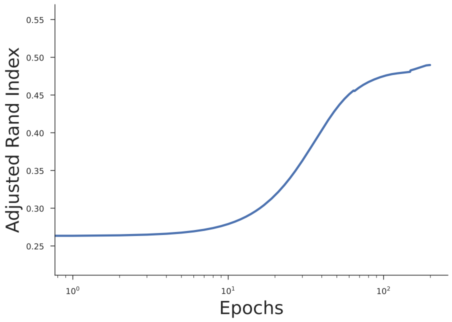

Figure 6. ARI of the model at identifying the different interventional regimes. The model is able to identify the different regimes on 50% of the samples.

4.3. Evaluation metrics

Reconstruction error The reconstruction error is the mean squared error between the input sample x and the output of the decoder $\hat{x}$. The optimal reconstruction error corresponds to the least-square error between the sample value and its conditional mean.

Regime identification Given the actual model, the latent space is not constrained in any way such that the latent z_i actually encode the interventional regime of variable X_i. Said differently, the latent variables z_i identify the interventional regime up to a permutation. We left the problem of aligning the latent variables with the true interventional regime, necessary for identifying the observational regime, for future work. We use the Adjusted Rand Index (ARI) to account for this permutation and evaluate the performance of the model at correctly identifying the different regimes.

4.4. Results

Due to time constraints, we were only able to run preliminary experiments using our pipeline and the results are promising. In Figure 5, we present the training curves of the model on a data sampled from a ANM with 10 variables. As we can see, the reconstruction error decreases steadily but does not converge to the least-square between the sample value and its conditional mean which indicates that the model is not able to perfectly model the mechanisms.

In Figure 6, we present the ARI of the model at identifying the different interventional regimes. We observe that by the end of the training, the model is able to correctly identify the sample regime on about half of the dataset. The previously observed inability of the model to perfectly model the mechanisms is likely due to the confounding effect stemming from the prediction of the wrong regime on the other half of the dataset. This issue seems to be a problem of optimizing the VAE itself and will be further investigate. In future work, we plan to run a more extensive set of experiment to validate our method and alongside an ablation study to confirm the necessity of some of the assumption we made in the data-generation process.

5. Conclusion

In this project, we propose to tackle the problem of causal modelling under unknown interventional regime as a latent variable model. We use a variational Auto-Encoder with discrete latent variables to jointly infer the interventional regime and the causal mechanisms in their different regime. It is too early to draw strong conclusion on the ability of our method to systematically identify the regime, but we can already see that the model is able to identify the interventional regime with a relatively good accuracy.

References

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable causal discovery from interventional data, 2020. URL https://arxiv.org/abs/2007.01754.

Chevalley, M., Schwab, P., and Mehrjou, A. Deriving causal order from single-variable interventions: Guarantees and algorithm, 2024. URL https://arxiv.org/abs/2405.18314.

Faria, G. R. A., Martins, A., and Figueiredo, M. A. T. Differentiable Causal Discovery Under Latent Interventions. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pp. 253–274. PMLR, June 2022. URL https://proceedings.mlr.press/v177/faria22a.html. ISSN: 2640-3498.

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., and Carin, L. Cyclical annealing schedule: A simple approach to mitigating kl vanishing, 2019. URL https://arxiv.org/abs/1903.10145.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax, 2017. URL https://arxiv.org/abs/1611.01144.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Komanduri, A., Wu, Y., Huang, W., Chen, F., and Wu, X. SCM-VAE: Learning Identifiable Causal Representations via Structural Knowledge . In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 1014–1023, Los Alamitos, CA, USA, December 2022. IEEE Computer Society. doi: 10.1109/BigData55660.2022.10021114. URL https://doi.ieeecomputersociety.org/10.1109/BigData55660.2022.10021114.

Lorch, L., Sussex, S., Rothfuss, J., Krause, A., and Schölkopf, B. Amortized inference for causal structure learning, 2022. URL https://arxiv.org/abs/2205.12934.

Pearl, J. *Causality*. Cambridge University Press, 2 edition, 2009.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. Causalvae: Disentangled representation learning via neural structural causal models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9588–9597, 2021. doi: 10.1109/CVPR46437.2021.00947.